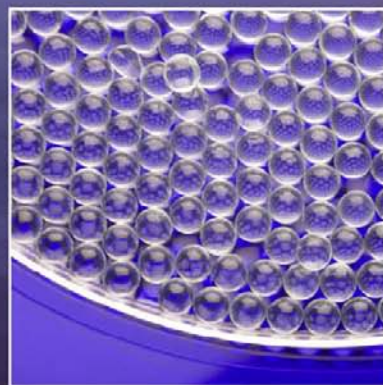Jeremy J. Ramsden

NANOTECHNOLOGY

An Introduction

Nanotechnology: An Introduction

Micro & Nano Technologies Series

Published Titles

9780815515739	Francois Leonard	*Physics of Carbon Nanotube Devices* (2009)
9780815515784	Mamadou Diallo, Jeremiah Duncan, Nora Savage, Anita Street & Richard Sustich	*Nanotechnology Applications for Clean Water* (2009)
9780815515876	Rolf Wüthrich	*Micromachining Using Electrochemical Discharge Phenomenon* (2009)
9780815515791	Matthias Worgull	*Hot Embossing* (2009)
9780815515838	Waqar Ahmed & M.J. Jackson	*Emerging Nanotechnologies for Manufacturing* (2009)
9780080964546	Richard Leach	*Fundamental Principles of Engineering Nanometrology* (2009)
9780815520238	Jeremy Ramsden	*Applied Nanotechnology* (2009)
9780815515944	Veikko Lindroos, Markku Tilli, Ari Lehto & Teruaki Motooka	*Handbook of Silicon Based MEMS Materials and Technologies* (2010)
9780815515869	Matthew Hull & Diana Bowman	*Nanotechnology Environmental Health and Safety* (2009)
9781437778236	Yugang Sun & John Rogers	*Semiconductor Nanomaterials for Flexible Technologies* (2010)
9781437778403	Victor Zhirnov & Ralph K. Cavin	*Microsystems for Bioelectronics* (2010)
9781437778489	Zeev Zalevsky & Ibrahim Abdulhalim	*Integrated Nanophotonic Devices* (2010)
9780815515456	Yi Qin	*Micromanufacturing Engineering and Technology* (2010)
9780080964478	Jeremy Ramsden	*Nanotechnology: An Introduction* (2011)

Forthcoming Titles

9780815515821	Regina Luttge	*Microfabrication for Industrial Applications* (2011)
9781437778632	Gurumurthy Ramachandran	*Assessing Nanoparticle Risks to Human Health* (2011)
9781455778621	Karthikeyan Subramani & Waqar Ahmed	*Emerging Nanotechnologies in Dentistry* (2012)

Nanotechnology: An Introduction

Jeremy Ramsden
Cranfield University

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience
broaden our understanding, changes in research methods, professional practices, or medical treatment
may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating
and using any information, methods, compounds, or experiments described herein. In using such
information or methods they should be mindful of their own safety and the safety of others, including
parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any
liability for any injury and/or damage to persons or property as a matter of products liability, negligence
or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in
the material herein.

Contents

Preface

There are already hundreds of books in print on nanotechnology in English alone, at all levels up to that of the advanced researcher and ranging in coverage from detailed accounts of incremental improvements in existing microtechnologies to far-reaching visionary approaches toward productive nanosystems. Furthermore, not only do many old-established scientific journals in the fields of physics and chemistry now carry nanotechnology sections, but numerous scientific journals dedicated solely to nanotechnology have appeared and new ones are launched each year. In addition, a flood of commercially oriented reports about present and future nanotechnology markets is constantly being produced; these are often weighty documents comprising a thousand pages or more, whose reliability is difficult to assess, since even if they make use of publicly available data, the processing of those data in order to arrive at predictions is typically carried out using unrevealed algorithms.

Faced with this huge and burgeoning literature, the newcomer to the field, who may well have a strong background in one of the traditional disciplines such as physics, mechanical or electrical engineering, chemistry or biology, or who may have been working on microelectromechanical systems (MEMS) (also known as microsystems technology, MST), is likely to feel confronted with a somewhat chaotic and colorful scene from which it is often difficult to extract meaning. The goal of this book is to tackle that problem by presenting an overview of the entire field, focusing on key essentials, with which the reader will be able to erect his or her own personal scaffold on which the amazingly diverse plethora of detailed information emerging from countless laboratories all over the world can be structured into some kind of coherent order. The emphasis is therefore on concepts; any attempt to complement this by capturing all the latest specific discoveries and inventions in nanotechnology, other than illustratively, would almost immediately become out of date; the aim of this book might briefly be stated as being "to make sense of nanotechnology"—to explain things thoroughly enough to be intelligible while still remaining introductory.

The book itself is structured around a robust anatomy of the subject. Following a basic introduction (Chapter 1), which includes a brief history of the subject, careful consideration is given to the meaning of the nanoscale (Chapter 2), on which everything else is dependent, since nanotechnology can most simply (but cryptically) be defined as "technology (or engineering) at the nanoscale". This chapter in itself constitutes a succinct summary of the entire field. Chapter 3 is devoted to interfacial forces, which govern key aspects of behavior at the nanoscale. Chapter 4 covers the nano/bio interface, which plays a fundamental role in the continuing evolution of nanotechnology, and Chapter 5 deals with the demanding issues of metrology in nanotechnology, which have also strongly influenced nanofabrication technology. In this chapter, the metrology of the nano/bio interface is covered in detail, since this is one of the newest and least familiar parts of the field. Nanomaterials (both nano-objects and nanostructured materials) are covered in Chapter 6—except carbon nanomaterials (and devices), which merit a separate chapter (9). Nanoscale

devices of all kinds (except those based on carbon)—mainly information processors and transducers, including sensors are the topic of Chapter 7, and strategies for their fabrication are covered in Chapter 8, devoted to the three fundamental approaches towards achieving nanoscale manufacture (nanofacture), namely the top–down methods rooted in ultraprecision engineering and semiconductor processing, the bottom-to-bottom approach that is closest to the original concept of nanotechnology (the molecular assembler), and the bottom–up (self-assembly) methods that have been powerfully inspired by processes in the living world. Problems of materials selection, design and so forth are treated in Chapter 10, especially how to deal with vastification; that is, the vast numbers of components in a nanosystem and the almost inevitable occurrence of defective ones. Chapter 11 is devoted to bionanotechnology, defined as the incorporation of biomolecules into nanodevices. The final chapter (12) deals with the impacts of nanotechnology: technical, economic, social, psychological and ethical. Each chapter is provided with a succinct summary at the end as well as suggestions for further reading. A glossary of nanotechnology neologisms is appended, along with a list of the most common abbreviations.

The primary readership is expected to be engineers and scientists who have previously been working in other fields but are considering entering the nano field and wish to rapidly acquire an appreciation of its vocabulary, possibilities and limitations. The secondary readership is anyone curious about nanotechnology, including undergraduates and professionals in other fields. The book should also appeal to those less directly connected with science and engineering, such as insurers and lawyers, whose activities are very likely to be connected with nanotechnology in the future, and traders, commodity brokers and entrepreneurs in general dissatisfied with remaining in ignorance of the technology that they are making use of. It is designed to equip the reader with the ability to cogently appraise the merits or otherwise of any piece of nanotechnology that may be reported in one form or another.

It is a distinct characteristic of nanotechnology that many of its features draw heavily from existing work in chemistry, physical chemistry, physics and biology. Hence, there is relatively little domain-specific knowledge associated with nanotechnology. Most of the themes in this book are covered in great detail in specialist literature that may not exclusively or even overtly be associated with nanotechnology. It seems, therefore, that nanotechnology is most aptly globally characterized as an attitude or mindset, comprising above all the desire both to understand the world at the atomic level and to create objects of beauty and utility by controlling matter at that level. Unlike the science of the subatomic level, however, nanotechnology necessarily concerns itself with superatomic levels as well, since the ultimate objects of its creation must be macroscopic in order to be of use to humanity. Hence, problems of emergent properties also form a part of nanotechnology.

The uniqueness of this book resides in its unifying viewpoint that draws many disparate pieces of knowledge together to create novel technologies. These nanotechnologies are in turn united by the distinctive attitude associated with nanotechnology.

Nanotechnology is inseparably associated with the emergence of qualitatively different behavior when a quantitative difference, namely increasing smallness,

becomes great enough: one might call this a Hegelian viewpoint of nanotechnology. Phenomena that are merely modified *pari passu* with diminishing size without any qualitative change should not, therefore, strictly rank as nanotechnology. This viewpoint avoids the pitfall of having to group practically all of physics, chemistry and biology under nanotechnology because of too vague a definition.

Inevitably, the author of a book of this nature is greatly indebted to countless colleagues and correspondents both at Cranfield and throughout the world. It would be invidious to mention some without mentioning all, and they are too numerous for the latter, but I hope that in this era of a voluminous research literature their inputs are adequately reflected in the reference list.

Jeremy J. Ramsden
Cranfield
May 2010

…Kicsinyben rejlik a nagy,
Olyan sok a tárgy, s létünk oly rövid.

IMRE MADÁCH

What is Nanotechnology?

1

CHAPTER CONTENTS

Nanotechnology is above all a mindset, a way of thinking about the world that is rooted in atomically precise perception. As such, it represents the apotheosis of man's ceaseless urge to understand the world and use that understanding for practical purposes. Well synonymized as "atomically precise technology", it encapsulates the vision of building "our earthly estate" atom-by-atom, controlling architecture, composition and hence physical properties with atomic resolution. "Hard" nanotechnologists promote a future world in which every artifact (and even food) can be constructed atom-by-atom from a feedstock such as acetylene, requiring in addition only energy and instructions. A more pragmatic view accepts that there are many intermediate stages in which partially atomically precise construction can improve existing artifacts and create new ones. Similarly, the resolute aim of "hard" nanotechnologists is to create productive nanosystems (PN) working with atomic precision—the nanoscale assemblers that would execute the instructions and build everything we need from the bottom upwards, whereas a more pragmatic view accepts that while in principle everything can be reproduced and many things imitated via atom-by-atom assembly, in many cases the improvement in properties or performance would be negligible and a hybrid approach will best serve the needs of

1

humanity. This is particularly likely to be the case for large artifacts (such as human dwellings or airplanes) and for relatively complex products such as food, which can be quite easily grown naturally.

In this chapter we shall first look at the basic definitions for nanotechnology, and sketch a concept system ("ontology") for the field. It is also possible to define nanotechnology ostensively, according to what is already generally considered to be nanotechnology, and extended by what is envisaged in the future. A further way of defining it is through its history. We also briefly look at the relation of nanotechnology to biology, which has been a powerful paradigm for convincing engineers that nanotechnology is possible—nanobiotechnology and bionanotechnology form the topics of subsequent chapters (4 and 11 respectively). General motivations for nanotechnology are considered—"Why nanotechnology?" Attention is drawn to the appended list of neologisms associated with nanotechnology (Appendix, p. 00).

1.1 DEFINITIONS AND CONCEPTS

1.1.1 Working Definitions

The simplest definition of nanotechnology is "technology at the nanoscale". The various definitions currently circulating can be reasonably accurately thus paraphrased. Obviously, this definition is not intelligible in the absence of a further definition, namely that of the nanoscale. Furthermore, definitions of components of nanotechnology, such as "nanofiber", also refer to the nanoscale; indeed every word starting with "nano", which we can generically write as "nanoX", can be defined as "nanoscale X". Therefore, unless we define "nanoscale", we cannot therefore properly define nanotechnology. A rational attempt to do so is made in Chapter 2. Here we note that provisionally, the nanoscale is considered to cover the range from 1 to 100 nm. Essentially this is a consensus without a strong rational foundation.

A slightly longer but still succinct definition of nanotechnology is simply "engineering with atomic precision", or "atomically precise technology" (APT). However, this definition does not explicitly include the aspects of "fundamentally new properties" or "novel" and "unique" that nanotechnologists usually insist upon, wishing to exclude existing artifacts that happen to be small. These aspects are encapsulated by the US National Nanotechnology Initiative's declaration that "the essence of nanotechnology is the ability to work at the molecular level, atom-by-atom, to create large structures with fundamentally new molecular organization . . . nanotechnology is concerned with materials and systems whose structures and components exhibit novel and significantly improved physical, chemical, and biological properties, phenomena, and processes due to their nanoscale size" [123].

The US Foresight Institute gives "nanotechnology is a group of emerging technologies in which the structure of matter is controlled at the nanometer scale to produce novel materials and devices that have useful and unique properties".

Function is stressed in: "the design, synthesis, characterization and application of materials, devices and systems that have a functional organization in at least one dimension on the nanometer scale". This is emphasized even more strongly in "nanotechnology pertains to the processing of materials in which structure of a dimension of less than 100 nm is essential to obtain the required functional performance" [36].

In all such definitions, there is the implicit meaning that, as for any technology, the end result must be of practical use. A dictionary definition of nanotechnology is "the design, characterization, production and application of materials, devices and systems by controlling shape and size in the nanoscale" [1]. An alternative definition from the same dictionary is "the deliberate and controlled manipulation, precision placement, measurement, modeling and production of matter in the nanoscale in order to create materials, devices, and systems with fundamentally new properties and functions" [1]. The emphasis on control is particularly important: it is this that distinguishes nanotechnology from chemistry, with which it is often compared. In the latter, motion is essentially uncontrolled and random, within the constraint that it takes place on the potential energy surface of the atoms and molecules under consideration. In order to achieve the desired control, a special, nonrandom *eutactic* environment needs to be available. How eutactic environments can be practically achieved is still being vigorously discussed. Finally, a definition of nanotechnology attempting to be comprehensive is "the application of scientific knowledge to measure, create, pattern, manipulate, utilize or incorporate materials and components in the nanoscale". This underlines the idea of nanotechnology as the consummation of Stage 4 in the sequence of technological revolutions that marks the development of human civilization (Table 12.1 in Chapter 12).

It is sometimes debated whether one should refer to "nanotechnology" or "nanotechnologies". The argument in favor of the latter is that nanotechnology encompasses many distinctly different kinds of technology. But there seems to be no reason not to use "nanotechnology" in a collective sense, since the different kinds are nevertheless all united by striving for control at the atomic scale. Both terms are, in fact, legitimate. When one wishes to emphasize diverse applications, the plural form is appropriate. The singular term refers above all to the mindset or attitude associated with the technology.

1.1.2 Towards a Concept System for Nanotechnology

Objects are perceived or conceived. The properties of an object (which may be common to a set of objects) are abstracted into characteristics. Essential characteristics (feature specifications) typically falling into different categories (e.g., shape, color) are combined as a set to form a concept; this is how objects are abstracted into concepts, and the set of essential characteristics that come together as a unit to form a concept is called the intension. The set of objects abstracted into a concept is called the extension. Delimiting characteristics distinguish one concept from another. Concepts are described in definitions and represented by designations. The set of designations

constitutes the terminology. Concepts are organized into concept systems. A concept system is often called an ontology (which literally means the science of being, but lately is often used in a more restricted sense, namely that of the study of categories).

Figure 1.1 shows (part of) an ontology for nanotechnology. To the right of the diagram one has products—an axis of tangible objects in order of increasing complexity: *materials, devices and systems*. To the left of the diagram one has processes. Note the relationships between *metrology* and *fabrication* (nanomanufacturing, usually abbreviated to nanofacture, also called atomically precise manufacturing, APM) and devices. An atomic force microscope is used to measure nanoscale features; every measuring instrument is necessarily a device; and pushing nano-objects around with a needle is the basis of bottom-to-bottom fabrication (see Section 8.3).

Especially the leaves of the tree might be associated with some ambiguity regarding their extensions. For example, devices can be characterized by the nature of their working medium: electrons, photons, etc. Yet many devices involve more than one medium: for example, nanoelectro*mechanical* devices are being intensively
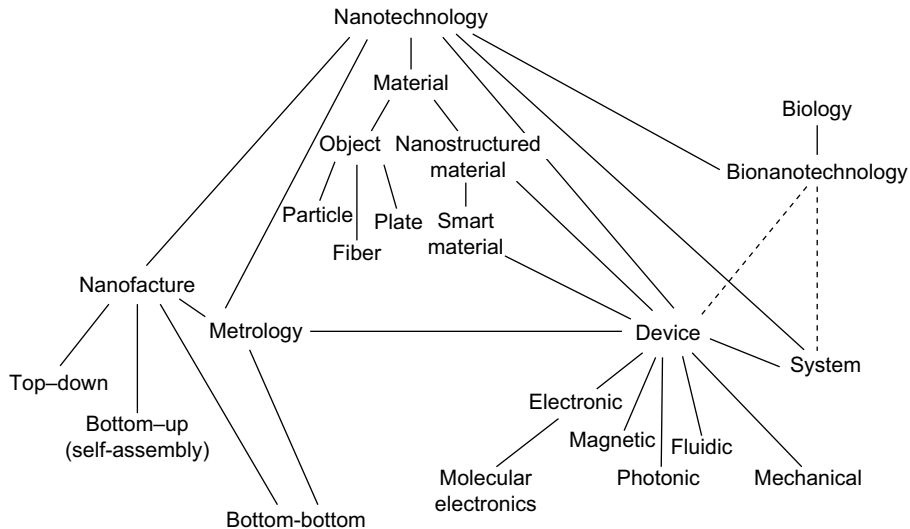
FIGURE 1.1

A concept system (ontology) for nanotechnology. Most of the terms would normally be prefixed by "nano" (e.g., nanometrology, nanodevice). A dashed line merely signifies that if the superordinate concept contributes, then the prefix must indicate that (e.g., bionanodevice, bionanosystem). Biology may also have some input to nanomanufacture (nanofacture), inspiring, especially, self-assembly processes. Not shown on the diagram is what might be called "conceptual nanotechnology", or perhaps better (since it is itself a concept), "virtual nanotechnology", which means the (experimental and theoretical) scrutiny of engineering (and other, including biological) processes at the nanoscale in order to understand them better; that is, the mindset or attitude associated with nanotechnology.

Table 1.1 Some nano concepts and their intensions and extensions		
Intension	**Concept**	**Extension**
One or more external dimensions in the nanoscale	Nano-object	Graphene, fullerene
One or more geometrical features in the nanoscale	Nanomaterial	A nanocomposite
Automaton with information storage and/or processing embodiments in the nanoscale	Nanodevice	Single electron transistor

researched as a way of achieving electronic switching; opto*electronic* control is a popular way of achieving photonic switching; and photochemistry in miniaturized reactors involves both nanophotonics and nanofluidics.

Table 1.1 describes a few of the concepts and their intensions and extensions. At the time of writing, the terminology of nanotechnology is still being intensively debated within national standards organizations as well supranational bodies such as the Comité Européen de Normalisation (CEN) and the International Standards Organization (ISO), hence no attempt has been made to be comprehensive here.

1.2 **AN OSTENSIVE DEFINITION OF NANOTECHNOLOGY**

An ostensive definition of current nanotechnology can be constructed from the most popular topics (Table 1.2), essentially extensions in the sense of Section 1.1.2.

A number of inferences can be drawn from this table, including the preeminence of carbon as a nanomaterial, the nanoparticle as a nano-object, and energy, composites (materials) and sensors as applications. Interestingly, "water" features highly on the list. The reasons for this will be apparent from reading Section 3.8.

A very simple (albeit privative) ostensive definition of nanotechnology is "If you can see it (including with the aid of an optical microscope), it's not nano", referring to the fact that any object below 100 nm in size is below the Abbe limit for optical resolution using any visible wavelength (equation 5.2). A nanoplate, however, would only be invisible if oriented exactly parallel to the line of sight.

1.3 **A BRIEF HISTORY OF NANOTECHNOLOGY**

Reference is often made to a lecture given by Richard Feynman in 1959 at Caltech [56]. Entitled "There's Plenty of Room at the Bottom", it expounds his vision of machines making the components for smaller machines (a familiar enough operation at the macroscale), themselves capable of making the components for yet smaller machines, and simply continuing the sequence until the atomic realm is

Table 1.2 Table of the relative importance (ranked by numbers of occurrences of words in the titles of papers presented at the Nano2009 Conference in Houston, Texas) of nanotechnology terms and applications

Rank	Term	Number of occurrences
1	Carbon, CNT	151
2	Nanoparticle, nanocrystal	138
3	Energy	96
4	(Nano)material	92
5	Nanotube	82
6	(Nano)composite	79
7	(Bio)sensor	55
8	Water	45
9	Device	33
10	Nanowire	33
11	Assembly	31
12	Silicon	30
13	Zinc (oxide)	26
14	Titanium (oxide)	25
15	Quantum	24
16	Silica	21
17	Phage	20
18	Bio	19
19	Photovoltaic	15
20	Nanorod	8
21	Graphene	7
22	Nanopore	7
23	Silver	7

reached. Offering a prize of $1000 for the first person to build a working electric motor with an overall size not exceeding 1/64th of an inch, Feynman was dismayed when not long afterwards a student, William McLellan, presented him with a laboriously hand-assembled (i.e., using the technique of the watchmaker) electric motor of conventional design that nevertheless met the specified criteria.

A similar idea was proposed at around the same time by Marvin Minsky: "Clearly it is possible to have complex machines the size of a flea; probably one can have them the size of bacterial cells ... consider contemporary efforts towards constructing small fast computers. The main line of attack is concentrated on "printing" or evaporation through masks. This is surely attractive; in one operation one can print thousands of elements. But an alternative, equally attractive, has been ignored. Imagine small machines fabricating small elements at kilocycle rates. (The speed of small mechanical devices is extremely high.) Again, one can hope to make thousands of elements per second. But the generality of the mechanical approach is much greater since there are many structures that do not lend themselves easily to laminar mask construction" [118]. One wonders whether Feynman and Minsky had previously

read Robert A. Heinlein's short story "Waldo", which introduces this idea (it was published in the August 1942 issue of "Astounding" magazine under the pseudonym Anson MacDonald).

Here we find the germ of the idea of the assembler, a concept later elaborated by Eric Drexler. The assembler is a universal nanoscale assembling machine, capable not only of making nanostructured materials, but also other machines (including copies of itself). The first assembler would have to be laboriously built atom-by-atom, but once it was working numbers could evidently grow exponentially, and when a large number became available, universal manufacturing capability, hence the nano-era, would have truly arrived (see also Chapter 8).

However, the idea of a minute device intervening at the level of elementary particles was conceived almost a hundred years earlier by James Clerk Maxwell when he conceived his "demon" for selectively allowing molecules to pass through a door, thereby entangling physics with information. Perhaps Maxwell should be considered as the real father of nanotechnology. The demon was described in Maxwell's *Theory of Heat* first published in 1871, but had already been mentioned in earlier correspondence of his.

1.3.1 Ultraprecision Engineering

It could well be said that the history of technological advance is the history of ever finer tolerances in machining metals and other materials. A classic example is the steam engine: James Watt's high-pressure machine that paved the way for the technology to move from a cumbersome and rather inefficient means of pumping water out of mines to an industrially useful and even self-propelling technology was only possible once machining tolerance had improved to enable pistons to slide within cylinders without leaking.

An approach to the nanoscale seemingly quite different from the Heinlein–Feynman–Minsky–Drexler vision of assemblers starts from the microscopic world of precision engineering, progressively scaling down to ultraprecision engineering (Figure 1.2). The word "nanotechnology" was itself coined by Norio Taniguchi in 1974 to describe the lower limit of this process [159]. He referred to "atomic bit machining".

1.3.2 Semiconductor Processing *qua* Microtechnology

The trend in ultraprecision engineering is mirrored by relentless miniaturization in the semiconductor processing industry. The history of the integrated circuit could perhaps be considered to start in 1904, when Bose patented the galena crystal for receiving electromagnetic waves, followed by Picard's 1906 patent for a silicon crystal. The thermionic valve and the triode were invented respectively by Fleming (in 1904) and de Forest (in 1906), which became the basis of logic gates, reaching zenith with the ENIAC, which contained about 20,000 valves. The invention of the point contact transistor in 1947 at Bell Laboratories essentially rendered the thermionic
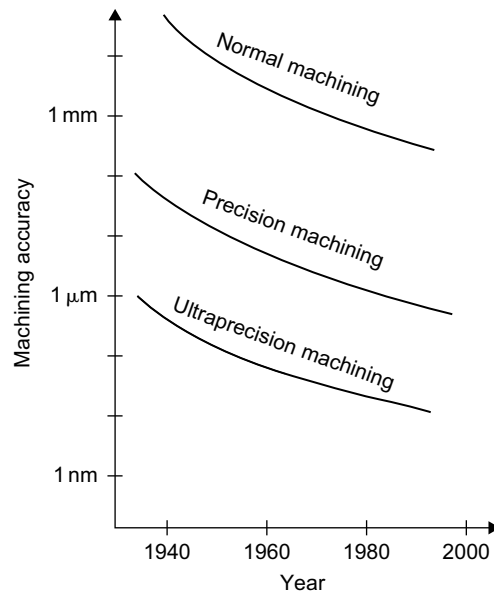
FIGURE 1.2

The evolution of machining accuracy (after Norio Taniguchi).

valve obsolete, but the first commercial use of transistors only occurred in 1953 (the Sonotone 1010 hearing aid), the first transistor radio appearing one year later. Meanwhile the idea of an integrated circuit had been proposed by Dummer at the Royal Signals Research Establishment (RSRE) in 1952, but (presumably) he was not allowed to work on it at what was then a government establishment, and the first actual example was realized by Kilby in 1958 at Texas Instruments, closely followed by Noyce at Fairchild in the following year. It is interesting to recall that the Apollo flight computer ("Block II") used for the first moon landing in 1969 was designed in 1964 (the year before Moore's law was first proposed), used resistor–transistor logic (RTL) and had a clock speed of 2 MHz. Intel introduced the first microprocessor, with about 2000 transistors, in 1971, the year in which the pioneering "LE-120A Handy" pocket calculator was launched in Japan. It took another decade before the IBM personal computer appeared (1981); the Apple II had already been launched in 1977. By 2000, we had the Pentium 4 chip with about 1.2×10^6 transistors fabricated with 180 nm process technology. In contrast, today's dual core Intel Itanium chip has about 1.7×10^9 transistors (occupying an area of about 50×20 mm) with a gate length of 90 nm. A 45 nm transistor can switch 3×10^{11} times per second—this is about 100 GHz. Experimental graphene-based devices achieve more than a THz. Despite the extraordinarily high precision fabrication called for in such devices,

modern integrated circuits are reliable enough for spacecraft (for example) to use commercial off-the-shelf (COTS) devices. The fabrication plants are not cheap—Intel's 2008 China facility is reputed to have cost \$2.5 $\times 10^9$: a mask alone for a chip made using 180 nm process technology costs about \$100,000, rising to one million dollars for 45 nm technology. Despite the huge costs of the plant, cost per chip continues to fall relentlessly: for example, a mobile phone chip cost about \$20 in 1997, but only \$2 in 2007.

The relentless diminution of feature size, and the concomitant increase of the number of transistors that can be fabricated in parallel on a single chip, has been well documented; structures with features a few tens of nanometers in size capable of being examined in an electron microscope were reported as long ago as 1960; device structures with dimensions less than 100 nm were already being reported in 1972, with 25 nm achieved in 1979. Incidentally, the lower size limit for practical semiconductor circuits is considered to be 20 nm; smaller sizes, hence higher transistor number densities per unit area, will only be achievable using three-dimensional design or quantum logic (Section 7.3). Thus we see that since the beginning of nanotechnology—identifying this with the conception of Maxwell's demon—nanotechnology has been intimately connected with information science and technology.

1.3.3 **Nanoparticles**

If we define nanotechnology ostensively, we have to concede a very long history: there is evidence that PbS nanocrystals were the goal of a procedure used since Greco-Roman times to color hair black [164]. Nanoparticulate gold has a long history, not least in medicine (see Section 4.2). The Flemish glassmaker John Utynam was granted a patent in 1449 in England for making stained glass incorporating nanoparticulate gold; and the Swiss medical doctor and chemist von Hohenheim (Paracelsus) prepared and administered gold nanoparticles to patients suffering from certain ailments in the early 16th century; a modern equivalent is perhaps the magnetic nanoparticles proposed for therapeutic purposes. The secret of the extraordinarily advanced metallurgical features of Damascus swords made more than 400 years ago has recently been found to be carbon nanotubes embedded in the blades [146].

With such a long history, it is perhaps hardly surprising that at present they represent almost the only part of nanotechnology with commercial significance. The fabrication of different kinds of nanoparticles by chemical means seems to have been well established by the middle of the 19th century (e.g., Thomas Graham's method for making ferric hydroxide nanoparticles [63]). Wolfgang Ostwald lectured extensively on the topic in the early 20th century in the USA, and wrote up the lectures in what became a hugely successful book, *Die Welt der vernachlässigten Dimensionen*. Many universities had departments of colloid science (sometimes considered as physics, sometimes as chemistry), at least up to the middle of the 20th century, then slowly the subject seemed to fall out of fashion, until its recent revival as part of

nanotechnology. The field somewhat overlapped that of heterogeneous catalysis, in which it was well known, indeed almost obvious, that specific activity (i.e., per unit mass) tended to increase with increasing fineness of division.

1.4 **BIOLOGY AS PARADIGM**

When Feynman delivered his famous lecture [56] it was already well known that the smallest viable unit of life was the cell, which could be less than a micrometer in size. It was surmised that cells contained a great deal of machinery at smaller scales, which has since been abundantly evidenced through the work of molecular biologists. Examples of these machines are molecule carriers (e.g., hemoglobin), enzymes (heterogeneous catalysts), rotary motors (e.g., those powering bacterial flagella), linear motors (e.g., muscle), pumps (e.g., transmembrane ion channels), and multienzyme complexes carrying out more complicated functions then simple reactions (e.g., the proteasome for degrading proteins, or the ribosome for translating information encoded as a nucleic acid sequence into a polypeptide sequence). When Drexler developed his explicit schema of nanoscale assemblers, allusion to nanoscale biological machinery was explicitly made as a "living proof-of-principle" demonstrating the feasibility of artificial devices constructed at a similar scale [43].

At present, probably the most practically useful manifestation of the biological nanoparadigm is self-assembly. In simple form, self-assembly is well known in the nonliving world (for example, crystallization). This process does not, however, have the potential to become a feasible industrial technology for the general-purpose construction of nanoscale objects, because size limitation is not intrinsic to it. Only when highly regular structures need to be produced (e.g., a nanoporous membrane, or a collection of monosized nanoparticles) can the process parameters be set to generate an outcome of a prespecified size. Nature, however, has devised a more sophisticated process, known to engineers as programmable self-assembly, in which every detail of the final structure can be specified in advance by using components that not only interlock in highly specific ways but are also capable of changing their structure upon binding to an assembly partner in order to block or facilitate, respectively, previously possible or impossible interlocking actions. Inspiration for harnessing programmable self-assembly arose from the work of virologists who noticed that pre-assembled components (head, neck, legs) of bacteriophage viruses would further assemble spontaneously into a functional virus merely upon mixing and shaking in a test-tube. This "shake and bake " approach appeared to offer a manufacturing route to nanodevices obviating: (1) the many difficulties involved in making Drexlerian assemblers, which would appear to preclude their realization in the near future; and (2) the great expense of the ultrahigh precision "top–down" approach, whether via UPMT or semiconductor processing. Even if assemblers are ultimately realized, it might be most advantageous to use them to assemble sophisticated "nanoblocks", designed to self-assemble into final macroscopic objects (see Section 8.3.2). In other words, self-assembly's

greatest potential utility will probably arise as a means to bridge the size gap between the nanoscopic products of assemblers and the macroscopic artifacts of practical use for humans. Self-assembly is covered in detail in Chapter 8.

1.5 WHY NANOTECHNOLOGY?

Nanotechnology is associated with at least three distinct advantages:

1. It offers the possibility of creating materials with novel combinations of properties.
2. Devices in the nanoscale need less material to make them, use less energy and other consumables, their function may be enhanced by reducing the characteristic dimensions, and they may have an extended range of accessibility.
3. It offers a universal fabrication technology, the apotheosis of which is the personal nanofactory.

The burgeoning worldwide activity in nanotechnology cannot be explained purely as a rational attempt to exploit "room at the bottom", however. Two other important human motivations are doubtless also playing a role. One is simply "it hasn't been done before"—the motivation of the mountaineer ascending a peak previously untrodden. The other is the perennial desire to "conquer nature". Opportunities for doing so at the familiar macroscopic scale have become very limited, partly because so much has already been done—in Europe, for example, there are hardly any marshes left to drain or rivers left to dam, historically two of the most typical arenas for "conquering nature"—and partly because the deleterious effects of such "conquest" are now far more widely recognized, and the few remaining undrained marshes and undammed rivers are likely nowadays to be legally protected nature reserves. But the world at the bottom, as Feynman picturesquely called it, is uncontrolled and largely unexplored. On a more prosaic note, nanotechnology may already offer immediate benefit for existing products through substitution or incremental improvement (Figure 1.3). The space industry has a constant and heavily pressing requirement for making payloads as small and lightweight as possible. Nanotechnology is ideally suited to this end user—provided the nanomaterials, devices and systems can be made sufficiently reliable (see Chapter 10).

1.5.1 Novel Combinations of Properties

There are two ways of creating nanomaterials. One is by adding nano-objects (nanoadditives) to a matrix. For example, organic polymer matrices incorporating carbon nanotubes can be light and very strong, or transparent and electrically conducting. The other is by fabricating materials *de novo*, atom-by-atom.

Since it is usually more expensive to create nanosized rather than microsized matter, one needs to justify the expense of downscaling the additives: as matter is divided

FIGURE 1.3

Flow chart to determine whether nanotechnology should be introduced into a product.

ever more finely, certain properties become qualitatively different (see Chapter 2). As examples, the optical absorption spectrum of silicon vapor is quite different from that of a silicon crystal, even though the vapor and crystal are chemically identical; when a crystal becomes very small, the melting point falls and there may be a lattice contraction (i.e., the atoms move closer together)—these are well understood consequences of Laplace's law, and may be useful for facilitating a sintering process. If the radius of the crystal is smaller than the Bohr radius of the electron in the bulk solid, the electron is confined and has a higher energy than its bulk counterpart; the optical absorption and fluorescent emission spectra shift to higher energies. Hence, by varying the crystal radius, the optical absorption and emission wavelengths can be tuned. Chemists have long known that heterogeneous catalysts are more active if they are more finely divided. This is a simple consequence of the fact that the reaction takes place at the interface between the solid catalyst and the rest of the reaction medium. Hence, for a given mass the finer the division the greater the surface area. This is

not in itself a qualitative change, although in an industrial application there may be a qualitative transition from an uneconomic to an economic process.

1.5.2 Device Miniaturization: Functional Enhancement

The response time of a device usually decreases with decreasing size. Information carriers have less far to diffuse, or travel ballistically, and the resonant frequency of oscillators increases (Section 2.7).

Clearly the quantity of material constituting a device scales roughly as the cube of its linear dimension. Its energy consumption in operation may scale similarly. In the macroscopic world of mechanical engineering, however, if the material costs are disregarded, it is typically more expensive to make something very small; for example, a watch is more expensive than a clock, for equivalent timekeeping precision. On the other hand when things become very large, as in the case of the clock familiarly known as Big Ben for example, costs again start to rise, because special machinery may be needed to assemble the components, and so on. We shall return to the issue of fabrication in Chapter 8.

Performance (expressed in terms of straightforward input–output relations) *may* thus be enhanced by reducing the size, although the enhancement does not always continue *ad libitum*: for most microelectromechanical systems (MEMS) devices, such as accelerometers, performance is degraded by downscaling below the microscale (Section 10.8), and the actual size of the devices currently mass-produced for actuating automotive airbags already represents a compromise between economy of material, neither taking up too much space nor weighing two much, and still-acceptable performance. On the other hand processor speed of VLSI chips is increased through miniaturization, since components are closer together and electrons have to traverse shorter distances.

Miniaturization may enable new domains of application by enhancing device accessibility. In other words, functionality may be enhanced by reducing the size. A typical example is the cellular (mobile) phone. The concept was developed in the 1950s at Bell Labs, but the necessary circuitry would have occupied a large multi-storey building using the then current thermionic valve technology and, hence, only became practical with large-scale circuit integration. Similarly, it would not be practicable to equip mass-produced automobiles with macroscopic accelerometers with a volume of about one liter and weighing several kilograms.

Functional enhancement also applies to materials. The breaking strain of a mono-lithic material typically increases with diminishing size (see, e.g., Section 2.7). This is usually because, for a given probability of occurrence of defects per unit volume, a smaller volume will inevitably contain fewer defects. This advantage may be countered if the surface itself is a source of defects because of the increased preponderance of surface (Section 2.2).

Fabrication procedures may also be enhanced by miniaturization: any moving parts involved in assembly will be able to operate at much higher frequencies than their macroscopic counterparts. New difficulties are, however, created: noise and

surfaces. The random thermal motion of atoms plays a much more deleteriously influential role than at the macroscale, where it can normally be neglected. The concept of the eutactic environment was introduced partly to cope with this problem. Bottom–up self-assembly, of course, requires noise to drive it; it is amplified up to constructive macroscopic expression by virtue of special features of the components being assembled. The preponderance of surfaces at the nanoscale, which can act as attractors for contamination and so forth, is probably best solved by appropriate energetic engineering of the surfaces (Section 3.2).

1.5.3 A Universal Fabrication Technology

Such a technology is usually considered to be based on nanoscale assemblers (i.e., personal nanofactories). They would enable most artifacts required by humans to be made out of a simple feedstock such as acetylene together with a source of energy (see Section 8.3). Note that there is an intermediate level of technological achievement in which objects are constructed with atomic precision, but without the need for construction equipment to be itself in the nanoscale (e.g., using current tip-based ultramicroscope technology and its derivatives). Nanofabrication (nanomanufacture or nanofacture) represents the ultimate in modularity. Using nanoblocks, purity, the bane of specialty chemical manufacturing, especially pharmaceuticals, is no longer as issue—extraneous molecules are simply ignored by the fabrication system.

1.6 SUMMARY

Nanotechnology is defined in various ways; a selection of already published definitions is given, from which it may be perceived that a reasonable consensus already exists. A more formal concept system is developed, in which care is taken to use the terms consistently. Nanotechnology is also defined ostensively (i.e., what objects already in existence are called "nano"?) and by its history. The role of biology is introduced as providing a living proof-of-principle for the possibility of nanotechnology; this has been of historical importance and continues to provide inspiration. Motivations for nanotechnology are summarized.

1.7 FURTHER READING

A.N. Broers, Limits of thin-film microfabrication, Proc. R. Soc. Lond. A 416 (1988) 1–42.

K.E. Drexler, Engines of Creation, Anchor Books/Doubleday, New York, 1986.

K.E. Drexler, Nanosystems: Molecular Machinery, Manufacturing, and Computation, Wiley-Interscience, 1992.

K. Maruyama, F. Nori, V. Vedral, The physics of Maxwell's demon and information, Rev. Mod. Phys. 81 (2009) 1–23.

W. Ostwald, Die Welt der vernachlässigten Dimensionen, Steinkopff, Dresden, 1914.

R. Zsigmondy, P.A. Thiessen, Das kolloide Gold, Akademische Verlagsgesellschaft, Leipzig, 1925.

The Nanoscale

CHAPTER CONTENTS

Before the advent of nanotechnology, and before the Système International (S.I.) was instituted, the bulk realm was typically referred to as macro and the atomic realm as micro. Scientific consideration of the microworld really began with the invention of the microscope, very appropriately named as it turned out, in view of the subsequent formalization of "micro" as the name of the multiplier 10^{-6}. The Abbe limit of resolution of the optical microscope (equation 5.2) is indeed of the order of one micrometer (the wavelength of visible light). Now that electron microscopes, and even more recently scanning probe microscopes, can resolve features a few nanometers in size, the name "microscope" has become anachronistic—it would be better to call these instruments ultramicroscopes or nanoscopes. The range intermediate between macro and micro was typically referred to as meso ("mesoscale" simply means "of an intermediate scale"). This had no precise definition (except in meteorology), but typically referred to a realm which, while being indubitably invisible (to the naked eye) and hence "micro", and not describable using macroscopic continuum models, could be adequately modeled without explicitly considering what were by then believed to be the ultimate constituents of matter—electrons, protons, neutrons and so forth—but by taking into account an appropriately coarse-grained discreteness. For example, according to this approach a protein might be modeled by considering each amino acid to be a blob with a characteristic chemical functionality according to its unique side chain ("residue"). One might go further and model an amino acid in a protein as a structureless blob with two integers characterizing, respectively, hydrogen bond-donating and hydrogen bond-accepting

capacities; nucleotides ("bases") in DNA (see Section 11.1) are commonly modeled as one of A, C, G and T, with the propensities to pair with, respectively, T, G, C and A. Given that this coarse graining neglects structure below the size of a few nanometers, it might be tempting to simply identify the nanoscale with the mesoscale. Working at the nanoscale would then appear as mainly a matter of computational convenience, since coarse-grained models are much more tractable than more detailed ones (consider Brownian dynamics *versus* molecular dynamics) and often require fewer assumptions. It may well be that a structure characteristic of this scale, such as a lipid bilayer, could be correctly determined from the quantum electrodynamical Hamiltonian, explicitly taking each electron and atomic nucleus into account, but this would be an extraordinarily tedious calculation (and one would need to mobilize the entire computing resources of the world even to reach an approximate solution for any practically useful problem). However, this usage of the term would not imply that there is any particular novelty in the nanoscale, whereas as one makes the transition from typically bulk properties to typically atomic and subatomic properties, one passes through a state that is characteristic neither of individual atoms nor of the bulk, and it would then appear to be reasonable to take the length range over which this transition occurs as the nanoscale.

The current *consensual* definition of the nanoscale is from 1 to 100 nm, although it is still disputed whether these limits should be considered exact or approximate. Finding a consensus is analogous to a multiobjective optimization procedure in which it is not possible to perfectly satisfy every objective considered independently because some of them conflict. A compromise is therefore sought that is as close as possible to all objectives. In many cases there is a continuum of possible states of each attribute (each attribute corresponding to an objective) and the Pareto front on which the compromise solutions lie is indeed a meaningful solution to the optimization problem. If, on the other hand, it is considered that there is only one correct solution for a given objective, and anything else is wrong, the consensus might be Pareto optimal but wrong in every regard—perhaps a little like not a single person in the population having the attributes of Quetelet's "average man". Nevertheless such a consensual definition might still be useful—at least it has the virtue of simplicity and probably does facilitate discussion of nanotechnology, albeit at a rather basic level, among a diverse range of people.

This chapter goes beyond the current consensus and examines the definition of the nanoscale more deeply. According to one of the definitions of nanotechnology discussed in Chapter 1, "nanotechnology pertains to the processing of materials in which structure of a dimension of less than 100 nm is essential to obtain the required functional performance" [36]. This functional performance is implicitly unattainable at larger sizes, hence the definition essentially paraphrases "the deliberate and controlled manipulation, precision placement, measurement, modeling and production of matter in the nanoscale in order to create materials, devices, and systems with fundamentally new properties and functions" [1]. The emergence of a qualitatively new feature as one reduces the size is a key concept associated with nanotechnology. In order to define the nanoscale, therefore, we need to search for such qualitative changes.

Table 2.1	Covalent radii of some atoms		
Element	**Symbol**	**Atomic number**	**Radius/nm [98]**
Carbon	C	6	0.077
Chlorine	Cl	17	0.099
Gold	Au	79	0.150
Hydrogen	H	1	0.031
Silicon	Si	14	0.117
Sodium	Na	11	0.154
Zinc	Zn	30	0.131

Given that in principle the prefix "nano" can apply to any base unit, not just the meter, one needs to justify the primacy of length in a definition of the nanoscale. Among the different categories of attributes—length, mass, voltage, and so on—length does enjoy a certain primacy, perhaps because man first measured things by pacing out their spacial dimension. And given that the meter roughly corresponds to one human step, we can *prima facie* accept that the nanoscale is based on the nanometer. As nanotechnologists seemingly never tire of relating, the prefix "nano" is derived from the Greek "$\nu\alpha\nu o\sigma$", meaning dwarf, and the meaning of "very small" has been formalized by the Système International (S.I.) of scientific units adopting nano as the name of the multiplier 10^{-9}. Thus one nanometer (1 nm) means precisely 10^{-9} m, in other words a millionth of a millimeter or a thousandth of a micrometer, which is somewhat larger than the size of one atom (Table 2.1).

2.1 THE SIZE OF ATOMS

An atom is important because it designates the ultimate (from a terrestrial viewpoint) particles in which matter exists. It was therefore very natural for Richard Feynman to suggest atom-by-atom construction of objects as the lower limit of miniaturization of engineering [56]. It would be highly impracticable to have to start with subatomic particles, such as protons, electrons, neutrons and so forth as building blocks, whereas atomically precise construction, as Feynman rightly perceived and emphasized, is an engineering problem (hence solvable in principle), not a scientific one (requiring the discovery of hitherto unknown laws). Table 2.1 gives the sizes of some atoms.

The scale of the individual atom might be considered as sufficient for a definition of the nanoscale, especially if nanotechnology were to be defined solely in the context of atom-by-atom assembly of objects. But nanotechnology already seems to be much more than this. The definitions of nanotechnology (Chapter 1) emphasize that novel, unique properties emerge at the nanoscale. This implies that merely assembling an otherwise known macro-object atom-by-atom warrants the name of nanotechnology by virtue of the novelty of the assembly process.

2.2 **MOLECULES AND SURFACES**

The dictionary definition of a molecule is typically "the smallest part of a substance that retains all the properties of the substance without losing its chemical identity and is composed of one or more atoms". This combines its etymological meaning as the diminutive of the Latin *moles*, mass (which on its own would make the word essentially synonymous with "particle") with Tyndall's definition as "a group of atoms drawn and held together by what chemists term affinity". This definition is readily applicable to typical covalent molecules such as most organic compounds; a molecule of the carbohydrate called glucose is precisely the particle composed of 6 carbon atoms, 6 oxygen atoms and 12 hydrogen atoms connected in such a way as to make what we know as glucose, and even a single such particle would taste sweet in the characteristic way of glucose, but none of these atoms could be removed without destroying the "glucoseness". Particles of other kinds of substances do not fit the definition so well. A single atom of a metal such as gold, although chemically gold, has a different optical absorption spectrum from that of bulk gold, and the same applies to numerous binary semiconducting compounds such as cadmium sulfide, CdS, which can be prepared as a vapor containing isolated CdS molecules (in the chemical sense). In bulk material, the individual atoms are close enough for the wave functions of their electrons to overlap, but to satisfy Pauli's exclusion principle, their energies must be slightly shifted, forming a band of states instead of a discrete level as in the isolated atom or molecule; see also Section 2.5.

The surface of a particle is qualitatively different from the bulk because it is less connected; the smaller the radius, the greater the proportion of underconnected atoms. Consider a "supersphere", a spherical aggregate of spheres, which we can take to be atoms (Figure 2.1). By simple geometric considerations, only one of the 19 atoms is not in contact with the surface. If the radii of the atom and the supersphere are r and R respectively, then the proportion of atoms in the shell must be $1 - [(R - r)/R]^3$. The mean connectivity, and hence cohesive energy, should vary inversely with a fractional power of this quantity. If R is expressed in units of r, then the surface to volume ratio is equal to $3r/R$: in other words, if $R = 3r$, equal numbers of atoms are in the bulk and at the surface. The nanoparticle is, therefore, one that is "all surface". As a result, the melting point T_m of small spheres is lowered relative to that of bulk material, according to

$$T_m/T_\infty = 1 - (C/R)^n \tag{2.1}$$

where C is a constant, and with the exponent $n = 1$ for metals as shown by experimental studies. In fact, melting is a rather complex phenomenon and if an approach such as that of Lindemann's criterion is adopted, due account of the difference between the surface and the bulk atoms might be important in determining the temperature of some practical process, such as sintering. Furthermore, it should be noted that if a nanoscale thin film is investigated, rather than particles (i.e., a nanoplate), there is no depression of the melting point relative to the bulk [89].
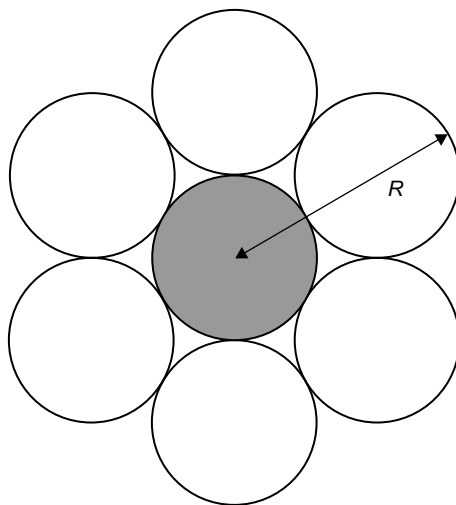
FIGURE 2.1

Cross-section of a spherical nanoparticle consisting of 19 atoms.

As Wulff has pointed out, for a crystal growing in equilibrium with its vapor or solution, the ratio of the surface tension γ of the growing phase to the distance from the center r should be constant. If the mechanical effect of the surface tension can be reduced to an isotropic pressure, then we have the Laplace law

$$\Delta P = 2\gamma/r \tag{2.2}$$

where ΔP is the pressure exerted on the crystal compared with zero pressure if the growing phase is an infinite plane. This pressure can do work on the lattice, compressing the crystal. If β is the isothermal compressibility coefficient, then

$$\Delta V/V = -\beta\Delta P = 3\Delta l/l \tag{2.3}$$

where V is the volume of the unperturbed crystal, and l a characteristic lattice parameter of the unperturbed (bulk) crystal. Conversely, the vapor pressure P of a particle should be inversely dependent on its radius R, as expressed by the Gibbs-Thomson (or Kelvin) law:

$$P(R)/P_\infty = e^{2\gamma v/(Rk_{\mathrm{B}}T)} \tag{2.4}$$

where P_∞ is the vapor pressure of a surface of infinite radius (i.e., flat), γ and v are, respectively, the surface tension (note that once surface tension is introduced, it should also be borne in mind that surface tension is itself curvature-dependent [152]) and molecular volume of the material, and k_{B} and T are, respectively, Boltzmann's constant and the absolute temperature. Equation (2.4) can be used to compare the

chemical potentials μ of two spheres of radii R_1 and R_2:

$$\mu_2 - \mu_1 = 2\gamma V \left(\frac{1}{R_2} - \frac{1}{R_1} \right) \qquad (2.5)$$

where V is the molar volume.

These equations do not predict an abrupt discontinuity at the nanoscale. To be sure, there are huge differences between the bulk and the nanoscale (e.g., melting temperature lowered by tens of degrees), but these "unique" properties approach bulk values asymptotically, and the detectable difference therefore depends on the sensitivity of the measuring apparatus, and hence is not useful for defining a nanoscale. On the other hand, sometimes there is evidence for qualitative change with diminishing size; as shown, for example, by the appearance of anomalous crystal structures with no bulk equivalent [137].

2.3 **NUCLEATION**

When a vapor is cooled it will ultimately condense, but if it is very pure it may be greatly supercooled because the appearance of a new phase—in this case, a liquid or a solid—is necessarily a discontinuous process and therefore requires a finite nucleus of the new phase to be formed via random fluctuations. At first sight this seems to be highly improbable, because the formation of the nucleus requires the energetically unfavorable creation of an interface between the vapor and the new condensed phase—indeed, were this not energetically unfavorable the vapor could never be stable. However, if the nucleus is big enough, the energetically favorable contribution of its nonsurface exceeds the surface contribution, and not only is the nucleus stable, but also grows further, in principle to an unlimited extent (see Figure 2.2). More
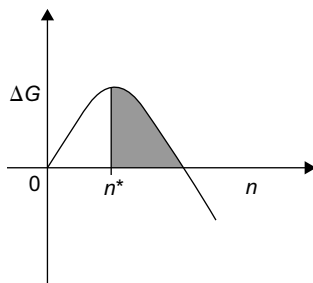
FIGURE 2.2

Sketch of the variation of free energy of a cluster containing n atoms (proportional to the cube of the radius). The maximum corresponds to the critical nucleus size n^*. Clusters that have managed through fluctuations to climb up the free energy slope to reach the critical nucleus size have an equal probability to shrink back and vanish, or to grow up to microscopic size. The shaded zone corresponds to a possible nanoscale (see text).

precisely, when atoms cluster together to form the new phase, they begin to create an interface between themselves and their surrounding medium, which costs energy $A\gamma$, where A is the area of the cluster's surface, equal to $(4\pi)^{1/3}(3nv)^{2/3}$, where n is the number of atoms in the cluster. At the same time each atom contributes to the (negative) cohesive energy of the new phase, equal to $nv\Delta G$, where ΔG is the difference in the free energies (per unit volume) between the uncondensed and condensed phases, the latter obviously having the lower energy. Summing these two contributions, at first the energy will increase with increasing n, but ultimately the (negative) cohesive energy of the bulk will win. Differentiating with respect to n and looking for maxima yields

$$n^* = -\frac{32\pi\gamma^3}{3v(\Delta G)^3}. \tag{2.6}$$

The critical nucleation size can lay reasonable claim to be considered as the boundary of the nanoscale from the mechano-chemical viewpoint. At least it represents a discontinuity *qua* crossover point. The critical nucleus size represents the boundary between "surface dominated" (an unstable state) and "bulk dominated" (a stable state).

A corollary of this definition is that nanotechnology is strongly associated with surface technology, since the nanoscale object is necessarily one in which all or almost all of its constituent atoms have the properties of the surface atoms of a bulk piece of what is atomically the same material.

Note that the nucleus, small as it is, has been characterized above using a thermodynamic continuum approach that is perfectly macroscopic in spirit; in other words, it is not apparently necessary to explicitly take atomistic behavior into account to provide a good description of the phenomena. The excellent agreement between the projections of equations based on (2.7) and experimental results on photographic latent images constituted from only a few silver atoms caused mild surprise when first noticed by Peter Hillson working at the Kodak research laboratories in Harrow, UK. Molecular dynamics simulations also suggest that the continuum approach seems to be valid down to scales of a few atomic radii.

2.4 CHEMICAL REACTIVITY

Consider a heterogeneous reaction $A+B \rightarrow C$, where A is a gas or a substance dissolved in a liquid and B is a solid. Only the surface atoms are able to come into contact with the environment, hence for a given mass of material B the more finely it is divided the more reactive it will be, in terms of numbers of C produced per unit time.

The above considerations do not imply any discontinuous change upon reaching the nanoscale. Granted, however, that matter is made up of atoms, the atoms situated at the boundary of an object are qualitatively different from those in the bulk
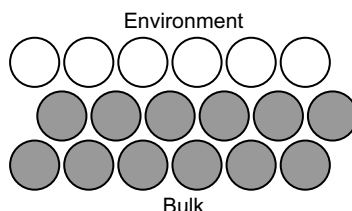
FIGURE 2.3

The boundary of an object shown as a cross-section in two dimensions. The surface atoms (white) are qualitatively different from the bulk atoms (gray), since the latter have six nearest neighbors (in two-dimensional cross-section) of their own kind, whereas the former only have four.

(Figure 2.3). A cluster of six atoms (in two-dimensional Flatland) has only one bulk atom (see Figure 2.1), and any smaller cluster is "all surface". This may have a direct impact on chemical reactivity: it is to be expected that the surface atoms are individually more reactive than their bulk neighbors, since they have some free valences (i.e., bonding possibilities). Consideration of chemical reactivity (its enhancement for a given mass, by dividing matter into nanoscale-sized pieces) therefore suggests a discontinuous change when matter becomes "all surface".

Another implication concerns solubility: the vapor pressure P of a droplet, and by extension the solubility of a nanoparticle, increases with diminishing radius r according to the Kelvin equation (cf. equation 2.4)

$$k_{\mathrm{B}}T\ln(P/P_\infty) = 2\gamma v/r \qquad (2.7)$$

In practice, however, the surface atoms may have already satisfied their bonding requirements by picking up reaction partners from the environment. For example, many metals become spontaneously coated with a film of their oxide when left standing in air, and as a result are chemically more inert than pure material. These films are typically thicker than one atomic layer, implying complex growth processes involving internal transport or diffusion. For example, on silicon the native oxide layer is about 4 nm thick. This implies that a piece of freshly cleaved silicon undergoes some lattice disruption enabling oxygen atoms to effectively penetrate deeper than the topmost layer. Thus it may happen that if the object is placed in the "wrong" environment, the surface compound may be so stable that the nanoparticles coated with it are actually less reactive than the same mass of bulk matter. A one centimeter cube of sodium taken from its protective fluid (naphtha) and thrown into a pool of water will act in a lively fashion for some time, but if the sodium were first cut up into one micrometer cubes, most of the metallic sodium would have already reacted with moist air before reaching the water.

Consider the prototypical homogeneous reaction

$$\mathrm{A} + \mathrm{B} \to \mathrm{C} \qquad (2.8)$$

and suppose that the reaction rate coefficient k_f is much less than the diffusion-limited rate, that is, $k_f \ll 4\pi(d_A + d_B)(D_A + D_B)$, where d and D are the molecular radii and diffusivities respectively. Then [147]:

$$\frac{dc}{dt} = k_f[\langle a \rangle \langle b \rangle + \Delta^2(\gamma_t)] = k_f \langle ab \rangle \tag{2.9}$$

where a and b are the numbers (concentrations) of A and B, the angular brackets denote expected numbers, and γ_t is the number of C molecules created up to time t. The term $\Delta^2(\gamma_t)$ expresses the fluctuations in γ_t: $\langle \gamma_t^2 \rangle = \langle \gamma_t \rangle^2 + \Delta^2(\gamma_t)$: supposing that γ_t approximates to a Poisson distribution, then $\Delta^2(\gamma_t)$ will be of the same order of magnitude as $\langle \gamma_t \rangle$. The kinetic mass action law (KMAL) putting $\langle a \rangle = a_0 - c(t)$ etc., the subscript 0 denoting initial concentration at $t = 0$, is a first approximation in which $\Delta^2(\gamma_t)$ is supposed negligibly small compared to $\langle a \rangle$ and $\langle b \rangle$, implying that $\langle a \rangle \langle b \rangle = \langle ab \rangle$, whereas strictly speaking it is not since a and b are not independent. Nevertheless, the neglect of $\Delta^2(\gamma_t)$ is justified for molar quantities of starting reagents (except near the end of the process, when $\langle a \rangle$ and $\langle b \rangle$ become very small), but not for reactions in ultrasmall volumes (nanomixers).

These number fluctuations, i.e. the $\Delta^2(\gamma_t)$ term, constantly tend to be eliminated by diffusion. On the other hand, because of the correlation between a and b, initial inhomogeneities in their spacial densities may (depending on the relative rates of diffusion and reaction) lead to the development of zones enriched in either one or the other faster than the enrichment can be eliminated by diffusion. Hence instead of A disappearing as t^{-1} (when $a_0 = b_0$), it is consumed as $t^{-3/4}$, and in the case of a reversible reaction, equilibrium is approached as $t^{-3/2}$. Deviations from perfect mixing are more pronounced in dimensions lower than three.

As the reaction volume decreases, the relative importance of fluctuations in the concentrations of the various chemical species increases, but this does not in itself constitute a qualitative change. In cases more complicated than the single reaction $A + B \rightarrow C$, however, in which additional parallel or sequential reactions are possible, the outcome may actually change. For example, consider what happens if an additional reaction $A + C \rightarrow D$ is possible. In the usual bulk situation A-rich regions will rapidly become coated by C-rich regions, which in turn will further react with A to form D. If, however, C gets diluted in the ocean of B before it can undergo further reaction with A, D will be formed in very poor yield (see Figure 2.4). Ways of achieving such rapid dilution could be to somehow divide the reaction space into very small volumes, or to impose a drastically violent mixing regime on the bulk, although friction with the walls of the containing vessel prevent turbulence from filling the entire volume.

Since nanotechnology is above all the technology of surfaces, it is natural to favor reactions taking place at surfaces, not the heterogeneous reactions in which one agent in a bulk three-dimensional space impinges on and reacts with another at a surface (as in electrochemistry), but the situation in which both reagents are confined to a two-dimensional surface. Merely as a consequence of having fewer directional possibilities to move (cf. Pólya's theorem), a reaction of type (2.8) will be significantly accelerated (see [140] for an experimental demonstration).
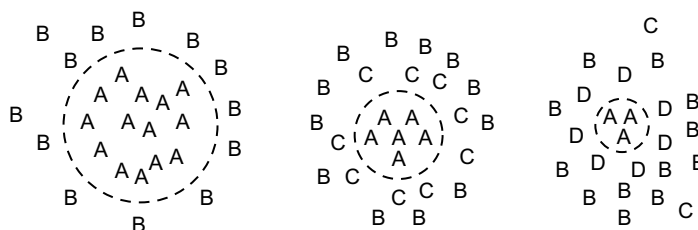
FIGURE 2.4

Illustration (cross-section) of a possible result of an initial mixture of chemical species A and B (on the left) when reactions A + B → C (in the middle) and A + C → D (on the right) are possible. Further explanation is given in the text.

2.5 ELECTRONIC AND OPTICAL PROPERTIES

Electronic Energy Levels. Individual atoms have discrete energy levels and their absorption spectra correspondingly feature sharp individual lines. It is a well-known feature of condensed matter that these discrete levels merge into bands, and the possible emergence of a forbidden zone (band gap) determines whether we have a metal or a dielectric.

Stacking objects with nanoscale sizes in one, two or three dimensions (yielding nanoplates, nanofibers and nanoparticles with, respectively, confinement of carriers in two, one or zero dimensions) creates a new class of superlattices or artificial atoms. These are exploited in a variety of nanodevices (Chapter 7). The superlattice gives rise to sub-bands with energies

$$E_n(k) = E_n^{(0)} + \hbar^2 k^2/(2m^*) \qquad (2.10)$$

where $E_n^{(0)}$ is the nth energy level, k the wavenumber, and m^* the effective mass of the electron, which depends on the band structure of the material.

Similar phenomena occur in optics, but since the characteristic size of photonic band crystals is in the micrometer range they are, strictly speaking, beyond the scope of nanotechnology.

Quantum Confinement. The general principle is that confinement occurs if a characteristic size (e.g., the thickness of a plate) is less than or equal to the electron coherence length. The shifting optical properties of very small dielectric particles (shift of the fundamental adsorption edge, disappearance of exciton absorption peaks) were first investigated by Berry [16, 17]. Later, Efros and Efros [47], Brus [27] and Banyai and Koch [13] provided explanations in terms of the confinement of charge carriers (electrons, defect electrons or "positive holes", and excitons). In this case, there is a clear criterion for confinement effects to be observable: the actual radius

of the particle (assumed spherical) must be less than the charge carrier's Bohr radius r_B (i.e., the smallest possible orbit for the (ground state) electron in hydrogen). Its magnitude is given by the expression

$$r_B = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} \tag{2.11}$$

where $m_e = 0.91 \times 10^{-30}$ kg is the mass of the electron. The Bohr radius of the free electron in vacuum is therefore 53 pm. In a semiconductor this formula can still be used, but ϵ_0 must be replaced by the actual dielectric constant of the material ϵ_s, and m_e by the effective mass m_{eff}, which depends on the band structure (it is defined as $m_{eff} = \hbar^2/[d^2E/dk^2]$), yielding (for the ground state)

$$r_B^* = \frac{4\pi\epsilon_s\hbar^2}{m_{eff}e^2} \tag{2.12}$$

This is the natural measure of the nanoscale for electronic properties, especially those involving adsorption and emission of light involving electronic transitions between the valence band and the conduction band. Typical values of r_B^* range from a few to a few hundred nanometers. Therefore, it is practically possible to create particles whose radius r is smaller than the Bohr radius. In this case the energy levels of the electrons increase (a similar argument applies to defect electrons, i.e., positive holes), and the greater the degree of confinement, the greater the increase. Hence the band edge of optical adsorption (and band-edge luminescent emission) shifts towards the blue end of the spectrum with decreasing r for $r < r_B$. This is sometimes called a quantum size effect in the scientific literature, and nanoparticles with this property are called quantum dots.

Since r_B varies greatly for different materials (and in some cases the electron and the hole have very different radii) it follows that the nanoscale (considering electronic properties) is material-dependent (see Table 2.2). As pointed out in Section 1.1, until now the nanoscale is often, albeit provisionally, taken as the range between 1 and 100 nm, according to which a nanoparticle would be any particle with a diameter falling in this range. However, a 50 nm CdTe particle would have the optical properties of bulk CdTe, and would not therefore rank as a quantum dot, nor *a fortiori* as a nanoparticle with respect to optical properties since they would not be novel, whereas InSb of the same size would.

An even simpler criterion for the optical nanoscale would be simply "invisibility". This could be determined by the Abbe resolution limit (equation 5.2), or the Rayleigh ratio $R(\theta)$ for light of wavelength λ scattered by particles of radius r:

$$R(\theta) = \frac{I_\theta d_s^2}{I_0 f V} = \frac{8\pi^4 r^2}{\lambda^4} / (1 + \cos^2\theta) \tag{2.13}$$

Table 2.2 Bohr radii and related parameters of some semiconductors [158]

Material	bandgap/eV[a]	ϵ_s/ϵ_0	$m_{\text{eff}}^{\text{electron}}/m_e$	$m_{\text{eff}}^{\text{hole}}/m_e$	r_B/nm[b]
CdS	2.42	5.4	0.21	0.8	7.5
InSb	0.17	17.7	0.0145	0.40	60
InP	1.35	12.4	0.077	0.64	11
Si	1.12	11.9	0.98	0.16	2.5
ZnO	3.35	9.0	0.27	–	1.7

[a] At 300 K.
[b] Of the electron; see equation (2.12); see also [172].

for unpolarized light, where I_0 is the incident irradiance, I_θ is the total intensity of scattered radiation observed at an angle θ and a distance d_s from the volume of scattering V, and the factor f accounts for polarization phenomena. The particles would be deemed to be in the nanoscale if a suspension of them showed no perceptible turbidity.

In order to establish quantitatively how the band-edge luminescent emission varies with particle diameter, consider a one-electron, one-dimensional time-dependent Schrödinger equation with energy eigenvalues E

$$\left[-\frac{\hbar^2}{2m_e}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x) \right]\Psi = E\Psi \tag{2.14}$$

with a potential $V(x) = 0$ for $0 \leq x \leq 2r$ (within the particle) and $V(x) = \infty$ for $x < 0$ and $x > 2r$. Using a trial wavefunction $\Psi(x) = \sin ax$ for the electron, within the particle we can solve to find $a^2 = 2m_eE/\hbar^2$. Ψ has to vanish at the particle boundaries, which can happen if the phase ax is an integer multiple of π, $ax = n\pi$, whence

$$E_n = \frac{\hbar^2 n^2 \pi^2}{8m_e r^2}, \quad n = 0, 1, \ldots; \tag{2.15}$$

we can call n the principle quantum number. Note that we have not specified exactly what value of m_e we should use in equation (2.14). It is an open question whether the effective mass of the electron is the same in a very small particle, already known to be subject to lattice contraction and possibly other distortions of the crystal structure (Section 2.2). The experimental verification of these formulas is subjected to a number of difficulties. Firstly it is quite difficult to prepare monodisperse quantum dots. The Stranski–Krastanov mechanism has been made use of for the preparation of monodisperse GaInAl dots for lasers via semiconductor processing technology (see Section 8.1.2), but most quantum dots are presently made by a chemical method, by reacting two soluble precursors together in solution (Section 6.1.2), with which it is difficult to get particle size distributions with a coefficient of variation much better than 10%. The surface atoms may have different electronic properties from the bulk by virtue of their lower coordination number. This notion was developed by Berry

[16, 17] as the main reason for the difference between bulk and nanoparticle optical properties. Particles in contact with another condensed phase, such as a suspending liquid, are likely to have additional electron levels caused by adsorbed impurities; it is extremely unlikely that the particle boundaries correspond to the electron potential jumping abruptly from zero to infinity.

Jellium. Quantum dots are dielectric nanoparticles, typically I–VII, II–VI or III–V semiconductors. Metal clusters can be modeled as jellium [96]. Each metal atom contributes a characteristic number ν of electrons to the free electron gas (a corollary of which is that the actual positions of the metal nuclei can be neglected) filling a uniform charged sphere corresponding to the actual cluster containing N metal atoms. The electronic structure is derived by solving the Schrödinger equation for an electron constrained to move within the cluster sphere under the influence of an attractive mean field potential due to the partially ionized metal atoms. The solution yields energy levels organized in shells and subshells with quantum numbers n and l respectively as with atoms. The order of the energy levels in jellium is a little different, however: $1s^2$, $1p^6$, $1d^{10}$, $2s^2$, $1f^{14}$, $2p^6$, etc.—hence even the first valence shell contains s, p, d and f orbitals. The degeneracy of each level is $2(2l+1)$.

A triumph of the jellium model was the correct prediction of "magic" numbers N_m—corresponding to clusters of exceptional stability. These are clusters with completely filled shells, and the N_m are the values of N fulfilling the condition

$$\sum^{n,l} 2(2l+1) = n\nu - q \tag{2.16}$$

where q is the charge of the cluster. Hence (for example) uncharged ($q = 0$)Cs ($\nu = 1$) clusters will have $N_m = 2, 8, 18, 20, 34, 40, 58, \ldots$; Al_{13}^- ($q = 3$) will be stable.

Superatoms. The electron affinity of Al_{13} is comparable to that of a chlorine atom and KAl_{13} is an ionically bound molecule analogous to KCl. Replacing one Al atom with C ($\nu = 4$) results in an inert species comparable to argon. An extended periodic table can be constructed from superatoms, defined as clusters that are energetically and chemically stable [34]. These entities are useful as nanoblocks for assembly into desirable structures (Section 8.3.2).

2.6 MAGNETIC AND FERROELECTRIC PROPERTIES

Ferromagnetism. In certain elements (e.g., Fe, Co, Ni, Gd) and alloys, exchange interactions between the electrons of neighboring ions lead to a very large coupling between their spins such that, above a certain temperature, the spins spontaneously align with each other. The proliferation of routes for synthesizing nanoparticles of ferromagnetic substances has led to the discovery that when the particles are below a certain size, typically a few tens of nanometers, the substance still has a large magnetic susceptibility in the presence of an external field but lacks the

remanent magnetism characteristic of ferromagnetism. This phenomenon is known as superparamagnetism.

Below a certain critical size r_s, the intrinsic spontaneous magnetization of a ferromagnetic material such as iron results in a single domain with all spins in the same direction; above that size the material is divided into domains with different spin directions. Kittel has given a formula to estimate the critical single domain size (diameter) in zero magnetic field:

$$d_s = 1.43\gamma_m/J_s^2 \tag{2.17}$$

where

$$\gamma_m = 4\sqrt{EK} \tag{2.18}$$

is the domain wall surface energy density, J_s the spontaneous magnetization, E the exchange energy and K the anisotropy constant. Typical values of d_s are 30 nm for Co, 50 nm for Fe, and 70 nm for iron oxides (hematite, Fe_2O_3, and magnetite, Fe_3O_4). These sizes could reasonably be taken to be the upper limit of the nanorealm, when magnetism is being considered. Incidentally, the magnitude of the Bohr magneton, the "quantum" of magnetism, is given by $\beta = e\hbar/(2m_e)$.

Within a single domain of volume V, the relaxation time (i.e., the characteristic time taken by a system to return to equilibrium after a disturbance) is

$$\tau = \tau_0 \exp(KV/k_BT) \tag{2.19}$$

where τ_0 is a constant ($\sim 10^{-11}$ s). Particles are called superparamagnetic if τ is shorter than can be measured; we then still have the high susceptibility of the material, but no remanence. In other words, ferromagnetism becomes zero below the Curie temperature. It is a direct consequence of the mean coordination number of each atom diminishing with diminishing size because there are proportionately more surface atoms, which have fewer neighbors. There is thus a lower limit to the size of the magnetic elements in nanostructured magnetic materials for data storage, typically about 20 nm, below which room temperature thermal energy overcomes the magnetostatic energy of the element, resulting in zero hysteresis and the consequent inability to store magnetization orientation information.

Since the relaxation time varies continuously with size (note that K may also be size-dependent), it is unsuitable for determining the nanoscale; on the other hand r_s could be used thus. It is known that the shape of a particle also affects its magnetization [132]. A similar set of phenomena occurs with ferroelectricity [31].

2.7 **MECHANICAL PROPERTIES**

The ultimate tensile strength σ_{ult} of a material can be estimated by using Hooke's law, *ut tensio, sic vis*, to calculate the stress required to separate two layers of atoms from one another, by equating the strain energy (the product of strain, stress and volume)

with the energy of the new surfaces created, yielding

$$\sigma_{\text{ult}} = (2\gamma Y/x)^{1/2} \tag{2.20}$$

where γ is the surface energy (Section 3.2), Y is Young's modulus (i.e., stiffness) and x is the interatomic spacing (i.e., twice the atomic radius, see Table 2.1). This is basically an idea of Dupré. A.A. Griffith proposed that the huge discrepancies between this theoretical maximum and actual greatest attainable tensile strengths were due to the nucleation of cracks, which spread and caused failure. Hence strength not only means stiffness, but also must include the concept of toughness that means resistance to crack propagation. By considering the elastic energy W_σ stored in a crack of length $2l$, namely

$$W_\sigma = \pi l^2 \sigma^2 / Y \tag{2.21}$$

where σ is the stress, and its surface energy

$$W_\gamma = 4l\gamma \tag{2.22}$$

(ignoring any local deformation of the material around the crack), a critical *Griffith length* can be defined when $\partial W_\sigma / \partial l$ equals $\partial W_\gamma / \partial l$:

$$l_G = 2\gamma Y/(\pi\sigma^2); \tag{2.23}$$

if $l < l_G$ the crack will tend to disappear, but if $l > l_G$ then it will rapidly grow. Note the existence of accompanying irreversible processes that accelerate the process, typically leading to microscopic failure, quite possibly with disastrous consequences. By putting $l_G = x$ into formula (2.23) one recovers the estimate of ultimate tensile strength, equation (2.20) (neglecting numerical factors of order unity).

If σ is taken to be a typical upper limit of the stress to which a structural component will be subjected in use, one finds values of l_G in the range of nanometers to tens of nanometers. Therefore, as far as mechanical properties of materials are concerned, l_G provides an excellent measure of the nanoscale. If the size of an object is less than l_G, it will have the theoretical strength of a perfect crystal. If it is greater, the maximum tensile strength will diminish $\sim l^{-1}$ (cf. equation 2.23), a result that was demonstrated experimentally by Griffith using glass fibers and by Gordon and others using whiskers of other materials such as silicon and zinc oxide.

In reality, equation (2.23) may somewhat underestimate the critical length because of the neglect of material deformation, which causes the work of fracture to exceed that estimated from the surface tension alone. Griffith's criterion is based on Dupré's idea, which completely ignores this deformation. Hence Bikerman [19] has proposed an alternative hypothesis, that a crack propagates when the decrease of strain energy around a critical domain exceeds the increase required to elastically deform this domain to the breaking point.

One consequence of these ideas is the impossibility of comminuting particles below a certain critical size d_{crit} by crushing. Kendall [90] derives, in a rather

simple way,

$$d_{\text{crit}} = 32YF/(3\sigma^2) \tag{2.24}$$

where F is the fracture energy; below this size crack propagation becomes impossible and the material becomes ductile.

Cracks can be arrested by interfaces, and it has long been realized that nature has solved the problem of creating strong (i.e., both stiff and tough) structural materials such as bones, teeth and shells by assembling composite materials consisting of mineral platelets embedded in protein matrix. The numerous interfaces prevent the propagation of cracks, and provided the stiff components of the composite are not bigger than the critical Griffith length, optimal strength should be obtainable. This has recently been subjected to more careful quantitative scrutiny [87] yielding an expression for the stiffness of the composite:

$$\frac{1}{Y} = \frac{4(1-\theta)}{G_p\theta^2 a^2} + \frac{1}{\theta Y_m} \tag{2.25}$$

where G_p and Y_m are respectively the shear modulus of the protein and Young's modulus of the mineral, θ is the volume fraction of mineral, and a is the aspect ratio of the platelets. It seems that the thickness of the platelets in natural materials (such as various kinds of seashells) roughly corresponds to l_G. By assuming that the protein and the mineral fail at the same time, an expression for the optimal aspect ratio of the mineral platelets can be derived, equal to $\sqrt{\pi Y_m \gamma / l}/\tau_p$, where τ is the shear strength of the protein matrix.

2.8 QUANTUM SMALLNESS

In the preceding sections of this chapter, the various crossover lengths that demarcate the nanorealm from other realms have emerged as upper limits of the nanorealm; above these lengths, the behavior is qualitatively indistinguishable from that of indefinitely large objects, and below them novel properties are observable. But does the nanorealm have a lower limit? This has been implicit in, for example, the discussion of the mesoscale at the beginning of this chapter, implying that the point where continuum mechanics breaks down and atoms have to be explicitly simulated demarcates the lower limit of the nanoscale. On the other hand, discussion about the possible design of nanoscale assemblers can hardly avoid the explicit consideration of individual atoms. Despite such challenges, it is probably fair to say that there has been considerably less discussion about the lower limit of nanotechnology compared with the upper one.

The idea that nanotechnology embodies a qualitative difference occurring at a certain point when one has shrunk things sufficiently links nanotechnology to Hegel, who first formulated the idea that a quantitative difference, if sufficiently large, can become a qualitative one in his *Wissenschaft der Logik*. In essence nanotechnology

is asserting that "less is different". In order to defend condensed matter physics from the criticism that it is somehow less fundamental than elementary particle physics, P.W. Anderson wrote a paper entitled "More is different" [7], in which he developed the idea of qualitative changes emerging when a sufficient number of particles are aggregated together (in other words chemistry is not simply applied physics, nor biology simply applied chemistry, but at each new level, implying simultaneous consideration of a larger number of particles than in the preceding one, new phenomena emerge, the existence of which could not have been predicted even from complete knowledge of the preceding level). Nanotechnology is a further example of the fundamental idea of the emergence of wholly new phenomena at a new level of description, except that now we are not *increasing* the quantity, but *decreasing* it. The various examples gathered in the preceding sections, such as the optical properties of semiconductor nanoparticles already mentioned, or the magnetic properties of nanoparticles, are all concrete manifestations of Hegel's idea of a qualitative change emerging when a quantitative change becomes sufficiently large.

There is, however, an even more fundamental difference that appears when we enter the quantum realm—the world of the absolutely small, as emphasized by Dirac: "It is usually assumed that, by being careful, we may cut down the disturbance accompanying our observations to any desired extent. The concepts of big and small are then purely relative and refer to the gentleness of our means of observation as well as to the object being described. In order to give an absolute meaning to size, such as is required for any theory of the ultimate structure of matter, we have to assume that there is a limit to the fineness of our powers of observation and the smallness of the accompanying disturbance—a limit which is inherent in the nature of things and can never be surpassed by improved technique or increased skill on the part of the observer. If the object under observation is such that the unavoidable limiting disturbance is negligible, then the object is big in the absolute sense and we may apply classical mechanics to it. If, on the other hand, the limiting disturbance is not negligible, then the object is small in the absolute sense and we require a new theory for dealing with it" [42].

Another way in which the quantum realm is absolutely different from the classical one was elaborated upon by Weisskopf [165]. In classical physics, laws predetermine only the general character of phenomena; the laws admit a continuous variety of realizations, and specific features are determined by the initial conditions. On the other hand, in the quantum realm individual atoms have well-defined specific qualities. Furthermore their identity is unique and immutable. "Two pieces of gold, mined at two different locations and treated in very different ways, cannot be distinguished from one another. All the properties of each individual gold atom are fixed and completely independent of its previous history" [165] (gold has only one stable isotope, it should be noted; this is not the case for all other elements, in which case the isotopic ratios could give clues to the provenance of the samples). Similarly, all electrons have the same charge, spin and mass. The same identity extends to the crystal structure of a substance (disregarding polymorphism), and indeed to other properties. This existence of well-defined specific qualities is alien to the spirit of classical physics.

Furthermore, these quantum "blocks" (atoms, electrons, etc.) are embodied in larger structures assembled from atoms, such as chemical molecules, which also possess quality, specificity and individuality.

The divisibility of *process* is also a main feature of classical physics [165]: every process can be considered as a succession of partial processes. Typically, a reversible change in thermodynamics is effected as a succession of infinitesimal steps. This gives rise to the Boltzmann paradox: since all possible motions in a piece of matter should share in its thermal energy, if there were an endless regression of the kind molecules → atoms → protons etc. → … then immense, indeed infinite energy would be needed to heat up matter, but this is evidently not the case. The existence of quanta resolves the Boltzmann paradox: the notion of a succession of infinitesimal steps loses meaning in the quantum realm.

When it comes to the living world, classical features seem to be even more apparent than in the macroscopic inanimate world—human perceptions such as the taste of a foodstuff depend on the recent history of what one has eaten, and the psychological circumstances at the time of eating, for example; and human identity is essentially our life history, in which our actual chemical constitution—anyway constantly changing—plays only a secondary role. Yet even life makes use of exquisitely identical macromolecules—nucleic acids (RNA and DNA) and proteins, and in that sense is also alien to the spirit of classical physics.

The ultimate aim of nanotechnology—bottom-to-bottom assembly in a eutactic environment, or programmable assembly at the atomic scale—is far closer to the true quantum world than the classical world. In order to facilitate bridging the gap between nanoscale artifacts and those suitable for human use, "nanoblocks" have been proposed as an intermediate level of object. These nanoblocks would be produced with atomic precision, but their own assembly into larger structures would be easier than working with atoms—self-assembly is likely to be the most convenient general route. Every nanoblock (of a given type) would be identical to every other one, regardless of where it had been produced. Yet each nanoblock would probably contain thousands or tens of thousands of atoms—comparable in size to the proteins of a living cell—and hence would not rank as absolutely small according to Dirac, yet would nevertheless possess quality, specificity and individuality. In this sense nanotechnology represents a kind of compromise between the classical and quantum realms, an attempt to possess the advantages of both. The advantages can be seen particularly clearly in comparison with chemistry, which attempts to create entities (chemical compounds) possessing quality, specificity and individuality, but in the absence of a eutactic environment the yield of reactions that should lead to a unique product are usually significantly below unity. The difference between chemistry and nanotechnology is analogous to the difference between analog and digital ways of representing information: in the latter the basic entities (e.g., zero and one) have a specific, immutable individuality (even though the voltage representing "one" in a digital computer may actually have a value between, say, 0.6 and 1.5 V), whereas in an analog device the information is directly represented by the actual voltage, which may be subject to some systematic bias as well as to unavoidable fluctuations.

Returning to the issue of measurement, Heisenberg has remarked that the mathematical description of a quantum object does not represent its behavior but rather our knowledge of its behavior. This brings clarity to another classical paradox—in not a single instance is it possible to predict a physical event exactly, since as measurements become more and more accurate, the results fluctuate. The indeterminist school deals with this by asserting that every physical law is of a statistical nature; the opposing school asserts that the laws apply exactly to an idealized world picture, to which an actual measurement can only approximate [133]. Clearly the latter viewpoint is appropriate to the quantum world, in which we can predict the probability of an event. Nanotechnology in effect creates a simplified version of the world, in which only a finite set of discrete states are available, whose occurrence (e.g., as a result of a nanofacturing process) can be predicted exactly.

Hence we can say that an ideal nano-object should be small enough to possess quality, specificity and individuality, like a quantum object, but large enough for its state not to be destroyed by measuring one of its attributes, such as its position. Here we seem to be approaching a fundamental definition of a nano-object, rather than merely a phenomenological one (albeit abstracted to essential characteristics), or an ostensive one.

The quantum limit corresponds to an irreducible lower limit of smallness but, depending on the phenomenon under consideration, it might be way beyond the size scale of a single atom; for example, the ultimate lower length scale is given by the Planck length (defined solely using fundamental constants) $\sqrt{\hbar G/c^3} \approx 1.6 \times 10^{-35}$ m.

2.9 SUMMARY

Consideration of what happens to things when we reduce their size reveals two kinds of behavior. In one group of phenomena there is a discontinuous change of properties at a certain size. This change can be very reasonably taken to demarcate the nanoscale. Note that qualitative changes in behavior with size only seem to occur "near the bottom". Hence there is no need to separate changes occurring at the nanoscale from those occurring at the micrometer, millimeter or even meter scale because there are none of the kind we are talking about. In the other group the properties change gradually without any discontinuous (qualitative) change occurring.

Evidently if nanotechnology is merely a continuation of the trend of ever better machining accuracy, as it was viewed by Taniguchi [159], we do not need to worry about qualitative differences, at least not at the hardware level. In this case the upper boundary of the nanoscale must be somewhat arbitrary, but one hundred nanometers seems entirely reasonable, from which it follows that it is really immaterial whether this boundary is taken to be approximate or exact. This definition would fit current usage in top–down nanomanufacture (see Chapter 8)—ultraprecision machining and semiconductor processing. In this usage, nano-objects and devices and the processes

Table 2.3 Summary of phenomenologically-based nanoscale-defining lengths

Domain	Defining length	Formula	Typical value/nm
Surfaces	(Geometry)		5
Nucleation	Critical nucleus size	(2.6)	5
Optics and electronics	Bohr radius	(2.12)	10
Magnetism	Single domain size	§2.6	50
Mechanics	Griffith length	(2.23)	50

used to make them are qualitatively the same as at bigger scales. The inclusion of phrases such as "where properties related to size can emerge" in definitions of nanotechnology (Section 1.1.1) is in this case superfluous, other than as a warning that something unexpected *might* happen.

In contrast, such unexpected phenomena constitute the essential part of the definition of nanotechnology in the first group, in which the nanoscale reveals itself as property-dependent. Furthermore, the scale may also depend on external variables such as temperature and pressure: what appears indubitably as a nanoscale phenomenon at one temperature might cease to be thus distinguished when hotter or colder. In order to permit comparisons between different sets of phenomena in this group, the characteristic size parameter can be normalized by the critical nanoscale-defining length of that phenomenon. Thus, for example, when comparing nanoparticles of different materials with respect to their optical properties, their dimensionless size would be given by r/r_B, where r is a characteristic length of the object under consideration, and the nanorealm concerns ratios $r/r_B < 1$. Table 2.3 summarizes some of these nanoscale-defining lengths (cf. Section 1.1.1).

Where there is no discontinuity, we can take the Hegelian concept of quantitative change becoming qualitative if great enough to justify the application of the term "nanotechnology". Usually this demands consideration of function (utility). Thus, even though the circuit elements in the current generation of very large-scale integrated circuits with features a few tens of nanometers in length work in exactly the same way as their macroscopic counterparts, new function (e.g., practicable personal cellular telephony) emerges upon miniaturization.

2.10 FURTHER READING

J.J. Ramsden, Less is different, Nanotechnol. Percept. 6 (2010) 57–60.

Forces at the Nanoscale

3

CHAPTER CONTENTS

We have established in the previous chapter that the smallness (of nano-objects) implies preponderance of surfaces over bulk. This implies that interfacial forces are particularly important in the nanorealm, governing the behavior of nano-objects and determining the performance of nanostructured materials and nanodevices. Of particular importance are the interfacial forces that are electrostatic in origin: Section 3.2 is mostly about them. The so-called local van der Waals or Casimir force is of unique relevance to the nanoscale—it is described first, in Section 3.1. As for the other fundamental forces, the gravitational force is so weak at the nanoscale—distance or mass—that it can be neglected. Conversely, the range of the strong nuclear force is much smaller than the nanoscale, and hence can also be neglected. The capillary force, which in the past has sometimes been considered as a "fundamental" force, is of course simply a manifestation of the wetting phenomena originating in the interfacial forces discussed in Section 3.2.

Familiar monolithic substances rely on strong metallic, ionic or covalent bonds to hold their constituent atoms together. These three categories correspond to the "elements" of mercury, sodium chloride and sulfur that Paracelsus added to the classical Greek ones of earth (solid—but possibly also encompassing granular matter), air (gas), fire (plasma) and water (liquid). Presumably mankind first encountered complex multicomponent materials (i.e., composites) in nature—wood is perhaps the best example. While the components individually (in the case of wood, cellulose and lignin) are covalently bonded, bonding at the cellulose–lignin interface depends on the relatively weak forces discussed in Section 3.2. Thus, knowledge of these interfacial forces are essential for understanding nanocomposites (Section 6.5). They also enable the self-assembly of nano-objects to take place (see Section 8.2.1).

3.1 THE CASIMIR FORCE

A cavity consisting of two mirrors facing each other disturbs the pervasive zero-point electromagnetic field, because only certain wavelengths of the field vibrations can fit exactly into the space between the mirrors. This lowers the zero-point energy density in the region between the mirrors, resulting in an attractive Casimir force between them [28]. The force falls off rapidly (as z^{-4}) with the distance z between the mirrors, and hence is negligible at the microscale and above, but at a separation of 10 nm it is comparable with atmospheric pressure (10^5 N/m), and therefore can be expected to affect the operation of nanoscale mechanical devices (Figure 3.1) and nanoscale assembly.

(a)   (b)
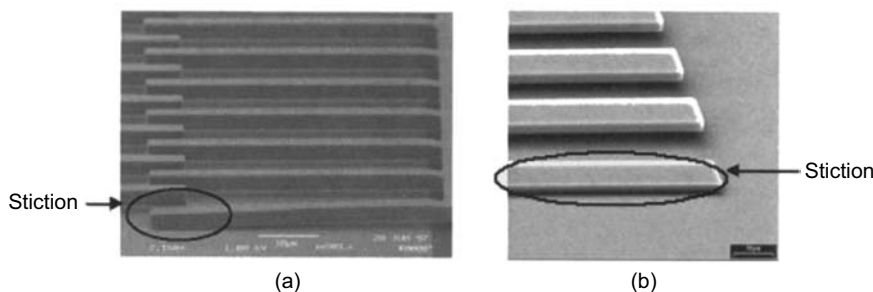
FIGURE 3.1

An illustration of the practical problems that may arise when micro or nano components approach each other to within nanoscale distances. (a) Stuck finger on a comb drive; (b) cantilever after release etch, adhering to the substrate. These effects are often collectively categorized as manifestations of stiction ("sticking friction"). Condensation of water and the resulting capillary force also contributes as well as the Casimir force. Reproduced with permission from [28].

3.2 **INTERMOLECULAR INTERACTIONS**

The forces described in this section are responsible for the relatively weak interactions between two objects, typically in the presence of an intervening medium. They are all electrostatic in origin. As Pilley has pointed out in his book on electricity, it is pervasive in bulk matter but not directly perceptible in everyday life. On the other hand, it is quite likely to be manifested at surfaces. When they carry electrostatic charges (they are sometimes called electrified interfaces)—for example, created by rubbing in air, or by ionization in solution—they will manifest Coulombic interaction. Even neutral surfaces will, however, interact via the Lifshitz–van der Waals family of forces. Finally there are the Lewis acid/base interactions, involving some electron exchange between the partners (e.g., the hydrogen bond), which might be considered as a weak covalent interaction.

3.2.1 **The Concept of Surface Tension**

Surface tension γ is formally defined as the free energy G required to create and extend an interface of area A:

$$\gamma = (\partial G / \partial A)_{T,P} \tag{3.1}$$

where the practically encountered constant temperature and pressure makes the Gibbs free energy the appropriate choice for G. In the Système International, the units of γ are N/m, which is the same as an energy per unit area (J/m^2). It is customary to refer to γ as a surface tension if the increase of area is reversible, and as a surface energy if it is not.

Generally speaking, work needs to be done to create an interface; it has a higher free energy than the bulk. The work of cohesion of a solid is

$$W^{(\text{coh})} = 2\gamma_1 A = -\Delta G^{(\text{coh})} \tag{3.2}$$

(see Figure 3.2). On the other hand, the work of adhesion (needed to separate two dissimilar substances 1 and 2) is given by (see Figure 3.2)

$$W_{12}^{(\text{adh})} = (\gamma_1 + \gamma_2 - \gamma_{12})A = -\Delta G^{(\text{adh})} \tag{3.3}$$

a formalism introduced in the 19th century by Dupré. γ_1 and γ_2 account for the old interfaces lost, and γ_{12} accounts for the new interface gained. Most of the subsequent difficulties experienced by the field of interfacial interactions have concerned the theoretical calculation (i.e., prediction) of terms involving two (or more) substances such as γ_{12}. The nanoscopic viewpoint is that the "microscopic" surface tension (or energy) γ_{12} depends on specific chemical interactions between the surfaces of the two substances 1 and 2.

Fowkes, Girifalco and Good introduced the reasonable assumption that the tension at the interface of substance 1 against substance 2 is lowered by the presence of 2 by an amount equal to the geometric mean of the tensions of the two substances
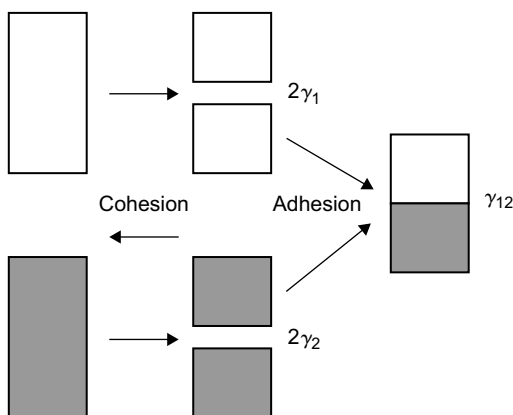
FIGURE 3.2

Cohesion and adhesion of substances 1 (white) and 2 (gray) (see text).

individually, hence equal to $\gamma_1 - (\gamma_1\gamma_2)^{1/2}$, and similarly the tension at the interface of substance 2 against substance 1 is $\gamma_2 - (\gamma_1\gamma_2)^{1/2}$. Summing these two terms, we have

$$\gamma_{12} = \gamma_1 + \gamma_2 - 2(\gamma_1\gamma_2)^{1/2} = (\sqrt{\gamma_1^{(LW)}} - \sqrt{\gamma_2^{(LW)}})^2 \qquad (3.4)$$

called the Girifalco–Good–Fowkes equation. This is equivalent to the work of adhesion being the geometric mean of the works of cohesion, i.e. $W_{12} = (W_{11}W_{22})^{1/2}$. The Dupré equation (3.3) then becomes

$$W_{12}^{(adh)} = 2(\gamma_1\gamma_2)^{1/2}A. \qquad (3.5)$$

Fowkes and van Oss developed the idea that the total interfacial energy is linearly separable into the dispersive (London–van der Waals), dipole-induced dipole (Debye), dipole–dipole (Keesom) and electron donor–acceptor terms, and Lifshitz has pointed out that the London–van der Waals, Debye and Keesom interactions are all of the same type (cf. the Hellman–Feynman theorem), with the same dependence of magnitude on separation between the two interacting substances, and hence

$$\gamma^{(total)} = \gamma^{(LW)} + \gamma^{(ab)} \qquad (3.6)$$

where LW denotes Lifshitz–van der Waals and ab denotes (Lewis) acid/base, and *a fortiori*

$$\gamma_{12}^{(total)} = \gamma_{12}^{(LW)} + \gamma_{12}^{(ab)}. \qquad (3.7)$$

Whereas the Lifshitz–van der Waals interaction is always attractive, the sign of the Lewis acid/base interaction depends on the relative proportions of Lewis acids and Lewis bases constituting the two interacting substances. Superscript $\ominus$ will be used to denote electron-donating (Lewis base) and superscript $\oplus$ will be used to denote electron-accepting (Lewis acid) moieties; van Oss has proposed that one might again

take the geometric mean, namely

$$\gamma^{ab} = 2(\gamma^{\ominus}\gamma^{\oplus})^{1/2}. \tag{3.8}$$

Two monopolar substances of the same sign will repel each other; attraction depends on the presence of cross-terms. By analogy with equation (3.5),

$$W_{12}^{(adh)} = 2[(\gamma_2^{\oplus}\gamma_1^{\ominus})^{1/2} + (\gamma_1^{\oplus}\gamma_2^{\ominus})^{1/2}]A \tag{3.9}$$

Hence the ab combining law is

$$\gamma_{12}^{(ab)} = 2[(\gamma_1^{\oplus}\gamma_1^{\ominus})^{1/2} + (\gamma_2^{\oplus}\gamma_2^{\ominus})^{1/2} - (\gamma_1^{\oplus}\gamma_2^{\ominus})^{1/2} - (\gamma_1^{\ominus}\gamma_2^{\oplus})^{1/2}]$$
$$= 2(\sqrt{\gamma_1^{\oplus}} - \sqrt{\gamma_2^{\oplus}})(\sqrt{\gamma_1^{\ominus}} - \sqrt{\gamma_2^{\ominus}}) \tag{3.10}$$

It takes account of the fact that $\ominus$ interacts with $\oplus$, which is why the ab interaction can be either attractive or repulsive. In typical biological and related systems, the Lewis acid/base interaction accounts for 80–90% of the total interactions. The most familiar manifestation is hydrogen bonding (e.g., double-stranded DNA (the double helix), globular proteins containing α-helices); π–π interactions (stacking of alternately electron-rich and electron-deficient aromatic rings) are frequently encountered in synthetic organic supermolecules.

Let us now consider two solids 1 and 3 in the presence of a liquid medium 2 (e.g., in which a self-assembly process, cf. Section 8.2.1, takes place). ΔG_{123} is the free energy per unit area of materials 1 and 3 interacting in the presence of liquid 2. Using superscript $\|$ to denote the interfacial interaction energies per unit area between infinite parallel planar surfaces,

$$\Delta G_{121}^{\|} = -2\gamma_{12} \tag{3.11}$$

and

$$\Delta G_{123}^{\|} = \gamma_{13} - \gamma_{12} - \gamma_{23}. \tag{3.12}$$

From the above equations we can derive:

$$\Delta G_{123}^{(LW,ab)\|} = \Delta G_{22}^{(LW,ab)\|} + \Delta G_{13}^{(LW,ab)\|} - \Delta G_{12}^{(LW,ab)\|} - \Delta G_{23}^{(LW,ab)\|} \tag{3.13}$$

where ΔG_{13} is the free energy per unit area of materials 1 and 3 interacting directly. It follows that:

- LW forces (anyway weak) tend to cancel out;
- the so-called "hydrophobic force" is a consequence of the strong cohesion of water ΔG_{22}. Attraction of suspended solids is only prevented by their hydrophilicity. The sign of ΔG_{12} with 2 = water, provides an unambiguous measure of the hydrophobicity of substance 1: $\Delta G_{12} < 0 \equiv$ hydrophilic; $\Delta G_{12} > 0 \equiv$ hydrophobic (see also Section 3.2.4).

$\Delta G_{123}^{\|}$ can be used to provide a rapid first estimate of whether adhesion between materials 1 and 3 will take place in the presence of medium 2. Tables 3.1 and 3.2 give some typical values (see Section 3.2.3 for information about their determination).

Table 3.1 Surface tension parameters of some solids [127]

Material	$\gamma^{(LW)}$/mJ m^{-2}	$\gamma^{\oplus}$/mJ m^{-2}	$\gamma^{\ominus}$/mJ m^{-2}
Synthetic polymers			
Nylon 6,6	36	0.02	22
PMMA	41	0	13
Polyethylene	33	0	0
Polyethylene oxide	43	0	64
Polystyrene	42	0	1.1
Polyvinylpyrrolidone	43	0	30
PVC	43	0.04	3.5
Teflon	18	0	0
Carbohydrates			
Cellulose	44	1.6	17
Dextran T-150	42	0	55
Metal oxides			
SiO_2	39	0.8	41
SnO_2	31	2.9	8.5
TiO_2	42	0.6	46
ZrO_2	35	1.3	3.6

Table 3.2 Surface tensions of some liquids (data mostly from [127])

Liquid	$\gamma^{(LW)}$/mJ m^{-2}	$\gamma^{\oplus}$/mJ m^{-2}	$\gamma^{\ominus}$/mJ m^{-2}
Water[a]	22	25.5	25.5
Glycerol	34	3.9	57
Ethanol	19	0	68
Chloroform	27	3.8	0
Octane	22	0	0
n-hexadecane	27.5	0	0
Formamide	39	2.3	40
α-bromonaphthalene	44	0	0
Diiodomethane	51	0	0

[a]Absolute values of $\gamma^{\oplus}$ and $\gamma^{\ominus}$ are not known at present; values are arbitrarily assigned to ensure that the known overall γ is correct (equation 3.8).

According to the Derjaguin approximation, a sphere of radius *r* (material 3) inter-acting with an infinite planar surface (material 1) has the following free energies of interaction as a function of *z*, the perpendicular distance between the plane and the nearest point of the sphere:

$$\Delta G^{(LW)}(z) = 2\pi \ell_0^2 \Delta G^{(LW)\parallel} r/z \tag{3.14}$$

where ℓ_0 is the equilibrium contact distance (about 0.15 nm);

$$\Delta G^{(da)}(z) = 2\pi \chi \, \Delta G^{(ab)\|} \exp[(\ell_0 - z)/\chi]r \tag{3.15}$$

where χ is the decay length for the ab interactions; and where electrostatic charges are present,

$$\Delta G^{(el)}(z) = 4\pi \epsilon_0 \epsilon \psi_3 \psi_1 \ln[1 + \exp(-\kappa z)]r \tag{3.16}$$

where ψ are the electrostatic surface potentials of materials 1 and 3 and $1/\kappa$ is the Debye length (inversely proportional to the square root of the ionic strength—this is why electrostatic interactions tend not to be very important in salty aqueous systems).

3.2.2 Critique of the Surface Tension Formalism

As Bikerman has pointed out [20], there are some difficulties with applying the concept of surface tension to solids. It is generally recognized that most real, machined solid surfaces are undulating as shown in Figure 3.3. Capillary pressure P_c:

$$P_c = 2\gamma/r \tag{3.17}$$

where r is the radius of the undulation, would tend to flatten such undulations since the pressure where r is small would engender a strain easily exceeding the yield strength of most metals, yet such undulations persist.

The equation for the three-phase line between a vapor (phase 1) and two liquids (phases 2 and 3) is well known to be

$$\frac{\sin \hat{1}}{\gamma_{23}} = \frac{\sin \hat{2}}{\gamma_{13}} = \frac{\sin \hat{3}}{\gamma_{12}} \tag{3.18}$$

where the angles $\hat{1}$, etc. enclose phase 1, etc. The assumption that solids have a surface tension analogous to that of liquids ignores the phenomenon of epitaxy (in which a crystal tends to impose its structure on material deposited on it). As an illustration of the absurdities to which the uncritical acceptance of this assumption may lead, consider a small cube of substance 2 grown epitaxially in air (substance 3) on a much larger cube of substance 1 (Figure 3.4). The angles A and B are both 90° and correspond to $\hat{3}$ and $\hat{2}$ respectively in equation (3.18). It immediately follows that the
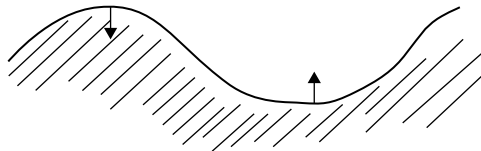

FIGURE 3.3

Sketch of the undulations of a somewhat idealized real machined surface. The arrows indicate the direction of capillary pressure.
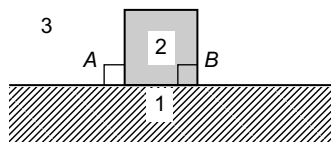
FIGURE 3.4

Illustration of a small cube of substance 2 grown epitaxially on a large cube of substance 1. After Bikerman [20].

surface tension of substance 1 equals the interfacial tension between substances 1 and 2, a highly improbable coincidence. Moreover, since $\hat{1}$ equals $180°$, the surface tension of substance 2 must be zero, which also seems highly unlikely.

Given these difficulties in the application of the surface tension concept to rigid solids, it is appropriate to exercise caution when using the approach described in Section 3.2.1. Many of the applications in nanotechnology (especially those concerned with the nano/bio interface) involve, however, soft matter, to which the application of the formalism may be reasonable.

3.2.3 Experimental Determination of Single-Substance Surface Tensions

The general strategy is to measure the advancing contact angles θ on the material 3 whose surface tension is unknown using three appropriate liquids with different surface tension components (determined, e.g., from hanging drop measurements; the shape of the drop, in particular its two radii of curvature r_1 and r_2, are measured for a pressure difference ΔP and the surface tension calculated from the Laplace law, equation (2.2), i.e., $\gamma = \Delta P/(1/r_1 + 1/r_2)$) [127]. With these values, three Young–Dupré equations:

$$\left[\gamma_2^{(LW)}/2 + (\gamma_2^{\oplus}\gamma_2^{\ominus})^{1/2}\right](1+\cos\theta) = \left(\gamma_1^{(LW)}\gamma_2^{(LW)}\right)^{1/2} + \left(\gamma_1^{\oplus}\gamma_2^{\ominus}\right)^{1/2} + \left(\gamma_1^{\ominus}\gamma_2^{\oplus}\right)^{1/2}$$
(3.19)

can be solved to yield the unknowns $\gamma_1^{(LW)}$, $\gamma_1^{\oplus}$ and $\gamma_1^{\ominus}$. Values of the surface tension parameters for some common materials have been given in Tables 3.1 and 3.2. Methods for computing ψ for ionized solids (e.g., polyions, protonated silica surfaces) are based on the Healy–White ionizable surface group model [72].

A major practical difficulty in experimentally determining surface tensions of solutes is that under nearly all real conditions, surfaces are contaminated in a highly nonuniform fashion. Figure 3.5 shows how surface roughness can be the origin of this nonuniformity. Adsorbed impurity molecules will experience different interfacial energies at different locations on the surface due to the different curvatures.
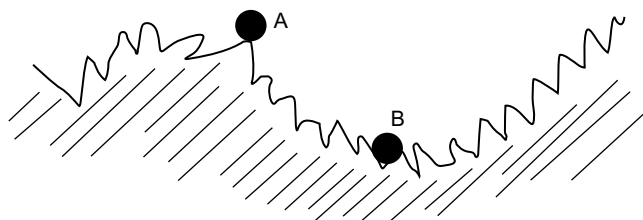
FIGURE 3.5

Cross-section of a possible real machined surface. Note the two distinct length scales of roughness. A and B are adsorbed impurity molecules. After Bikerman [20].

3.2.4 Wetting and Dewetting

Wetting means the spreading of a liquid over a solid surface; dewetting is its converse, the withdrawal of liquid from a surface (the term "wetting" is used regardless of the nature of the liquid—hence "dry" in this context means not that water is absent, but that all liquid is absent). They are basic processes in countless natural and industrial processes. Although pioneering work in characterizing the interfacial tensions upon which wetting depends was reported 200 years ago by Young, the processes are still relatively poorly understood. Wetting is mostly considered from a mesoscopic viewpoint and therefore fits well into the framework of nanotechnology. Few experimental techniques are available for investigating the important solid/liquid interfaces: the contact angle method is simple and probably still the most important, but only a handful of laboratories in the world have shown themselves capable of usefully exploiting it. The history of dewetting, a phenomenon of no less industrial importance than wetting, is much more recent: quantitative experimental work dates from the early 1990s.

It is essentially intuitive to expect that the spreading of a liquid on a solid surrounded by vapor depends on γ_{SV} (S = solid, V = vapor, L = liquid). The quantitative relationship was given by Young in 1805 (cf. equation 3.19):

$$\gamma_{LV} \cos\theta = \gamma_{SV} - \gamma_{SL} \qquad (3.20)$$

The degree of wetting is inversely proportional to the contact angle θ; $\theta = 0$ corresponds to complete wetting. Young's equation (3.20) can be easily derived by noting that the surface tension can be written as a force per unit distance. The interfacial forces acting on the triple line T, where three phases S, L, V (solid, liquid, vapor) meet must sum to zero in a given direction (x, parallel to the interface) (Figure 3.6). More formally, it follows from the condition that (at equilibrium) the energies must be invariant with respect to small shifts dx of the position of T. The structure of T may be very complex. For example, for water containing dissolved electrolyte, the local ion composition may differ from that in the bulk; soft solids may be deformed
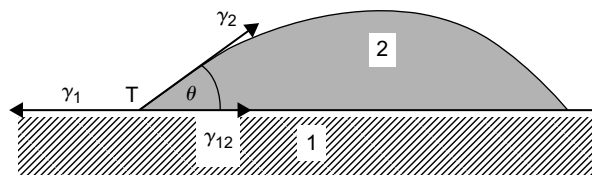
FIGURE 3.6

A drop of a liquid (substance 2) on a solid (substance 1). The vapor is neglected here (i.e., $\gamma_1 \equiv \gamma_{SV}$, etc.).

in the vicinity of T. Although Young's equation ignores these details it provides a remarkably accurate description of contact angles.

3.2.5 Length Scales for Defining Surface Tension

The region in which deviations from "far field" quantities occur is known as the core, with radius $r_C \sim 10$ nm. Hence for typical drops (with a radius $R \sim 1$ mm) used in the contact angle determinations from which surface tension can be calculated, the curvature of T may be neglected; atomic scale heterogeneity on the subnanometer scale may also be neglected.

3.2.6 The Wetting Transition

Complete wetting is characterized by $\theta = 0$, which implies (from equation 3.20)

$$\gamma_{LV} = \gamma_{SV} - \gamma_{SL} \tag{3.21}$$

at equilibrium (out of equilibrium, this relation may not hold). The Cooper–Nuttall spreading coefficient S is

$$S = \gamma_{\tilde{S}V} - \gamma_{SL} - \gamma_{LV} \tag{3.22}$$

where $\gamma_{\tilde{S}V}$ is the interfacial tension of a dry solid. Three regimes can thus be defined:

1. $S > 0$. This corresponds to $\gamma_{\tilde{S}V} > \gamma_{SV}$, i.e. the wetted surface has a lower energy than the unwetted one. Hence wetting takes place spontaneously. The thickness h of the film is greater than monomolecular if $S \ll \gamma_{LV}$. The difference $\gamma_{\tilde{S}V} - \gamma_{SV}$ can be as much as 300 mJ/m^2 for water on metal oxides. Such systems therefore show enormous hysteresis between advancing and receding contact angles. Other sources of hysteresis include chemical and morphological inhomogeneity (contamination and roughness).
2. $S = 0$. Occurs if $\gamma_{\tilde{S}V}$ practically equals γ_{SV}, as is typically the case for organic liquids on molecular solids.
3. $S < 0$. Partial wetting. Films thinner than a certain critical value, usually ~ 1 mm, break up spontaneously into droplets (cf. Section 8.1.2).

3.3 **CAPILLARY FORCE**

It follows from the previous considerations that an attractive interaction between a liquid (of density ρ) and a solid will cause that liquid to rise to a height h within a vertical tube of diameter r made from the solid, emerging from a reservoir of the liquid until the gravitational pull on the liquid column equals the interfacial attraction:

$$h = \frac{}{2\gamma_{LV}\cos\theta_{SL}}\rho gr \tag{3.23}$$

where g is the acceleration due to gravity. For a small diameter tube, the rising of the liquid can be appreciable and easily visible to the naked eye. This is the classical manifestation of the capillary force. Since water is both a strong hydrogen bond donor and a strong hydrogen bond acceptor, it is able to interact attractively with a great many different substances, especially minerals, which tend to be oxidized.

Consequences for Nano-Objects. It is part of everyday experience that a human emerging from a bath carries with them a film of water—depending on the hydrophobicity of their skin—that may be of the order of 100 μm thick, hence weighing a few hundred grams—negligible compared with the weight of its bearer. A wet small furry creature like a mouse has to carry about its own weight in water. A wet fly would have a surplus weight many times its own, which is presumably why most insects use a long proboscis to drink; bringing their bodies into contact with a sheet of water usually spells disaster. The inadvertent introduction of water (e.g., by condensation) into microsystems may completely degrade their performance (cf. Figure 3.1); nanosystems are even more vulnerable.

The Force Environment of Nano-Objects and Nanodevices. Apart from the capillary force referred to above, we should also consider objects wholly immersed in a liquid. In this environment viscosity, rather than inertia, dominates. In contrast to the Newtonian mechanics appropriate to describing the solar system, the movement of a bacterium swimming in water is governed by a Langevin equation comprising frictional force and random fluctuations (Brownian motion). This also encompasses the realm of soft matter, a good definition of which is matter whose behavior is dominated by Brownian motion (fluctuations).

3.4 **HETEROGENEOUS SURFACES**

The expressions given in the preceding sections have assumed that the surfaces are perfectly flat and chemically homogeneous. Most manufactured surfaces are, however, rough and may be chemically inhomogeneous, either because of impurities present from the beginning, or because impurities acquired from the environment have segregated due to roughness (Figure 3.5).

3.4.1 **Wetting on Rough and Chemically Inhomogeneous Surfaces**

If the roughness is in the nanorange—a few nanometers to a micrometer—the (subnanometer-sized) molecules of the liquid interact locally with planar segments of the surface, while still yielding a unique contact angle, denoted by θ^*, supposed different from the contact angle θ of the same liquid with a perfectly smooth planar surface. Now suppose that the liquid drop on the rough surface is perturbed by a small horizontal displacement by a distance dx in the plane of the surface. Because of the roughness, characterized by the ratio $\mathfrak{r}$ of actual to apparent surface areas, the real distance is $\mathfrak{r}dx$, $\mathfrak{r} > 1$. According to Young's law (3.20), the work done is

$$dW = (\gamma_{SL} - \gamma_{SV})\mathfrak{r}dx + \gamma_{LV}\cos\theta^* dx. \tag{3.24}$$

The drop should then find an equilibrium state at which $dW/dx = 0$, yielding

$$(\gamma_{SV} - \gamma_{SL})\mathfrak{r} = \gamma_{LV}\cos\theta^*. \tag{3.25}$$

Comparison with (3.20) yields Wentzel's law:

$$\cos\theta^* = \mathfrak{r}\cos\theta. \tag{3.26}$$

In other words, since $\mathfrak{r} \geq 1$, roughness will always make a hydrophobic surface more hydrophobic and a hydrophilic surface more hydrophilic. This is the basis of technologies to manufacture superhydrophobic and superhydrophilic surfaces. Typically, natural superhydrophobic materials such as the leaves of the lupin (Figure 3.7) have roughnesses at multiple length scales.

A similar approach has been used to derive the Cassie–Baxter law for surfaces chemically inhomogeneous (on the nanoscale). Suppose that such a surface is constituted from fractions f_i of N different materials, which individually in pure form yield contact angles θ_i with the liquid under test. Then,

$$\cos\theta^\dagger = \sum_{i=1}^{N}\cos\theta_i \tag{3.27}$$

where $\theta^\dagger$ is the effective contact angle. The advent of nanofabrication has yielded intriguing situations of drops resting on the tops of arrays of nano- or micropillars, let us suppose with a height r and covering a fraction f of the surface. If the drop remains on the top of the pillars (a situation that has been called the fakir effect), the surface would presumably be sensed as smooth (with the air–liquid interface between the pillars remaining parallel to the mean plane of the substrate), but chemically heterogeneous, and the Cassie–Baxter law would be applicable, simplifying to $\cos\theta^\dagger = f(1 + \cos\theta) - 1$. If on the other hand the drop completely wets both the horizontal and vertical surfaces, the surface is sensed as rough and the Wentzel law would be applicable. As the "intrinsic" contact angle varies, the drop can minimize its energy by either obeying the Wentzel law (for a hydrophilic material) or the Cassie–Baxter law (i.e., displaying the fakir effect, for hydrophobic surfaces). The crossover point between the two regimes is given by $\cos\theta = (f - 1)/(r - f)$.

FIGURE 3.7

Drops of rainfall on the leaves of the lupin showing the superhydrophobicity.

3.4.2 Three-Body Interactions

The main structural unit of proteins is the alpha helix. Simple polyamino acids such as polyalanine will fold up spontaneously in water into a single alpha helix. Many proteins, such as the paradigmatical myoglobin, can be adequately described as a bundle of alpha helical cylinders joined together by short connecting loops. The alpha helix is held together by hydrogen bonds between the ith and $i + 4$th amino acid backbones. Denatured myoglobin has no difficulty in refolding itself spontaneously in water, correctly reforming the alpha helices, but one might wonder why the process is so robust when the denatured polypeptide chain is surrounded by water molecules, each of which is able to donate and accept two hydrogen bonds (albeit that at room temperature, 80–90% of the possible hydrogen bonds between water molecules are already present—see Section 3.8). The paradox is that although refolding is an intramolecular process, the overwhelming molar excess of water should predominate, ensuring that the backbone hydrogen bonds are always fully solvated by the surrounding water and hence ineffectual for creating protein structure.

One of the intriguing features of natural amino acid sequences of proteins is their blend of polar and apolar residues. It was formerly believed that the protein folding problem involved simply trying to pack as many of the apolar residues into the protein interior as possible, in order to minimize the unfavorable free energy of interaction between water and the apolar residues. Nevertheless, a significant number (of the

order of 50% of the total, for a folded globular protein of moderate size) of apolar residues remains on the folded protein surface.

The presence of an apolar residue in the vicinity of a hydrogen bond is a highly effective way of desolvating it. Folding success actually involves the juxtaposition of appropriate apolar residues with backbone hydrogen bonds. The effectiveness of desolvation of a hydrogen bond can be computed by simply counting the number of apolar residues within a sphere of about 7 Å radius centered midway between the hydrogen bond donor and the hydrogen bond acceptor [54]. This approach, which can be carried out automatically using the atomic coordinates in the protein data bank (PDB), reveals the presence of *dehydrons*, underdesolvated (or "underwrapped") hydrogen bonds. Clusters of dehydrons are especially effective "sticky patches" on proteins.

3.5 WEAK COMPETING INTERACTIONS

In any assembly process starting from a random arrangement, it is very likely that some of the initial connexions between the objects being assembled are merely opportunistic and at a certain later stage will need to be broken in order to allow the overall process to continue. For this reason, it is preferable for the connexions to be weak (i.e., the types considered in Section 3.2; e.g., hydrogen bonds) in order to enable them to be broken if necessary. The impetus for breakage comes from the multiplicity of competing potential connexions that is inevitable in any even moderately large system.

One example is provided by superspheres (see Section 8.2.9). Another is biological self-assembly (e.g., of compact RNA structures, Section 8.2.11)—bonds must be broken and reformed before the final structure is achieved. This is particularly apparent because these molecules are synthesized as a linear polymer, which already starts to spontaneously fold (which means forming connexions between parts of the polymer that are distant from each other along the linear chain) as soon as a few tens of monomers have been connected. As the chain becomes longer, some of these earlier bonds must be broken to allow connexions between points more distant along the chain to be made. Since, as hydrogen bonds, they have only about one tenth of the strength of ordinary covalent bonds, they have an appreciable probability of being melted (broken) even at room temperature. Furthermore, the polymer is surrounded by water, each molecule of which is potentially able to participate in four hydrogen bonds (although at room temperature only about 10% of the maximum possible number of hydrogen bonds in water are broken—see also Section 3.8). Hence there is ceaseless competition between the intramolecular and intermolecular hydrogen bonds.

3.6 COOPERATIVITY

When considering the interactions (e.g., between precursors of a self-assembly process), it is typically tacitly assumed that every binding event is independent.

Similarly, when considering conformational switches, it has been assumed that each molecule switches independently. This assumption is, however, often not justified: switching or binding of one facilitates the switching or binding of neighbors, whereupon we have cooperativity (if it hinders rather than facilitates, then it is called anticooperativity). A cooperative process can be conceptualized as two subprocesses: nucleation and growth. Let our system exist in one of two states (e.g., bound or unbound, conformation A or B), which we shall label 0 and 1. We have [149]

$$\text{nucleation:} \quad \cdots 000 \cdots \overset{\sigma S}{\rightleftharpoons} \cdots 010 \cdots \quad (3.28)$$

and

$$\text{growth:} \quad \cdots 001 \cdots \overset{S}{\rightleftharpoons} \cdots 011 \cdots . \quad (3.29)$$

Let $\{1\} = \theta$ denote the probability of finding a "1"; we have $\{0\} = 1 - \theta$. The parameter λ^{-1} is defined as the conditional probability of "00" given that we have a "0", written as (00) and equal to $\{00\}/\{0\}$. It follows that $(01) = 1 - (00) = (\lambda - 1)/\lambda$. According to the mass action law (MAL), for nucleation we have

$$S = \frac{\{011\}}{\{001\}} = \frac{(11)}{(00)} \quad (3.30)$$

from which we derive $(11) = S/\lambda$, and hence $(10) = 1 - (11) = (\lambda - S)/\lambda$. Similarly for growth

$$S = \frac{(01)(10)}{(00)^2} = (\lambda - 1)(\lambda - S). \quad (3.31)$$

Solving for λ gives

$$\lambda = [1 + S + \sqrt{(1 - S)^2 + 4\sigma S}]/2. \quad (3.32)$$

To obtain the sought-for relation between θ and S, we note that $\theta = \{01\} + \{11\} = \{0\}(01) + \{1\}(11)$, which can be solved to yield

$$\lambda = \frac{1 + (S - 1)/\sqrt{(1 - S)^2 + 4\sigma S}}{2}. \quad (3.33)$$

Cooperativity provides the basis for programmable self-assembly (Section 8.2.8), and for the widespread biological phenomenon of "induced fit", which occurs when two molecules meet and "recognize" each other, whereupon their affinity is increased.

3.7 PERCOLATION

Percolation can be considered to be a formalization of gelation. Let us consider the following. Initially in a flask we have isolated sol particles, which are gradually connected to each other in nearest-neighbor fashion until the gel is formed. Although the particles can be placed anywhere in the medium subject only to the constraint of

hard-body exclusion, it is convenient to consider them placed on a two-dimensional square lattice (imagine a Gō board). Two particles are considered to be connected if they share a common side (this is called site percolation). Alternatively, neighboring lattice points are connected if they are bonded together (bond percolation). In principle the lattice is infinite but in reality it may merely be very large. Percolation occurs if one can trace the continuous path of connexions from one side to the other. Initially all the particles are unconnected. In site percolation, the lattice is initially considered to be empty, and particles are added. In bond percolation, initially all particles are unconnected and bonds are added. The problem is to determine what fraction of sites must be occupied, or how many bonds must be added, in order for percolation to occur. In the remainder of this section we shall consider site percolation. Let the probability of a site being occupied be p (and of being empty, $q = 1 - p$). The average number of singlets per site is $n_1(p) = pq^4$ for the square lattice, since each site is surrounded by four shared sides. The average number of doublets per site is $n_2(p) = 2p^2q^6$, since there are two possible orientations. A triplet can occur in two shapes, straight or bent, and so on. Generalizing,

$$n_s(p) = \sum_t g(s,t)p^s q^t \tag{3.34}$$

where $g(s,t)$ is the number of independent ways that an s-tuplet can be put on the lattice, and t counts the different shapes. If there is no "infinite" cluster (i.e., one spanning the lattice from side to side) then

$$\sum_s sn_s(p) = p \tag{3.35}$$

The first moment of this distribution gives the mean cluster size

$$S(p) = \frac{\sum s^2 n_s(p)}{p} \tag{3.36}$$

Writing this as a polynomial in p using equation (3.34), it will be noticed that for $p < p_c$ the value of the series converges, but for $p > p_c$ the value of the series diverges. The "critical" value $p = p_c$ corresponds to the formation of the "infinite" cluster. For site percolation on a square lattice, $p = p_c = 0.5$; the universal Galam–Mauger formula [59]

$$p_c = a[(d-1)(q-1)]^{-b} \tag{3.37}$$

with $a = 1.2868$ and $b = 0.6160$ predicts p_c, with less than 1% error, for all known lattices of connectivity q embedded in a space of dimension d. The larger the lattice, the sharper the transition from not percolating to percolating. For a 3×3 lattice there can be no percolation for two particles or less, but the probability of randomly placing three particles in a straight line from edge to edge is evidently one third.

3.8 THE STRUCTURE OF WATER

So much nanotechnology takes place in water, including microfluidic-enabled reactors, many self-assembly processes, as well as nanobiotechnology and bionanotechnology, that it is important to recall some of the salient features of this remarkable liquid.

Water—H–O–H—can participate in four hydrogen bonds (HB). The two electron lone pairs (LP) on the oxygen atom are electron donors, hence HB acceptors. The two hydrogens at the ends of the hydroxyl groups (OH) are HB donors, hence electron acceptors. The equilibrium

$$H_2O_{\text{fully bonded}} \rightleftharpoons OH_{\text{free}} + LP_{\text{free}} \qquad (3.38)$$

is balanced such that at room temperature about 10% of the OHs and LPs are nonbonded, i.e. free. It is especially noteworthy that the concentrations of these two free species are 7–8 orders of magnitude greater than the concentrations of the perhaps more familiar entities H^+ and OH^-, and their chemical significance is correspondingly greater.

The OH moiety has a unique infrared absorption spectrum, different according to whether it is hydrogen-bonded or free, which can therefore be used to investigate reaction (3.38). A striking example of how the equilibrium can be controlled is given by the spectroscopic consequences of the addition of cosolutes. If sodium chloride is added to water, the Na^+ and Cl^- ions can respectively accept and donate electron density to form quasihydrogen bonds to appropriate donors and acceptors in roughly equal measure, and the intensity of the OH band in the infrared spectrum does not change. If sodium tetraphenylborate is added, the borate ion is prevented by its bulky phenyl ligands from interacting with the water, resulting in fewer bonded OH groups; hence in order to maintain the equilibrium (3.38) the concentration of free lone pairs must diminish. Conversely, if tetrabutylammonium chloride is added, there will be an excess of bonded OH.

On the basis of extensive experiments on hydrogels, Philippa Wiggins has proposed that two kinds of water, low and high density with respectively more and less hydrogen bonding, can be created by surfaces in contact with aqueous solutions. Although the theoretical interpretation of the phenomenology is as yet incomplete, whatever its origins it must inevitably have profound effects on nanoscale fluidic circuits, in which essentially all of the fluid phase is in the vicinity of solid surfaces.

3.9 SUMMARY

The large ratio of surface to volume characteristic of nanoscale objects and devices places interfacial forces in a prominent position in governing their behavior. Although subject to criticism, the surface tension formalism allows the magnitudes of these forces between objects made from different materials in the presence of different

liquids to be quickly estimated from tables of experimentally derived "single substance surface tensions".

In aqueous systems, Lewis acid/base interactions, most notably hydrogen bonding, typically dominate the interfacial forces.

Real, that is morphologically and chemically heterogeneous, surfaces require some modification to the basic theory. In some cases a simple mean-field correction may be adequate; in others nanostructure must explicitly be taken into account. This is especially strikingly shown in protein interactions, which are actually three-body in nature and depend on the subtle interplay of solvation and desolvation.

Multiple forces of different strength and range may be operating simultaneously. This provides the basis for programmability. The assembly of objects into constructions of definite size and shape is only possible if programmability is incorporated, and in the nanoscale this can typically only be achieved by judicious design at the level of the constituent atoms and groups of atoms of the objects.

Multibody connexions provide the basis for cooperativity, an essential attribute of many "smart" devices. Another collective phenomenon is percolation, which is a paradigm for the assembly process known as gelation.

3.10 FURTHER READING

C. Binns, Prodding the cosmic fabric with nanotechnology, Nanotechnol. Percept. 3 (2007) 97–105.

M.G. Cacace, E.M. Landau, J.J. Ramsden, The Hofmeister series: salt and solvent effects on interfacial phenomena, Q. Rev. Biophys. 30 (1997) 241–278.

P.G. de Gennes, Wetting: statics and dynamics, Rev. Mod. Phys. 57 (1985) 827–863.

J.S. Rowlinson, B. Widom, Molecular Theory of Capillarity, Oxford: Clarendon Press, 1982.

The Nano/Bio Interface

4

CHAPTER CONTENTS

As nanotechnology becomes more widely written and talked about, the term "nano/bio interface", as well as several variants such as the "nano–bio" and "bio–nano interfaces", have emerged. This chapter gives an extended definition of what the nano/bio interface is and what it might involve.

One of nanotechnology's protagonists, Eric Drexler, has robustly countered scepticism that Feynman's concept of making machines that make smaller machines could be continued down to the atomic level (Section 1.3) by adducing the many examples of biological mechanisms operating at the nanoscale [43]. This biological "living proof-of-principle" has since been vastly reinforced by the acquisition of ever more detailed knowledge of the mechanisms of these biological machines. This "living proof-of-principle" may be considered as a *virtual* nano/bio interface.

Apart from this rather abstract meaning, the nano/bio interface clearly must signify the interface between a nanostructured nonliving domain and the living domain. It therefore constitutes a special case of the nonliving/living interface, in which the scale of salient features on the nonliving side is constrained to fall within a certain range. Nothing is, however, specified regarding the scale of the living side. This could be the very largest scale, that of an entire ecosystem or of society. Below that, it might

be useful to consider (multicellular) organisms, organs, tissues, cells and molecules. Clearly very different phenomena enter at each scale.

Let us consider the distinction between information flowing from the nano domain to the bio domain and vice versa. These two sitiations are denoted respectively as the nano–bio and bio–nano interfaces. The meanings of the two can be quite different. For example, considering the interface between nanotechnology and society, the nano–bio interface denotes the impact nanotechnology has on society; for example, how it might be completely transformed by the advent of molecular manufacturing. Conversely, the bio–nano interface denotes the introduction of regulatory measures for nanotechnology. Considering the general environment, the nano–bio interface implies, for example, the response of soil microbes to nanoparticles added for remediation purposes; and conversely the bio–nano interface implies the destruction of nanoparticles added for remediation purposes by soil microbes. At the scale of an organism (e.g., a human being), the nano–bio interface signifies the technology of scaling atomic-scale assembly up or out to provide human-sized artifacts; conversely the bio–nano interface corresponds to the man–machine interface, which a human being would use to control a nanoscale assembly process. Insofar as digitally encoded control is likely to be used, there should be no difference in principle between the interface for controlling a nanoscale process and one for controlling a macroscale process.

Feynman [56] mentions Albert R. Hibbs's suggestion of a miniature "mechanical surgeon" (nowadays referred to as a nanoscale robot or nanobot) able to circulate within the bloodstream and carry out repairs *in situ*. Hogg [77] conceives such a device as being about the size of a bacterium, namely a few hundred nanometers in diameter. The nano/bio interface is literally (physically) delineated by the outer surfaces of the device and its nanoscale appendages; that is, the zone between the device and its living host. More generally, this meaning of the nano/bio interface refers to any situation in which a nanomaterial or a nanodevice is in contact with living matter.

If the radius of the "bio" side of the interface is less than that of the "nano" side (Figure 4.1), we can refer to bio–nano; conversely if the radius of the "nano" side of the interface is less than that of the "bio" side, we can refer to nano–bio.

A further meaning of "nano/bio interface" is the means with which humans interact with a nanodevice. While the protagonists of nanobots typically envisage
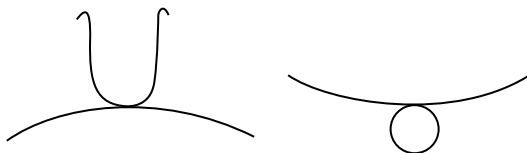

FIGURE 4.1

Left hand: a sharply curved "finger" of a protein abuts an inorganic surface. Right hand: a sharply curved inorganic nanoparticle abuts a living cell membrane.

a fully autonomous device, appraising its surroundings, processing the information internally and carrying out appropriate tasks accordingly, Hibbs presumed that information about its surroundings would be transmitted to a human surgeon, who upon analyzing the data would then direct the repair work, much in the way that certain operations are already carried out today by remotely controlling tools fitted to the end of an endoscope, for example. In this case, the requisite nano/bio interface will be the same as the man–machine interface made familiar through digital information processing technology: the man–machine interface has been a preoccupation of computer scientists ever since the inception of information technology (IT). At present, the issue scarcely arises for nanotechnology, since we do not yet have sophisticated nanosystems that need to be interfaced. The closest current realizations of assemblers operating with atomic precision are tip-based scanning probe devices that are digitally controlled; their nano/bio interface is indeed a standard IT man–machine interface. This kind of control will, however, be far too cumbersome for assemblers that are themselves nanosized—the lags in regulation would tend to generate chaos—hence the assemblers will need to operate with a great deal of autonomy. Although the current generation of screen-based graphical user interfaces (GUI) might be slightly more convenient than punched tape or cards, the laborious letter by letter entry of instructions or data via a keyboard remains slow, frustrating and error-prone. While hardware continues to advance exponentially (Moore's law), software and the man–machine interface continue to lag behind and limit human exploitation of IT.

Also encompassed within this meaning of the nano/bio interface is how a nanomachine can produce at the human scale. For example, a single nanostructured microsystem may synthesize only attograms of a valuable medicinal drug, too little even for a single dose administered to a patient. The solution envisaged is scaleout, i.e., massive parallelization—the chemical reactor equivalent of parallel computing. This has already been realized, albeit on a modest scale involving less than a dozen microreactors.

4.1 THE "PHYSICAL" NANO/BIO INTERFACE

From the biological viewpoint, the nano/bio interface can be considered at three scales at least (examples given in parentheses):

1. the organismal scale (e.g., wearing clothing made from a nanomaterial)
2. the cellular scale (e.g., cell adhesion to a nanomaterial)
3. the molecular scale (e.g., protein adsorption to a nanomaterial).

There is an additional interface above the organismal scale, namely between society and nanotechnology, which comprises several unique features and will be dealt with separately (Chapter 12). The nano/bio interface can be considered a special case of the living–nonliving (or bio–nonbio) interface, in which the nonliving side has nanoscale features. If it is a nanomaterial (a nanostructured material or a nano-object),

then many of the issues have been dealt with in a general way under the rubric of bio-compatibility. If it is a nanodevice, then the issues are likely to belong to metrology (Chapter 5) or to bionanotechnology (Chapter 11). If it is a nanosystem, then effectively it will be interacting with a system on the living side, and we are again in the realm of the interface between society and nanotechnology (Chapter 12). The focus in the following sections is how the presence of nanoscale features introduces unique behavior not present otherwise.

4.1.1 **Organisms**

It is a moot point whether nanostructure is perceptible by organisms. Certainly it is not directly visible, given that the organs of sight use visible light. We have rather to ask whether we could otherwise become aware of (e.g., feel) a difference between a textile woven from a nanostructured fiber and one that was not, for example. Indirectly there might be no difficulty; a cravat whose fibers are nanostructured to make them water-repellent would be immediately distinguishable using the naked eye from an ordinary one. Natural fibers of course have structure at the nanoscale, but they are not nanostructured in the sense of being deliberately engineered at that scale. Similar considerations apply to foodstuffs: two aliments of the same overall composition may well feel different when first placed in the mouth if they have different structure at the nanoscale, but if this structure has not been deliberately engineered then we are not talking about nanotechnology according to the definitions (Section 1.1). Experiments to determine whether deliberately differently engineered structures are differently perceived do not yet appear to have been carried out in a sufficiently systematic fashion to be useful; indeed this aspect of the nano/bio interface is, as yet, practically unexplored.

The study of the deleterious biological consequences of nano-objects, especially nanoparticles, penetrating into the body constitutes the field of nanotoxicology (see Section 4.3). Particles may gain access to the body by inhalation, ingestion, through the skin, or following the introduction of material in some medical procedure such as implantation. If they are introduced into the blood they are likely to be taken up by the macrophage system (see also Section 4.1.3). Toxic, or at any rate inflammatory, effects may arise indirectly through the adsorption of blood proteins on the surface of a particle and their subsequent denaturation (see Section 4.1.4). In other words, the effects on the overall organism depend on processes involving suborganismal scales down to that of molecules.

4.1.2 **Tissues**

The tissue–substratum interface has been intensively investigated for many decades, starting long before the era of nanotechnology, in the search to develop more biocompatible orthopedic and dental implant materials. It has long been realized that a rough prosthesis will integrate better into existing bone than a smooth one. Contrariwise, an ultrasmooth stent is known to be less prone to adsorb proteins from the bloodstream

than a rough one. This knowledge has not yet, however, been systematically refined by correlating nanoscale morphological features with the rate and quality of assimilation. Although surfaces have been treated so as to incorporate certain molecules into them, this has been done in a (bio)chemical fashion rather than in the spirit of atomically precise engineering, and there has been no comprehensive systematic study of nanoscale features evoking a response not seen otherwise, and disappearing if the features are enlarged up to the microscale.

The tissular nano/bio interface is of acute interest to the designers of nanoparticles for drug delivery. The main general challenge is achieving long circulation times of the nanoparticles in the bloodstream, which is in turn achieved by evading opsonization. This is thus a problem of protein adsorption, dealt with in detail in Section 4.1.4. There is also the question of how the nanoparticles are eliminated after they have done their work. Since solubility increases with increasing curvature according to the Gibbs–Thomson law (equation 2.4), the smaller the particle radius, the faster it will dissolve. Designers of surgical instruments are often concerned with tribology aspects. Although naively it might seem that a perfectly smooth surface would have the lowest coefficient of friction, asperity serves to diminish contact area and, hence, may even be advantageous. Nano-engineered asperity has, as yet, been incompletely explored as a route to controlling the friction experienced by objects in contact with tissue. An interesting research direction seeks to harness biological lubricants, notably the ubiquitous glycoprotein mucin, in order to control friction [170]. A further aspect of the tissue–nanomaterial interface is biomineralization. The tough, hard shells of marine organisms such as the abalone have a composite structure essentially consisting of mineral nanoplates cemented together by a protein matrix [87]. The exact mechanism of their formation has yet to be elucidated but presumably is essentially an extracellular process. Bone is apparently constructed on similar principles, although it has a more sophisticated hierarchical structure.

4.1.3 Cells

This section considers two themes: (1) a nano-object small enough to penetrate inside a cell; and (2) the response of a supported cell to nanostructured features of its substratum (support).

Penetration of Nano-Objects into the Cell's Interior

It has been a long-standing frustration of electrophysiologists that the electrodes typically used to probe electrical activity in the interior of the brain are gigantic in comparison with the size of neurons and dendrites, hence observations not only imply gross damage due to initial installation of the electrode, but also that signals are recorded simultaneously from dozens of neurons. The availability of carbon nanotubes, which can be excellent electrical conductors (Section 9.2), offers the possibility of creating an electrode considerably smaller than a single cell. This has the potential to revolutionize experimental neurophysiology. The main practical obstacle is the difficulty of manipulating the nanotubes: they must be attached to some

kind of back plate. If the nano-object is freely floating, such as the drug delivery nanoparticle already referred to, the main problem is controlling uptake. The tiniest nanoparticles might be able to enter the cell through some of the larger transmembrane channels. Larger nanoparticles may be engulfed (e.g., by macrophages) as if they were viruses. They may also become wrapped by the outer lipid membranes of the cell, as has been observed in a model system [117]. The penetration of a drug-bearing nanoparticle into the cell's interior may be precisely the goal of the particle designer, but nanoparticles used in other applications for cosmetic purposes, for example as ultraviolet-absorbing sunscreen applied to the skin, may, by similar mechanisms, also end up inside the cell where they may have toxic consequences. The investigation of these consequences constitutes what is currently the most active part of the field of nanotoxicology (Section 4.3).

The Response of a Cell to External Nanostructured Features

The fundamental premise engendering interest in living cells interacting with nanostructured surfaces is that there is strong evidence that cell surfaces are themselves heterogeneous at the nanoscale, whence the hypothesis that by matching artificial substratum patterns with natural cell surface ones, a degree of control over the cell can be achieved. "Patterns" are here to be understood as denoting statistical regularity only. Perhaps the degree of regularity of cell surface features determining the binding to substrata is intermediate between wholly regular and random; that is, maximally complex (Figure 4.2).

While some responses may have a purely physical nature, at least in part (the characteristic spreading, i.e., the transformation from sphere to segment, Table 4.1, shown by some cell types placed on a planar surface could result from cell viscoelasticity opposing adhesion to the surface), others involve cellular intelligence: it is an adaptive response, the mechanism of which involves the reception of environmental information by the cell surface and the transmission of that information to the cell's
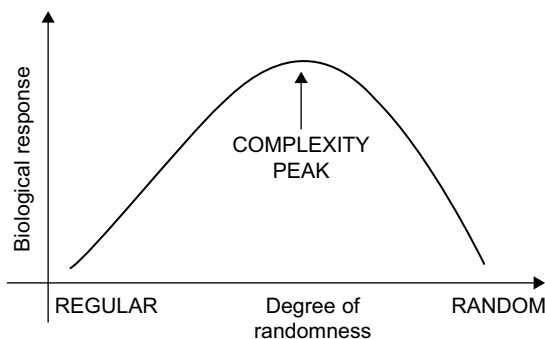
FIGURE 4.2

Proposed relationship between environmental complexity and biological response.

Table 4.1 Cell environmental responses, in roughly increasing order of complexity. See Section 5.5 for the techniques used to measure the responses

Response	Level	Timescale	_a_
Adhesion	Energetic	s	1,2
Spreading (morphology)	Energetic	min	1,2
Growth alignment	Energetic?	h	1
Microexudate secretion	Gene expression	min	2,3
Growth factor secretion	Gene expression	min	3
Alteration of metabolism	?	?	3
Differentiation	Gene expression	days	1,2,3,4
Speciation (i.e., cancer)	Chromosome rearrangement	years	1,2,3,4

a Techniques useful for determining responses are: 1, microscopy (usually optical, but may include scanning probe techniques); 2, nonimaging interfacial techniques (e.g., optical waveguide lightmode spectroscopy; 3, biochemical techniques (e.g., immunocytochemistry); 4, nucleic acid arrays.

genome, following which action in the form of activating or deactivating the synthesis of certain proteins (i.e., changes in gene expression) results. These proteins then engender a certain response. Sometimes it is the unique history of cell–environment interactions that determines cell behavior.

Eukaryotic Cells

This hypothesis of maximum complexity (Figure 4.2) originated in the many observations that the behavior of cells depends on the nature of the basement membrane supporting them. A classic example is the different patterns of neurite outgrowths from neurons supported on different extracellular matrix materials such as laminin and tenascin. An early example of cells brought into contact with artificial materials were the experiments of Carturan et al. on immobilizing yeast within inorganic gels [30]. Eukaryotic cells are typically a few to several tens of μm in diameter. They are enveloped by a lipid bilayer (the plasmalemma) and the shape is controlled by the cell itself. In suspension they tend to adopt the shape of lowest surface:volume ratio (viz., a sphere) as expected from purely mechanical considerations but on a solid surface tend to spread (i.e., transform into a segment). There is already a vast literature on the interaction of individual living cells with *microstructured* surfaces, defined as having features in the range 100 nm–100 μm. The main result from this large body of work, is that the cells tend to align themselves with microscale grooves. The reader appraising this literature should be warned that sometimes these surfaces are referred to as "nano" without any real justification (since the features are larger than nanoscale).

The goal of investigations of the nano/bio interface could be to demonstrate that unique features of cell response arise as feature sizes fall below 100 nm (taking the consensual definition of the upper limit of the nanoscale). Surprisingly, this has hitherto been done in only a very small number of cases. One of the few examples is the work by Teixeira et al. [160], who showed that when the ridges of a grooved

pattern were shrunk to 70 nm, keratocyte alignment was significantly poorer than on ridges 10 times wider (whereas epithelial cell alignment was unchanged). Alternatively, following the spirit of Chapter 2, one could define one or more "biological response nanoscales" by presenting cells with a systematically diminishing set of structural features and observing at what scale the cells started to respond. Alignment is of course only one, perhaps the most trivial, of the huge variety of measurable responses. Table 4.1 summarizes some of the others.

A rather diverse set of observations may be systematized by noting that the fundamental response of a cell to an environmental stimulus is an adaptive one. Following Sommerhoff [154], one can distinguish three temporal domains of adaptation: behavioral (short-term); ontogenic (medium-term); and phylogenetic (long-term). Although Sommerhoff was discussing organisms, these three domains have their cellular counterparts. At the cellular level, behavioral adaptation is largely represented by energetically driven phenomena [101], perhaps most notably the spreading transition characteristic of many cells adhering to substrata (Figure 4.3). Morphological changes have been shown [91], and the rate of spreading has been quantitatively shown, to depend on nanoscale features of the substratum [8].

The secretion of microexudates, growth factors, etc. would appear to involve the reception of chemical and/or morphological signals by the cell surface and their transmission to the nucleus in order to express hitherto silent genes. Indeed, this secretion appears to be part of an adaptive response to condition an uncongenial surface. Work in molecular biology has established that the amino acid triplet arginine–glycine–aspartic acid (RGD), characteristically present in extracellular matrix proteins such as fibronectin, is a ligand for integrins (large transmembrane receptor molecules present in the cell membrane), binding to which triggers changes depending on cell type and circumstances. Such phenotypic changes fall into the category of ontogenic adaptation, also represented by stem cell differentiation. Recently reported work has shown

FIGURE 4.3

Idealized cell (initial radius R) spreading on a substratum F in the presence of medium C. Upper panel: cross-sections, showing the transition from sphere to segment (h is the ultimate height of the cell). Lower panel: plans, as might for example be seen in a microscope (the dark region shows the actual contact area).

that 30-nm diameter titania nanotubes promoted human mesenchymal stem cell adhesion without noticeable differentiation, whereas 70–100-nm diameter tubes caused a ten-fold cell elongation, which in turn induced cytoskeletal stress and resulted in differentiation into osteoblast-like cells [126]. Phylogenetic adaptation involves a change of genetic constitution, and would correspond to, for example, the transformation of a normal cell to a cancerous one. There are well-established examples of such transformations induced by certain nanoparticles and zeolites. Two important criteria for deciding whether a nanoparticle might be carcinogenic seem to be:

1. Whether the particle is insoluble (this does not, of course, apply to nanoparticles made from known soluble carcinogens active when dispersed molecularly).
2. Whether the particle is acicular and significantly longer than macrophages (see Figure 4.3).

If both these criteria are fulfilled the macrophages ceaselessly try to engulf the particles and they become permanent sites of inflation, which may be the indirect cause of tumorigenesis. Nevertheless, given the long induction periods (typically decades) that elapse between exposure to the particles and the development of a tumor, elucidation of the exact mechanism remains a challenge. Incidentally, there appear to be no known examples of nanostructured substrata that directly induce cancer through contact. However, it should be kept in mind that the mechanism for any influence may be indirect (as in [126]; note also that an extracellular matrix protein might bind to a certain nanotextured substratum, change its conformation—see Section 4.1.4—and expose RGD such that it becomes accessible to a cell approaching the substratum, triggering ontogenic changes that would not have been triggered by the protein-free substratum) and is generally unknown. The secretion of macromolecular, proteinaceous substances that adsorb on the substratum is an especially complicating aspect of the nano/bio interface, not least because any special nanostructured arrangement initially present is likely to be rapidly eliminated thereby.

Prokaryotic Cells

In contrast to eukaryotes, prokaryotes (archaea and bacteria) are typically spherical or spherocylindrical, smaller (diameters are usually a few hundred nm) and enveloped by a relatively rigid cell wall predominantly constituted from polysaccharides, which tends to maintain cell shape. Prokaryotes have received rather less attention than eukaryotes, although the interaction of bacteria with surfaces is also a topic of great importance for the biocompatibility of medical devices, which is normally viewed in terms of adhesion rather than adaptation. Nevertheless, if one considers bacterial communities, the formation of a biofilm, based on a complex mixture of exudates, and which usually has extremely deleterious consequences, should be considered as an adaptive response at least at the ontogenic level, since the expression pattern of the genome changes significantly. A worthy (and hitherto unreached) goal of investigation, therefore, is whether one can prevent biofilm formation by nanostructuring a substratum.

4.1.4 **Biomolecules**

A protein is comparable in size or smaller to the nanoscale features that can nowadays be fabricated artificially. The problem of proteins adsorbing to nonliving interfaces has been studied for almost 200 years and an immense body of literature has been accumulated, much of it belonging to the realm of physical chemistry and without any specific nanoscale ingredient. A good deal of the phenomenology can be satisfactorily interpreted on the basis of the Young–Dupré equations (Chapter 3), which allow one to link experimentally accessible single substance surface tensions (e.g., via contact angle measurements on protein adsorbents (substrata), and on thin films carefully assembled from proteins) to interfacial energies. Wetting is a typical mesoscale phenomenon with a characteristic length ~30 nm, thus averaging out much molecular detail. This approach has allowed the systematization of a great deal of data for different proteins adsorbing on different substrata in the presence of different liquid media, rationalizing the interaction in terms of the interfacial free energy ΔG_{123}, the subscripts 1, 2, and 3 denoting adsorbent, liquid medium, and adsorbate respectively. As schematized in Figure 4.4, the adsorption process involves firstly surmounting a repulsive energy barrier of height ΔG_a (the profile of which determines the adsorption kinetics, see Chapter 3), followed by residence at the interface in a potential well of depth ΔG_b. As complex objects, proteins typically undergo changes during residence on the surface, such as dehydration of their zone of contact with the substratum and/or denaturation.

The model represented by Figure 4.4 was developed from thermodynamic principles without special assumptions regarding the structures of adsorbent, adsorbate and intervening medium. A clue that this model is too simplistic to represent reality was actually already discovered over 100 years ago by Hofmeister. A nanoscale approach to the proteinaceous nano/bio interface takes cognizance of the following:

1. Details of the molecular structure on the interface, with explicit recognition of the solvent (water), need to be taken into account.
2. The surface tensions of highly curved features (e.g., nanoscale ridges) will differ from the values associated with planar surfaces.
3. Proteins typically have highly heterogeneous surfaces at the nanoscale [29].

This cognizance may be termed the (bio)physical chemistry of the nano/bio interface. Item (1) leads to the Hofmeister effect [39], items (2) and (3) may lead to a different balance of forces and, as with living cells, it has been hypothesized (and demonstrated) that matching protein heterogeneity with artificial substratum nanoscale heterogeneity leads to anomalous behavior in protein adsorption [2]. At present, there is very little theoretical prediction of phenomenology at the nano/bio interface. Although the behavior of a single protein approaching a substratum can nowadays be considered to be reasonably well understood and predictable, real biomedical problems involve a multiplicity of proteins. It is well-known that substrata exposed to blood experience a succession of dominantly adsorbed proteins (the Vroman effect); until now this has not been comprehensively investigated using nanostructured substrata, and indeed to do so purely empirically without any guiding theory would
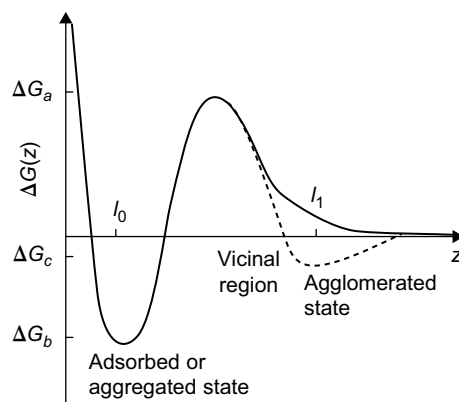
FIGURE 4.4

Sketch of the interfacial interaction potential $\Delta G_{123}(z)$ experienced by a protein (or other nano-object) approaching a substratum. The potential is the sum of different contributions, individually varying smoothly and monotonically, and the actual shape depends on their relative magnitudes and decay lengths. In this hypothetical (but typical) example, at moderate distances z from the substratum the net interaction is repulsive, dominated by long-range hydrophilic repulsion (at low ionic strength, electrostatic repulsion might be dominant). Sometimes (as shown by the dashed portion of the curve) a secondary minimum appears; low-energy objects unable to surmount the barrier ΔG_a may reside at a separation ℓ_1. At short distances, the attractive Lifshitz–van der Waals interaction dominates; adsorbed objects reside at a separation ℓ_0. At very short distances the Born repulsion dominates. Further explanation is given in the text.

be a daunting task. As part of this research direction, one should include the phenomenon of the protein "corona" hydrodynamically associated with a nano-object suspended in a proteinaceous medium (e.g., blood), or with a surface exposed to such a medium. This corona can be expected to shift its composition as initially adsorbed proteins are exchanged for others.

The mechanism of such *protein exchange* processes, in particular their dependence on the interfacial free energies, is still very imperfectly understood, but it seems (at least, according to present knowledge, there is no reason to think otherwise) that by appropriately modifying the surface tensions and taking into account the heterogeneities of both adsorbate and adsorbent (possibly by simply summing all possible combinations of pairwise interactions) one would be able to correctly predict the entire phenomenology, including toxicology aspects. Once in residence on the surface, the protein may exchange its intramolecular contacts for protein–substratum contacts (Figure 4.5), without necessarily any change of enthalpy, but the entropy inevitably increases because the extended, denatured conformation occupies a much larger proportion of the Ramachandran map than the compact, native conformation.
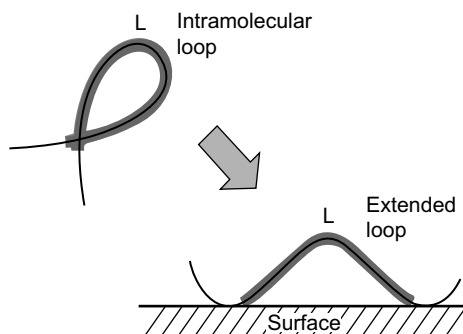
FIGURE 4.5

Surface-induced protein denaturation, showing the substitution of the intramolecular contacts in the native conformation by substratum–protein contacts and the concomitant transition from a compact globular conformation to an extended denatured one.

4.2 NANOMEDICINE

One of the most important manifestations of the nano/bio interface is the application of nanotechnology to that branch of applied biology known as medicine. Mention has already been made of Feynman's inclusion of "microscopic surgeons" in his vision of what came to be called nanotechnology [56]. The dictionary definition of medicine is "the science and art concerned with the cure, alleviation, and prevention of disease, and with the restoration and preservation of health". As one of the oldest of human activities accessory to survival, it has of course made enormous strides during the millennia of human civilization. Formally it was well captured by Hippocrates' dictum, "Primum nil nocere"; during the past few hundred years, and especially during the past few decades, it has been characterized by an enormous technization, and the concomitant enormous expansion of its possibilities for curing bodily malfunction. The application of nanotechnology, the latest scientific–technical revolution, is a natural continuation of this trend; the definition of nanomedicine is the application of nanotechnology to medicine.

4.2.1 A Concept System for Nanomedicine

One might usefully divide the applications of nanotechnology to medicine into "auxiliary" and "direct". The three main activities encompassed by the former are: (1) any application of nanotechnology assisting drug discovery; (2) any application of nanotechnology assisting drug manufacture; and (3) the indirect application of nanotechnology to medical diagnostics. The main part of the third activity is taken up by the enhanced processing power of nanoscale digital information processors, applied to areas such as pattern recognition and telemedicine, which might be defined as diagnosis and treatment by a medical doctor located geographically remotely from the

patient. The first activity covers a very broad spectrum, including superior digital data processing capability (which impacts laboratory information management systems (LIMS) and laboratory automation) as well as nanobiology (see Appendix, p. 247 for the definition). The second is mainly concerned with miniature—albeit micro rather than nano—chemical mixers and reactors; it is hoped that the microsystems paradigms for chemical synthesis will make it feasible to generate quasicustomized versions of drugs, if not for individual patients at least for groups of them. It would be appropriate to include under this heading advanced materials, many of which are nanocomposites with superior barrier properties, used for packaging drugs.

The latter ("direct") encompasses *in vitro* and *in vivo* applications. The most important *in vitro* applications are: materials ("tissue scaffolds") for helping to grow, in culture, tissues destined for implanting in the human body (in so far as this depends on superior understanding of the behavior of living cells, it can be said to be part of nanobiology); and "labs-on-chips"—miniature mixers, reactors, etc. for carrying out analyses of clinical fluids. A typical lab-on-a-chip belongs to microsystems rather than nanotechnology, however, and is therefore outside the scope of this book, although there may be nanotechnology incorporated within the microfluidic channels (e.g., a nanoparticulate catalytic surface lining a tube), in which case one can speak of micro-enabled nanotechnology. A very interesting area is the development of miniature DNA sequencing devices based on the physical, rather than the chemical, properties of the nucleotides [175]. It might be recalled that an early hope of that quintessentially "nano" metrology device, the atomic force microscope (Section 5.1), was to determine the base sequence of DNA by directly imaging the polymer, at sufficient resolution to be able to distinguish the four different bases by shape. This has turned out to be far more difficult than originally envisaged, but other methods based on examining a single or double DNA chain are being invented and examined.

In vivo applications encompass nanostructured materials, such as materials for regenerative medicine (which might well be rather similar to the tissue scaffolds used *in vitro*) and materials for implantable devices, which might include prostheses for replacing bones and stents for keeping blood vessels open [145]. Nanotechnology is attractive for implants because, through creating novel combinations of material properties, it should be possible to achieve performances superior to those available using traditional materials.

The most active current field of nanomedicine is the use of nanoparticles of progressively increasing sophistication. The two main classes of applications are as contrast agents (mainly for imaging purposes but more generally for diagnostics) and as drug delivery agents. In both cases the particles are administered systemically. The major challenge is to ensure that they circulate long enough before opsonization to be able to accumulate sufficiently in the target tissue. If the particles are superparamagnetic (see Section 2.6) they can be steered by external electromagnetic fields, accelerating their accumulation at the target. Since cancer cells typically have a significantly faster metabolic rate than that of host tissue, substances will tend to be taken up preferentially by cancer cells, assuming that there is a pathway for uptake. In other

cases, and also to enhance the uptake differential between cancerous and noncancerous tissue, the nanoparticles are "functionalized"; that is, their surface is specially treated in some way (e.g., by coating the surface with a ligand for a receptor known to exist in the target cells).

The simplest medicinal nanoparticles are those made of a single substance, for example magnetite, which, when taken up by the tissue, can be subjected to an external electromagnetic field such that they will absorb energy, become hot, and transfer the heat to the cells surrounding them, thereby killing them (necrosis). The particles can themselves be considered as the drug, which requires the external field for activation. Other examples of single-substance drug particles are gold and silver ones. Nanoparticulate (potable) gold was already introduced as a therapeutic agent by Paracelsus about 500 years ago. Nanoparticulate silver is nowadays widely used (e.g., in toothpaste) as a general antibacterial agent.

At the next level of sophistication, the particles constitute a reservoir for a small organic-molecular drug. Prior to use the particles (which may contain numerous pores of an appropriate nature) are saturated with the drug, which starts to dissociate once the particles are introduced into the human body. At the current level of the technology, this approach works best if the characteristic dissociation time is long compared with the time needed to become accumulated at the target. Hemoglobin is the paradigm for this type of particle but is still far from having been successfully imitated by an artificial device (see Chapter 11).

Slightly more sophisticated are particles in which the reservoir is initially closed, the closure being designed to resist opening under conditions likely to be encountered before final arrival at the target. An example is a calcium carbonate nanoshell enclosing a stomach drug designed to be administered orally. The calcium carbonate will dissolve in the acidic conditions of the stomach, thereby releasing the drug. This is an example of a responsive particle, which might even be called adaptive under a highly restrictive set of conditions. The ideal "smart" adaptive nanoparticle will: sense its surroundings for the presence of its target; bind to the target when it reaches it; sense its surroundings for the local concentrations of the therapeutic agent it carries; and release its burden if the surrounding concentration is low. At present, most research in the field is concerned with devising novel adaptive materials for drug delivery (nanostructure is especially useful here because it allows multiple attributes to be combined in a single material).

Any system based on a reservoir has a finite capacity and, hence, therapeutic lifetime. For many therapeutic regimens this may not be a disadvantage, since the drug may only be required for a limited duration. For long-term therapy, the goal is to create nano-objects able to manufacture the drug from substances they can gather from their immediate environment (i.e., veritable nanofactories).

The ultimate in sophistication of the nano-object is the nanoscale robot or "nanobot". Microscopic or nanoscopic robots are an extension of existing ingestible devices that can move through the gastrointestinal tract and gather information (mainly images) during their passage. As pointed out by Hogg [77], minimal capabilities of such devices are: (chemical) sensing; communication (receiving information

from, and transmitting information to, outside the body, and communication with other nanobots); locomotion—operating at very low Reynolds numbers (around 10^{-3}), implying that viscosity dominates inertia; computation (e.g., recognizing a biomarker would typically involve comparing sensor output to some preset threshold value; due to the tiny volumes available, highly miniaturized molecular electronics would be required for constructing on-board logic circuits of practically useful data processing power); and of course power—it is estimated that picowatts would be necessary for propelling a nanobot at a speed of around 1 mm/s. Such nanobots could also incorporate one or more drug reservoirs, or miniature devices for synthesizing therapeutic substances using materials found in their environment. It is very likely that to be effective, these nanobots would have to operate in large swarms of billions or trillions. Figure 4.6 sketches a putative nanobot.

In vivo applications of nanotechnology also encompass the implantation of devices for sensing biochemicals (physiological markers). The smaller the device, the more readily it can be implanted. The technology involved is basically the same as that being developed for nanobots.

It may well be that the most important future role for nanobots will be to carry out tricky repairs that at present require major surgery; for example, clearing plaque from the walls of blood vessels or renewing calcified heart valves. The body has wonderful repair mechanisms of its own but for some reason certain actions appear to be impossible.
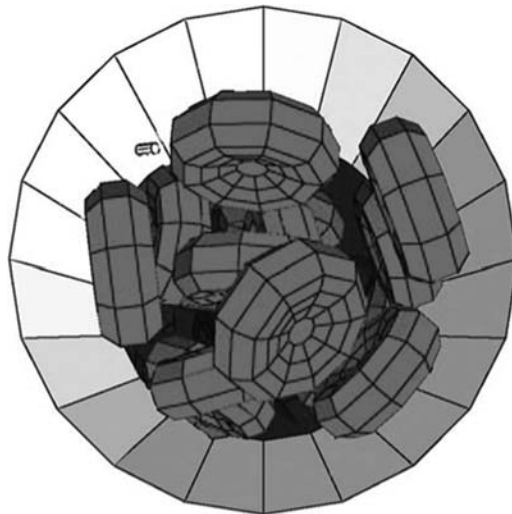
FIGURE 4.6

Sketch of the likely form of a future nanobot (the small bacterium-sized cylinder in the upper left quadrant) drawn to scale in a blood vessel containing erythrocytes. Reproduced with permission from [77].

4.2.2 **Wider Implications**

It is customary nowadays to take a global view of things, and in assessing the likely impact of nanotechnology on medicine this is very necessary. Nanotechnology is often viewed to be the key to far-reaching social changes, and once we admit this link then we really have to consider the gamut of major current challenges to human civilization, such as demographic trends (overpopulation, aging), climate change, pollution, exhaustion of natural resources (including fuels), and so forth. Nanotechnology is likely to influence many of these, and all of them have some implications on human health. Turning again to the dictionary, medicine is also defined as "the art of restoring and preserving health by means of remedial substances *and the regulation of diet, habits, etc.*" With this in mind, it would be woefully inadequate if the impact of nanotechnology on medicine were restricted to consideration of the development of the topics discussed in the preceding parts of this section.

4.3 **NANOTOXICOLOGY**

There is currently great concern about the potential adverse impacts of nano-objects on human health. Nanotoxicology is defined as the study of the toxicology of nanomaterials; while allergic reactions to certain morphologies and combinations of atoms cannot be ruled out, overwhelming attention is currently given to nano-objects, because they can be inhaled or ingested or otherwise enter the body, triggering inflammatory responses. Clearly the field is part of nanomedicine; on the one hand the developers of nanoparticle-based therapeutics need to respect the Hippocratic doctrine of "Primum nil nocere" (a major issue is what happens to the billions or trillions of systemically introduced nanoparticles after they have accomplished their therapeutic task); and on the other hand the possible effects of exposure to nanoparticles on those engaged in preparing products containing them is an important part of occupational medicine.

The toxicity of nano-objects falls into two categories: they can be considered as chemicals with enhanced reactivity due to their small size (see Section 2.4), hence if the substance from which they are made is toxic or can become toxic in the body, the nano-object is likely to be toxic—as an example consider mercury (Hg), which if ingested as millimeter-sized droplets typically passes through the body unmetabolized, but which in nanoparticulate form is likely to be readily oxidized to highly toxic Hg^+ or Hg^{2+}; or they are not toxic *per se*, but initiate an immune response, ranging from mild temporary inflammation to persistent chronic inflammation and may be ultimately carcinogenesis.

The mechanism of the latter category is denaturation of proteins adsorbed onto the nano-object from its environment (see Figure 4.5), which may be the cytoplasm, intracellular space, or a fluid such as blood (whether denaturation actually takes place depends on the chemical nature of the nano-object's surface [129]). The denatured protein-coated nano-object is then identified as nonself by the immune system, which initiates the processes that should lead to elimination of the foreign object. One of the chief problems is caused by nanofibers. If insoluble and excessively elongated,

as is blue asbestos and most carbon nanotubes, they can be neither solubilized nor ingested by the macrophages (Figure 4.7), which nevertheless persist indefinitely in a futile attempt to destroy them, engendering sickness.

Although most attention is focused on the potential dangers of manufactured nano-objects inhaled, ingested or adsorbed through the skin, two other sources may be of greater significance. One is wear of implanted prostheses such as hip replacements, which generates micro- amd nanoparticles in abundance (e.g., Figure 4.8). The other is the environment, which contains both natural nanoparticles such as the ultra-fine component of desert sand, salt aerosols produced by the evaporation of sea spray and volcanic ash, and artificial but unintentionally produced aerial nanoparticles such as the products of combustion.

Note that our planet has an oxidizing atmosphere, and has had one probably for at least 2000 million years. This implies that most metals, other than gold, platinum and so forth (the noble metals), will be oxidized. Hence, many kinds of metallic nanoparticles will not be stable in nature.

A key issue is whether our human defence mechanisms can adapt to cope with new nano-objects emanating from the nanotechnology industry. It seems that they cannot in the case of long insoluble fibers (e.g., amphibole asbestos and carbon nanotubes). In the case of other nano-objects such as smoke and mineral particles the dose (concentration, duration, intermittency, etc.) and accompanying lifestyle seem to play a primordial role.

Although a great deal of work has been (for a review see, e.g., [148]) and continues to be done on the biological effects of nano-objects, the present effort seems both uncoordinated (with much duplication of effort worldwide alongside the persistent presence of significant gaps in knowledge) and misguided. Firm general principles as adumbrated above already seem to have been established, and there is no real need to verify them *ad nauseam* by testing every conceivable variant of nano-object. Such tests are easy to do at a basic level (the typical procedure seems to be: introduce
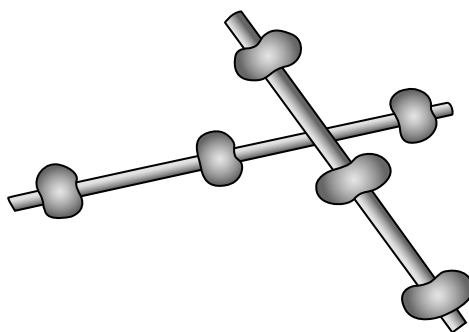
FIGURE 4.7

Sketch of a typical scanning electron micrograph of two amphibole asbestos needles within the pleural cavity occupied by several macrophages. Reproduced with permission from [128].
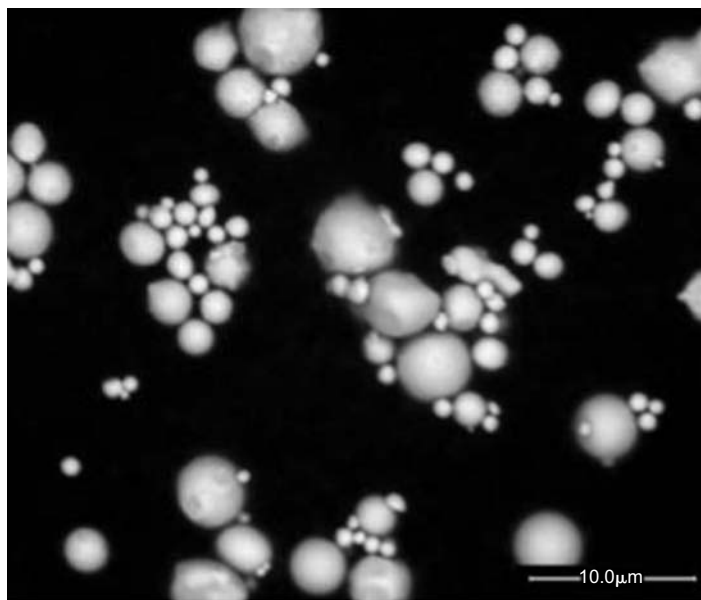
FIGURE 4.8

Scanning electron micrograph of CoCr particles within tissues adjacent to a human implant, retrieved at revision surgery. Reproduced with permission from [148].

a vast quantity of nano-objects into the test species, observe behavior, physiological variables, etc., and finally establish in which cells or tissues the nano-objects were accumulated, if not excreted). To arrive at meaningful results, however, far more attention needs to be paid to chronic exposure under realistic conditions. Furthermore, given that we already know that the occurrence of tumors traceable to exposure to nano-objects typically manifests itself after a delay of several decades, tests need to be long-term. The increasingly dominant mode of funding scientific research in the developed world, namely project grants lasting for two or three years, is not conducive to the careful, in-depth studies that are required to achieve real advances in understanding (neither in nanotoxicology nor, it may be remarked, in other fields of scientific endeavor); rather, this mode encourages superficial, pedestrian work with easily foreseen outcomes. Hopefully organizations like the National Institutes of Health in the USA will continue to enable a scientist to devote a lifetime of truly exploratory work to his or her chosen topic.

4.4 SUMMARY

The nano/bio interface comprises three meanings. First is the conceptual one: the "living proof of principle" that nanoscale mechanisms (the subcellular molecular

machinery inside a living cell) exist and can function. Within this meaning there is also an inspirational aspect: living cells are known to distinguish between natural structures differing from one another at the nanoscale, suggesting that artificial mimics may also be used to invoke specific living responses. Second is the man–machine interface aspect: how can humans control atomic-scale assembly? Conversely, how can atomic-scale assembly be scaled up to provide artifacts of human dimensions? Third is the literal physical boundary between a living organism and a nanomaterial, device or system. This applies both to nanobiotechnology (the application of nanotechnology to biology, e.g., implantable medical devices) and bionanotechnology (the use of biological molecules in nanodevices, i.e. the topic of Chapter 11). This "bio–physical" interface has several characteristic scales from the biological viewpoint: organismal, cellular and biomolecular. Each scale is examined, considering the nano/bio as a special case of the general problem of the bio–nonbio (living–nonliving) interface. Perhaps the most important manifestation of nanobiotechnology is nanomedicine. This is defined and discussed, concluding with the toxicology of nano-objects. The immense possibilities of this branch of nanotechnology includes such things as implanted inertial tracking devices providing feedback from trunk and leg muscle movements, allowing one to regulate one's posture correctly.

4.5 **FURTHER READING**

R. Arshady, K. Kono (Eds.), Smart Nanoparticles in Nanomedicine, Kentus Books, London, 2006.

J.J. Ramsden, et al. The design and manufacture of biomedical surfaces, Ann. CIRP 56/2 (2007) 687–711.

P.A. Revell, The biological effects of nanoparticles, Nanotechnol. Percept. 2 (2006) 283–298.

This page intentionally left blank

Nanometrology

CHAPTER CONTENTS

The ultimate goal of nanometrology is to provide the coordinates and identity of every constituent atom in a nano-object, nanostructured material or nanodevice. This goal raises a number of problems, not least in the actual representation and storage of the data. Thanks to techniques such as X-ray diffraction (XRD), one can readily determine atomic spacings to a resolution of the order of 0.1 nm and this information, together with the external dimensions of an object, effectively achieves our goal provided the substance from which the object is made is large, monolithic and regular. This achievement must, however, be considered as only the beginning

since nanomaterials (and nanodevices considered as complex nanomaterials) may be fabricated from a multiplicity of substances, each one present as a domain of unique and irregular shape. Techniques such as XRD require averaging over a considerable volume to achieve an adequate signal to noise ratio for the finest resolution—hence the stipulations that the object being examined be large and regular. As an illustration, consider the challenge of reverse engineering a very large-scale integrated circuit on a "chip"—bearing in mind that this is essentially a two-dimensional structure. X-ray diffraction of the chip could not yield useful information for the purpose of making an exact replica of the circuit. The challenge of nanometrology is to achieve atomic scale resolution for such arbitrary structures (the chip is not of course truly arbitrary—the structures form a functional circuit; but knowledge of the function alone would not suffice to help one reconstruct the details, not least because there are likely to be several hardware routes to achieving comparable function).

Any successful manufacturing technology requires appropriate metrology, and the atomic-scale precision implicit in nanotechnology places new demands on measuring instruments. Reciprocally, the development of the quintessential nanometrology tool, namely the scanning probe microscope (better called an ultramicroscope or nanoscope, see Appendix, p. 248-9) has powerfully boosted nanotechnology itself, since these instruments have become the backbone of efforts to develop bottom-to-bottom manufacturing procedures (Section 8.3). The focus of this chapter is the nanometrology of surfaces. For more general issues regarding ultraprecision measurement instrumentation, the reader may refer to the recent book by Leach (see Section 5.7).

Morphology and chemistry are not independent at the nanoscale. Depending on how it is cut, even the planar face of a crystal of a binary compound MX can vary dramatically from pure M to pure X. "Roughness" or texture at the nanoscale may actually be constituted from an intricate array of different crystal facets. The chemical effect of this morphology depends on the characteristic length scale of the phenomenon being investigated. Living cells, for example, are known to be highly sensitive to the crystallographic orientation of a substratum. This has been demonstrated by cell growth experiments on single crystals: epithelial cells attached themselves and spread only on the (011) faces of calcium carbonate tetrahydrate and not on the (101) faces within tens of minutes following initial contact, but after 72 hours all cells on the (011) faces were dead, but well-spread and living on the (101) faces [70]. However, it should be noted that these two faces mainly differ in the surface distribution of the lattice water molecules, to which the living cell may be especially sensitive. Note that cells actively secrete extracellular matrix (ECM) proteins when in contact with a substratum, which are then interposed to form a layer between the cell and the original substratum material; hence the observed effects could have been due to the different conformations adopted by these ECM proteins due to the different chemistries and morphologies of the different crystal faces (cf. Section 4.1.4).

Dynamic measurements are likely to become of increasing importance as process as well as structure also falls under the scrutiny of the metrologist. *In situ* techniques are required here; rastering is also to be avoided.

Metrology of the nano/bio interface presents challenges of immense complexity, often requiring completely different approaches from those developed for wholly inanimate systems.

Although the functional characterization of nanodevices (device performance) could be legitimately included under nanometrology, it is really too vast and varied a field; selected aspects are considered in Chapters 7 and 10.

5.1 TOPOGRAPHY

Methods can be divided into "contact" and "noncontact". Microscopically, there is little ambiguity about the notion of contact but in the nanoscale one is aware that the Born repulsion prevents atoms from moving arbitrarily close to each other. Hence, although the atomic force microscope (AFM)—Section 5.1.1—is often considered to be a miniature, nanoscale stylus scanned over a sample the repulsion between the atoms of the AFM tip and sample asperity is actually action at a distance, hence the method could equally well be classified as noncontact. Scanning near-field optical microscopy (SNOM), although usually grouped with AFM because the instrumentation is very similar, uses a "light (optical) stylus", for which the notion of contact with the sample is even more tenuous.

5.1.1 Contact Methods

Stylus-based profilers have long been used by engineers to determine the surface roughness of objects. A sharp-pointed stylus equipped with some means of determining its vertical displacement is simply dragged over the surface perpendicular to the surface plane. The vertical motions (deflexions from a mean) of the stylus are considered to more or less faithfully mimic the surface topography, which is recorded through a position transducer attached to the stylus. This device was being progressively miniaturized, and the ability to determine subnanometer vertical displacement was achieved for the first time by the Topografiner, invented by scientists at the US National Standards Institute [173] This nanoscale vertical resolution was achieved using the electron tunneling effect, a quantum phenomenon (given the existence of two levels having the same energy, there is a finite probability for an electron occupying one of the energy levels to pass to the other one (if unoccupied), depending exponentially on the spacial distance separating the levels). The current measured between an electrically conducting stylus and an electrically conducting sample can therefore be converted into sample topography.

Another indispensable technological advance was the perfection of piezoelectric motion controllers in the (x, y) plane (i.e., that of the surface) and in the z direction (perpendicular to the surface). The stylus could then be raster-scanned very close to the surface. A feedback circuit can be arranged to appropriately adjust the z displacement required to keep the tunneling current constant. The perfected instrument was called the scanning tunneling microscope (STM) [21]. The principle, of miniature

styli moving over the sample surface and at each position returning some informa-
tion about topography, or friction, or chemical nature, etc., has meanwhile been vastly
extended to cover dozens of different scanning probe microscopies (SPM), as the
family is called. The most important (in the sense of being the most widely used) is
called the atomic force microscope (Figure 5.1).

It is a considerable advantage over electron microscopy that many probes can
operate in air at atmospheric pressure, and even in liquids. On the other hand, the
forces applied to the sample features are relatively large, and unless extreme care is
taken may distort or even destroy the sample during imaging. Furthermore, the pres-
ence of liquid (even thin films of water coating the sample surface and scanning tip)
gives rise to capillary forces, which may, for example, pull the tip towards the sample
surface (Section 3.3; cf. Figure 3.1). The technique is, however, being continuously
improved. An important innovation has been the introduction of "tapping mode", in
which the cantilever oscillates vertically, thereby minimizing contact of the tip with
the sample, and permitting the use of lock-in amplification to reduce noise.

AFM Resolution. The vertical resolution is limited only by the piezoelectric crystal
that moves the sample relative to the tip, and the arrangement for detecting cantilever
deflexion. An important consideration is that for the finest nanotexture, such as that
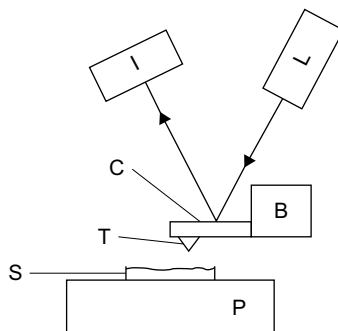
FIGURE 5.1

The atomic force microscope. The sample S is mounted on the platform P, in relation to
which the block B can be moved in the x, y plane (parallel to the surface of S) and z
direction (perpendicular to the surface of S). A flexible cantilever C is mounted on the block,
and on the end of the cantilever is a sharp tip T. In order to record the vertical displacement
of the tip as a result of the asperity of S, the beam from a diode laser L is reflected off the
cantilever onto a split photodiode I. The tip is scanned across the sample (i.e., in the x, y
plane) while remaining almost in contact with it; sample asperity displaces the tip vertically,
and the corresponding deflexions of the cantilever are faithfully recorded as the ratio of the
signals from the two halves of the photodiode, from which sophisticated signal processing
allows the three-dimensional topography of the surface to be extracted. A quadruply split
photodiode enables sideways deflexions of the cantilever due to friction to be recorded
as well.

mimicking a protein surface, the *lateral* features would be smaller than the 30 to 40 nanometer radius of the typical stylus of an SPM: current technology is able to routinely mass produce silicon or silicon nitride tips with radius R equal to a few tens of nanometers. Hence, in the image generated by the microscope the apparent lateral dimensions of features will be broadened (see, e.g., [150]). If r is the true feature radius, the apparent lateral dimension L of an object imaged by a tip of radius R is given by

$$L = 4(Rr)^{1/2} \tag{5.1}$$

This problem can to some degree be overcome by independently measuring the precise shape of the tip (e.g., with a scanning electron microscope) and then processing the deflexion profiles recorded using SPM in order to deconvolute the influence of tip shape—this is the exact analogy of the modulation transfer function (MTF) approach used in optical image processing. Alternatively, finer styli (tips) can be used, for example made by controlled etching of standard tips. Such tips enable subnanometer resolution to be obtained in principle. These ultrafine tips are, however, very fragile and hence easily broken during scanning. Furthermore, if imaging in liquid, impurities dissolved in the liquid may rapidly deposit on the tip. The effects on a conventional tip of $R \sim 30$ nm might be neglected, but such deposition is likely to significantly increase the radius of an ultrafine tip. Carbon nanotubes (see Chapter 9), being extremely thin and rigid, offer a promising alternative to etched tips provided a convenient way of manipulating and attaching them to the microscope cantilever, and possibly replacing them *in situ* when broken or contaminated, can be found.

Contrast Enhancement. A significant challenge in the field is the metrology of fabricated surfaces, especially the characterization of chemical variegation. If only morphology needs to be characterized, then the problem can be solved by either contact or noncontact techniques. The chemical contrast may, however, be insufficient to yield a clear distinction between the different components. A practical approach to overcome this problem is selective post-fabrication processing that only affects one of the components. This principle is applicable more widely than just to the imaging of complex surfaces. For example, iron is a catalyst for carbon nanotube (CNT) growth using (plasma-enhanced) chemical vapor deposition, and hence if there is a heterogeneous deposit (islands) of Fe on the surface, each island will serve as an initiator of columnar CNT formation.

When the variegation resides in differing organic functionalities, the contrast between regions of differing functionalities may be too low. In this case, a useful technique may be to allow high contrast objects smaller than the smallest feature size to selectively bind to one functionality. Now that a plethora of very small nanoparticles is available commercially, and others can be chemically synthesized by well-established methods (see, e.g., [135]), this method has become very useful and practicable. Examples of the decoration of block copolymer films are given in references [121] and [109]. Using fairly extensive tabulations of single-substance surface tensions, the adhesive force between two substances in the presence and absence of a

fluid (in which the nanoparticles will usually be suspended) can be readily estimated (Section 3.2) and used to select a suitable nanoparticle.

Scanning Ion Current Microscopy. A useful extension to the original SPM concept is scanning ion current microscopy (SICM) [62], in which the surface to be characterized is immersed in electrolyte containing a counterelectrode, and the scanning probe is a miniature working electrode inside a very fine glass capillary (Figure 5.2). The closer the end of the capillary is to the surface, the smaller the ion current, which can therefore be used to generate a map of the topography. This method provides an innovative solution to the problem of excessive lateral force being applied to the sample and is particularly useful for mechanically imaging ultrafragile samples, such as a living cell, since the capillary never comes into actual contact with the surface.

Near-Field Optical Microscopy. An important addition to the scanning probe microscopy family is the scanning near-field optical microscope (SNOM), also known as the near-field scanning optical microscope (NSOM). The scanning arrangements remain the same, but now an optical fiber brings light very close to the surface. Transmission, reflexion, and fluorescence can all be measured. The principle is shown in Figure 5.3. The obtainable resolution is below the diffraction limit applicable to far-field optics (equation 5.2). The resolution depends on the fineness of the construction, especially the diameter of the optical fiber-based dielectric probe illuminating the sample. The relative motion, with subnanometer control, between sample and dielectric probe is accomplished using piezoelectric crystals as in the other scanning probe microscopies. A related technique is thermal radiation scanning tunneling microscopy (TRSTM) [166].
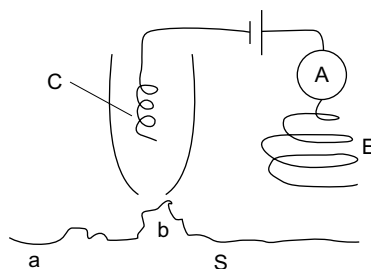
FIGURE 5.2

The scanning ion current microscope (SICM). A capillary C containing the working electrode is moved relative to the sample S. The magnitude of the current (measured by the ammeter A) between the working electrode and the large counterelectrode E depends on the gap between the tip of the capillary and the sample surface. Sample and electrodes are bathed in electrolyte. When C is above feature b, the gap is small, the resistance is high and the current is low. Above feature a, the gap is relatively large, the resistance low, and the current high.
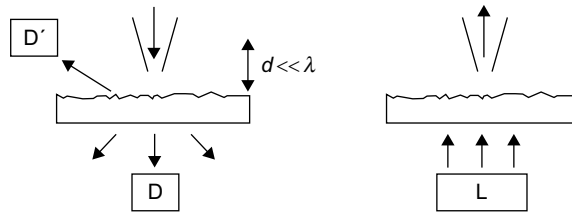
FIGURE 5.3

Scanning near-field optical microscopy (SNOM), also known as near-field scanning optical microscopy (NSOM). On the left, SNOM in illumination mode: a dielectric probe (e.g., a tapered and surface-metallized optical fiber) positioned at a distance $d \ll \lambda$ from the surface illuminates the sample from above. Either the transmitted or the reflected light is collected in the far field (detectors D or D$'$, respectively). On the right, SNOM in collection mode: the sample is illuminated from far below (source L). A dielectric probe in the near field collects the light transmitted through the sample.

5.1.2 Noncontact (Optical) Methods

Optical profilers are analogous to the mechanical stylus instruments but use focused beams to detect the location of the surface. They are therefore unlikely to have the resolution required to characterize nanotexture.

Light Scattering Techniques. These are in principle more useful, especially for characterizing statistical (ir)regularity. Conventional scattering techniques include specular reflexion, total integrated scattering, and angle-resolved scattering; the newer speckle techniques (speckle contrast, speckle pattern illumination, and angle- or wavelength-dependent speckle correlation [105]) are of particular interest. In the speckle pattern illumination method [106], based on doubly scattered coherent light, the (specularly reflecting) surface is illuminated with a monochromatic speckle pattern, whose phase distribution is then modulated by the rough surface. In polychromatic speckle autocorrelation [107], the (diffusely scattering) surface is illuminated with a collimated, partially coherent (i.e., polychromatic) light beam, either discrete (produced by a combination of laser diodes) or continuous (produced by superbright light-emitting diodes, for example).

The distribution of collimated, typically partially coherent light scattered from a diffusely reflecting surface is suitable for determining its statistical roughness up to about a quarter of a wavelength (i.e., about 150 nm for typical visible light sources). If the surface is specularly reflecting, the illuminating light should itself be a speckle pattern, whose phase distribution is modulated by the asperity.

Evanescent optical wave-based techniques, especially its most recent variant, optical waveguide lightmode spectroscopy (OWLS), can yield structural data on ultrathin layers, including geometric thickness, refractive indices, molecular orientation (e.g., [139]), the refractive index profile perpendicular to the interface [115] and the lateral distribution of adsorbed objects (e.g., [8]), with some limitations regarding substrata (Table 5.1). These techniques are especially useful because they can be used

to measure processes *in situ* with good time resolution (Section 5.3) and the nano/bio interface (Section 5.5), described in more detail in those sections.

Imaging Nanostructures. Ever since the invention of the microscope in the 17th century, science has been confronted with the challenge of exploring phenomena that are not directly visible to the human eye. The same extension of the senses applies to "colors" only revealed using infrared or ultraviolet radiation, sounds of a pitch too low or too high to be audible, and forces too slight to be sensed by the nerves in our fingers. Although artists sometimes maintain that there is a qualitative distinction between the visible and the invisible, scientists have not found this distinction to be particularly useful; for them, the problem of "visualizing" atoms is only technical, not conceptual.

Improvements in lenses, and other developments in microscope design, eventually enabled magnifications of about 2000-fold to be reached. With that, objects around 100 nm in size could just be visualized by a human observer peering through the eyepiece of the microscope. The classical microscope runs into the fundamental limitation of spacial resolving power Δx, due to the wavelike nature of light (Abbe's limit):

$$\Delta x = \lambda/2(\text{N.A.}) \tag{5.2}$$

where λ is the wavelength of the illuminating light and N.A. is the numerical aperture of the microscope condenser. To address this problem, one can

- reduce the wavelength of the light
- operate in the near field rather than the far field, as in SNOM (Figure 5.3)
- renounce direct imaging
- use a totally different approach (profilers, Section 5.1.1).

Wavelength Reduction. Although shorter-wavelength varieties of radiation (ultraviolet, X-rays) are well known, as the wavelength diminishes it becomes very hard to construct the lenses needed for the microscope. However, one of the most important

Table 5.1 Summary of the ranges of applicability of optical techniques for investigating thin films

Technique	Thin film material		
	Transparent dielectrics	Opaque materials	Metals
Scanning angle reflectometry (SAR)	√	√	√
Ellipsometry	√	√	√
Surface plasmon resonance (SPR)			√
Optical waveguide lightmode spectroscopy (OWLS)	√		

results emerging from quantum mechanics is the de Broglie relationship linking wave and particle properties:

$$\lambda = h/p \qquad (5.3)$$

where λ is the wavelength associated with a particle of momentum $p = mv$, where m and v are the mass and velocity, respectively, and h is Planck's constant, with a numerical value of 6.63×10^{-34} J s. Hence, knowing the mass and velocity of a particle, we can immediately calculate its wavelength.

The electron had been discovered not long before the formulation of the de Broglie relationship, and was known to be a particle of a certain rest mass ($m_e = 9.11 \times 10^{-31}$ kg) and electrostatic charge e. We know that opposite charges attract, hence the electron can be accelerated to a desired velocity simply by application of an electric field. In other words, the wavelength can be tuned as required! Furthermore, ingenious arrangements of magnetic fields can be used to focus electron beams. The transmission electron (ultra)microscope was invented by Ernst Ruska and Max Knoll in the 1930s. Nowadays, high-resolution electron microscopy can indeed image matter down to atomic resolution. The space through which the electrons pass, including around the sample, must be evacuated, because gas molecules would themselves scatter, and be ionized by, fast-moving electrons, completely distorting the image of the sample. If the sample is very thin, the modulation (according to electron density) of electrons transmitted through the sample can be used to create an electron density map (transmission electron microscopy, TEM). Otherwise, a finely focused beam can be raster-scanned over the sample and the reflected electrons used to create a topographical image (scanning electron microscopy, SEM, first developed by Manfred von Ardenne, also in the 1930s, although not commercialized until the 1960s). In this case, if the sample is not electrically conducting, a thin layer of a metal, typically palladium, must be evaporated over its surface to prevent the accumulation of those electrons that are not reflected, which is liable to obscure the finest features.

Alternatively, if the sample is a semiconductor with a not-too-large band gap, it might be practicable to heat it in order to make it sufficiently conducting. Continuous incremental improvements in the technology of scanning electron microscopy now make it possible to obtain images in the presence of air at a pressure of a few thousandths of an atmosphere. This is called environmental scanning electron microscopy (ESEM). Some resolution is thereby sacrificed, but on the other hand it is not necessary to dehydrate the sample, nor is it necessary to coat it with a metal if it is nonconducting—the remaining air suffices to conduct excess electrons away. A combination of transmission and scanning microscopies (STEM, scanning transmission electron microscopy) was first realized by Manfred von Ardenne but again only commercialized several decades later. It has recently been demonstrated that a commercial instrument, the Nion UltraSTEM, has been practically able to achieve the nanometrologist's goal of imaging and chemically identifying every atom in an arbitrary three-dimensional structure.

Instead of electrons, ions can also be used as imaging particles (FIB, fast ion bombardment). Originally developed as a way of controllably etching parts of samples in order to be able to more completely investigate their inner structures using electron microscopy, the instrumentation is now sufficiently advanced to enable it to be used as an ultramicroscopy technique in its own right.

5.2 **CHEMICAL SURFACE STRUCTURE (CHEMOGRAPHY)**

Many of the available techniques are essentially nonimaging approaches because they lack lateral resolution in the nanoscale. Often nanoscale resolution is only achievable for regular structures. The best-known approach of this type is probably X-ray diffraction (which could equally well be considered as a technique for determining structure). A beam of X-rays is made to impinge on the sample making an angle θ with a plane of atoms within it, and the spacial distribution of the scattered X-rays is measured. Because the wavelength λ of X-rays is of the order of interatomic-plane distance d (tenths of a nanometer), crystalline material, or at least material with some order in its atomic arrangement, diffracts the beam. The key condition for constructive interference of the reflected beam is Bragg's law:

$$d\sin\theta = n\lambda, n = 1, 2, \dots. \tag{5.4}$$

This metrology technique was developed soon after the discovery of X-rays by Röntgen in 1895, in other words long before the era of nanotechnology.

The main family of classical surface chemical analytical methods involves firing one kind of photon (or electron) at the sample and observing the energy of the photons (or electrons) whose emission is thereby triggered.

In Auger electron spectroscopy (AES), an incident electron beam ejects an electron from a core level; the resulting vacancy is unstable and is filled by an electron from a high level, releasing energy that is either transferred to another (Auger) electron (from a yet higher level), or emitted as an X-ray photon. The measurement of the spectrum of the Auger electrons is called AES. In energy dispersive X-ray spectroscopy (EDS, EDX) it is the X-ray photons whose spectrum is measured. Both these techniques are capable of good lateral resolution (within the nanoscale), because the incident electron beam can be finely focused. EDS is typically carried out within a scanning electron microscope equipped with a suitable X-ray detector. It yields quantitative elemental abundances with an accuracy of around 1 atom%. AES, on the other hand, can additionally identify the chemical state of the element. All these techniques yield an average composition within a certain depth from the surface of the sample, which is a complicated function of the scattering of the incident and emergent radiations. Typically AES samples only the first few nanometers from the surface, whereas EDS averages over a somewhat greater depth, which might be as much as 1 μm.

Techniques such as X-ray fluorescence and X-ray photoelectron spectroscopy, in which the incident photon is an X-ray, have insufficient lateral resolution to be useful for mapping nanotexture.

A different kind of technique is secondary ion mass spectrometry (SIMS), in which a beam of energetic ions (typically gallium or oxygen) focused on the sample knocks out ions from it, which are detected in a mass spectrometer. Recent advances in ion beam technology have resulted in the introduction of nanoSIMS, with a lateral resolution of a few tens of nanometers. The relationship between the detected ion abundances and the original sample composition strongly depends on the overall constitution of the sample, and hence quantification is usually a difficult challenge. One advantage of SIMS is that the incident ion beam can be used to systematically etch away the sample, allowing the depth profile of the chemical composition to be obtained.

Atom probe field ion microscopy (APFIM) requires an acicular sample (hence limiting its practical applications); upon applying a high electric field between the counterelectrode and the sample, some atoms evaporate from the sample as ions and move towards a detector, their position on which is directly related to their original position in the sample.

Continuing advances in instrumentation now make it feasible to nondestructively map the inner sections of three-dimensional objects from which they can be completely reconstructed (tomography). Although initially demonstrated by Hounsfield with X-rays, it is now available for a variety of the techniques discussed above, including APFIM, TEM, STEM, AFM, etc., with the capability of nanoscale resolution in all three coordinates.

Atomic force microscopy (Figure 5.1) can also be used to determine the chemical structure since the force–distance relationships as the tip is made to approach and then retracted from a particular patch on the surface are characteristic of the chemistry of that patch (Chapter 3). Careful measurement of those relationships at each patch (Figure 5.4), although laborious, allows a complete map of the chemical variegation of the surface to be obtained. Advances in nanoscale tip fabrication already offer considerable flexibility in choosing tips made from different materials in order to maximize the contrast between different patches. If a rather stiff cantilever is used, its motion will tend to follow the topography of the Born repulsion. A more flexible cantilever will be sensitive to the longer-range, but weaker, electron donor–acceptor and electrostatic interactions (Section 3.2), which depend upon the chemical composition of the sample at the point being measured.

Finally, we note that traditional vibrational spectroscopies (infrared and Raman–Mandelstam) are useful for identifying nano-objects, typically in reflexion mode. Sum frequency generation (SFG) spectroscopy is highly surface-specific. In this technique a beam of visible light ($h\nu_1$) and a (tunable) beam of monochromatic infrared light ($h\nu_2$) are directed onto the surface. The intensity of the sum frequency $h(\nu_1 + \nu_2)$ is measured while varying ν_2; it is enhanced if ν_1 or ν_2 is near a resonance (e.g., due to a bond vibration). Summation, normally forbidden if there is an inversion

FIGURE 5.4

Prototypical force–displacement cycle on the atomic force microscope. At the selected point on the sample surface, the tip is slowly made to approach the sample (starting from position O, at which it is not in contact with the sample). At point a the detector response moves from a to a', corresponding to movement of the monitoring beam reflected off the cantilever from an initial position near the perimeter of the four-segment diode detector towards the center of the four segments, without the cantilever being deflected. During the displacement $a' \rightarrow b$ the beam is centered on the detector and a practically constant response, indicating a practically constant deflexion, that is a practically constant force, is observed. Moving the tip still further into the sample, from $b \rightarrow c$, causes a significant increase in detector response; the displacement in this region is the sum of cantilever deflexion and layer deformation. Hysteresis can be quantified as the force difference between the mean of $a' \rightarrow b$ and $d \rightarrow e$.

center, is allowed at a surface since any inversion symmetry must be broken, hence the technique is surface-specific. It is especially useful for buried interfaces, such as the surface of a solid in contact with water. Strong pulsed lasers are required because the summation efficiency is very low.

5.3 THE METROLOGY OF SELF-ASSEMBLY

Hitherto we have been implicitly considering time-invariant situations. In this case the permanent fixing of the structure required by electron microscopy or the relatively slow rastering procedure required by the "contact" methods render them unsuitable.

A typical self-assembly scenario (see Section 8.2.1) is to disperse nano-objects randomly in a liquid and allow them to adsorb and assemble on a solid substratum (i.e., at the solid/liquid interface). In order to establish basic relations between the self-assembled structure and the characteristics of the nano-objects, the liquid and the substratum, some non-perturbing, *in situ* means of (a) counting the added particles; (b) determining their rate of addition; and (c) determining the structure of the self-assembled film are required.
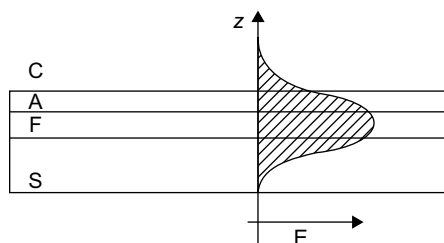
FIGURE 5.5

The electromagnetic field distribution of a zeroth-order guided wave in a four-layer slab waveguide (S, support; F, high refractive index film; A, self-assembled adlayer; C, cover medium (liquid in which the nano-objects would be initially suspended). Note the exponentially decaying evanescent fields in the zones S and (A, C). The highest sensitivity of the phase velocity of the guided modes to adlayer structure is obtainable with thin waveguides whose thickness (i.e., that of the F-layer) is close to the cut-off limit of the zeroth modes, but in this case only two orthogonal modes can be excited in the waveguide.

Optical methods (scanning angle reflectometry (SAR), ellipsometry) rely on monitoring changes in the reflectance of the solid/liquid interface due to accumulation of the nano-objects. Optical waveguide lightmode spectroscopy (OWLS) relies on perturbation of the evanescent field generated by waves guided along the substratum by the nano-objects, provided their polarizability is different from that of the liquid (Figure 5.5).

If the substratum is a thin metal film, optically excited surface plasmon resonance (SPR) can also be used to monitor the presence of the nano-objects, although it is less sensitive and less informative than OWLS. If an electrode-coated piezoelectric crystal can be used as the substratum, changes in the resonant vibration frequency of the crystal and the dissipation of its oscillation can also provide useful information (the quartz crystal microbalance, QCM).

5.4 THE REPRESENTATION OF TEXTURE

The ultimate representation of any surface is to list the height (relative to some reference plane) of each atom or cluster on the surface. This would amount to a topographical image and is, of course, often obtained, albeit with lower resolution. If, however, one wishes to correlate the texture of the surface with some functional parameter, the information needs to be interpreted (e.g., in order to determine the laws governing the effects of nanotexture on the response of—for example—living cells growing on it) and one needs to go beyond mere images. Hence, ways of reducing the data are sought. The simplest way of describing an uneven surface, reducing it to a single parameter, is to compute its roughness. At first sight roughness appears

to be rather random and suitable for a statistical treatment. Although profilometry is capable of yielding a continuous description of the height $h(x)$ of a surface scanned in the x direction (parallel to the plane of the surface), in this digital age we should assume that it is measured discretely and has a value h_i for the ith measured patch.

5.4.1 Roughness

The quantification of roughness then amounts to a straightforward statistical analysis of the set of height data as summarized in Table 5.2. This list of parameters is not exhaustive. Others are in use, including some derived from the Abbott curve, a plot of the height of the surface as a function of the fraction of the surface exceeding that height. Most of these parameters are described in international standards (e.g., ISO 13565). A general and sophisticated surface characterization can be accomplished by determining the power spectral density (PSD) $\mathcal{S}$ of surface features, which encapsulates information about the lateral distribution of the individual roughness components absent in the simple parameters given in Table 5.2:

$$\mathcal{S}(f_x, f_y) = \lim \frac{1}{L^2} \left(\int_0^L \int_0^L h(x,y) e^{-i2\pi(f_x x + f_y y)} \mathrm{d}x \mathrm{d}y \right)^2 \tag{5.5}$$

where the f are the spacial frequencies in the (x, y) plane of the two-dimensional surface profile $h(x, y)$ defined for a square of side L. The power spectral density is the Fourier transform of the autocorrelation function.

If there is some evidence of statistical self-similarity, the corresponding fractal dimension may be a compact way of representing the roughness. If the power spectral

Table 5.2 Table of common roughness parameters. h_i is the height of the ith spot on the surface, with reference to some arbitrary fixed level; N is the number of measured spots. Additional parameters are the skew or skewedness (the third moment of the distribution of h) and the kurtosis (the fourth moment)

Symbol	Name	Definition	Synonyms
$\langle h \rangle$ or $\bar{h}$	Mean height	$(1/N) \sum_i^N h_i$	Zeroth moment; d.c. component
R_a	Arithmetical mean roughness	$(1/N) \sum_i^N \lvert h_i - \langle h \rangle \rvert$	First moment; center-line average (CLA)
R_q	Root mean square roughness	$\sqrt{\sum_i^N (h_i - \langle h \rangle)^2 / N}$	Second moment; interface width
R_t	Maximum peak to valley	$\max h_i - \min h_i$	Maximum roughness depth
R_z	Mean peak to valley height	R_t averaged over segments	
$\mathfrak{r}$	roughness ratio	$\dfrac{\text{actual area}}{\text{apparent area}}$	

density is plotted (as its logarithm) against the logarithm of the spacial frequency, and if there is a straight line portion extending over two or three orders of magnitude of frequency, then a fractal dimension d_C can be derived from the slope. A perfectly flat surface has a (Euclidean) dimension of 2; a fractal surface would have a dimension > 2. The formal definition of fractality relates to the capacity of the surface, and in particular how that capacity scales with the size of objects placed on the surface. One can determine it by measuring the number n of balls required to cover the surface, using a series of differently sized balls, and plotting $\log n$ against $\log r$, where r is the radius of a ball. For a perfectly flat surface, n will be proportional to $1/r^2$, but for the fractal surface one will be able to place proportionally more smaller balls, and the slope $-d_C$ of the log–log plot becomes greater than 2. Experimentally, the fractal surface dimension could be determined by measuring the quantity of nanoparticles of a range of sizes ranging from a few nanometers to a few micrometers required to jam the surface.

A fractal dimension is actually the capacity, as originally defined by Kolmogorov [97]:

$$d_C = \lim_{\varepsilon \to 0} \log N(\varepsilon)/\log(1/\varepsilon) \tag{5.6}$$

where $N(\varepsilon)$ is the minimum number of p-dimensional cubes of side ε needed to cover the set; p is the dimension of Euclidean space and the set in question is a bounded subset of that space.

The fractal concept highlights the importance of characteristic length as a parameter of surface roughness, even though very few real surfaces have self-similarity extending over a sufficiently wide range of lengths to warrant being described as fractals. The fact is that in many of the complex environments with respect to which biocompatibility is being determined, a single roughness parameter is quite inadequate to characterize surface topography. That is perhaps the main reason why the present state of knowledge regarding correlations between surface roughness and cell response is still in a fairly primitive state.

Lacunarity is an extension of the fractal concept that attempts to capture information about the spacial correlations of real surfaces (Section 5.4.3).

5.4.2 One-Dimensional Texture

The techniques discussed in Sections 5.1 and 5.2 are aimed at mapping out the chemical and topographical variegation to as fine a resolution as that of any variegation that may be present. Raster techniques such as the scanning probe microscopies can identify variegation pixel by pixel and produce a one-dimensional string of information *a priori* suitable for analyzing according to the methods suggested in this section (the to-and-fro motion of the scanning probe may impose a certain kind of correlation on the symbolic sequence if there are two-dimensional features present; i.e., thicker than a single scanning line). The main task is to determine the presence of compositional and/or topographical correlations within the plane of the surface.

Preliminary to the discussion of nanotexture proper, which implies a two-dimensional arrangement of features important for molecular and cellular recognition, let us then consider the one-dimensional situation, which is of course much simpler. The purpose of this section is to look at some relatively straightforward parameters that might be used as means to quantify texture. A one-dimensional sequence of symbols (each symbol representing a topographical feature, such as a height, or a chemical element in a particular oxidation state) can be completely defined by the set of probability distributions $W_1(yx)dy$, the probability of finding y, the value of the symbol, in the range $(y, y + dy)$ at position x (if one is moving along the sequence at a uniform rate, this might be equivalent to a time t), $W_2(y_1x_1, y_2x_2)dy_1dy_2$, the joint probability of finding y in the range $(y_1, y_1 + dy_1)$ at position x_1 and in the range $(y_2, y_2 + dy_2)$ at position x_2, and so on for triplets, quadruplets and higher multiplets of values of y. If there is an unchanging underlying mechanism generating the sequence, the probabilities are stationary and the distributions can be simplified as $W_1(y)dy$, the probability of finding y in the range $(y, y + dy)$; $W_2(y_1y_2x)dy_1dy_2$, the joint probability of finding y in the ranges $(y_1, y_1 + dy_1)$ and $(y_2, y_2 + dy_2)$ when separated by an interval $x = x_2 - x_1$; etc. If successive values of y are not correlated at all, i.e.,

$$W_2(y_1x_1, y_2x_2) = W_1(y_1x_1)W_1(y_2x_2) \tag{5.7}$$

etc., all information about the process is completely contained in W_1, and the process is called purely random.

Markov Chains
If, however, the next step of a process depends on its current state, i.e.,

$$W_2(y_1y_2x) = W_1(y_1)P_2(y_2|y_1x) \tag{5.8}$$

where P_2 denotes the conditional probability that y is in the range $(y_2, y_2 + dy_2)$ after having been at y_1 at an earlier position x, we have a Markov chain, defined as a sequence of "trials" (that is, events in which a symbol can be chosen) with possible outcomes $\mathbf{a}$ (possible states of the system), an initial probability distribution $\mathbf{a}^{(0)}$, and (stationary) transition probabilities defined by a stochastic matrix P. A random sequence, i.e., with a total absence of correlations between successive symbols, is a zeroth order Markov chain. Without loss of generality we can consider a binary string, that is a linear sequence of zeros and ones. Such sequences can be generated by choosing successive symbols with a fixed probability. The probability distribution for an r-step process is

$$\mathbf{a}^{(r)} = \mathbf{a}^{(0)}P^r \tag{5.9}$$

The Markovian transition matrix, if it exists, is a compact way of representing texture.

If upon repeated application of P the distribution $\mathbf{a}$ tends to an unchanging limit (i.e., an equilibrial set of states) that does not depend on the initial state, the Markov chain is said to be ergodic, and

$$\lim_{r \to \infty} P^r = Q \tag{5.10}$$

where Q is a matrix with identical rows. Now

$$PP^n = P^n P = P^{n+1} \tag{5.11}$$

and if Q exists it follows, by letting $n \to \infty$, that

$$PQ = QP = Q \tag{5.12}$$

from which Q (giving the stationary probabilities, i.e., the equilibrial distribution of **a**) can be found. Note that if all the transitions of a Markov chain are equally probable, then there is a complete absence of constraint, in other words the process is purely random.

The entropy of a Markov process is the "average of an average" (i.e., the weighted variety of the transitions). For each row of the (stochastic) Markov transition matrix, an entropy $H = -\sum_i p_i \log_2 p_i$ (Shannon's formula) is computed. The (informational) entropy of the process as a whole is then the average of these entropies, weighted by the equilibrial distribution of the states.

Runs

Another approach to characterize texture is to make use of the concept of a run, defined as a succession of similar events preceded and succeeded by different events. Let there again be just two kinds of elements, 0 and 1, and let there be n_0 0s and n_1 1s, with $n_0 + n_1 = n$. r_{0i} will denote the number of runs of 0 of length i, with $\sum_i r_{0i} = r_0$, etc. It follows that $\sum i r_{0i} = n_0$, etc. Given a set of 0s and 1s, the numbers of different arrangements of the runs of 0 and 1 are given by multinomial coefficients and the total number of ways of obtaining the set $r_{ji}(j = 1, 2; i = 1, 2, \ldots, n_0)$ is [119]

$$N(r_{ji}) = \begin{bmatrix} r_0 \\ r_{0i} \end{bmatrix} \begin{bmatrix} r_1 \\ r_{1i} \end{bmatrix} F(r_0, r_1) \tag{5.13}$$

where the terms denoted with square brackets are the multinomial coefficients, which give the number of ways in which n elements can be partitioned into k subpopulations, the first containing r_0 elements, the second r_1, etc.:

$$\begin{bmatrix} r \\ r_i \end{bmatrix} = \frac{n!}{r_0! r_1! \ldots r_k!}, \quad \text{with } \sum_i^{i=k} r_i = n \tag{5.14}$$

and the special function $F(r_0, r_1)$, given in Table 5.3, is the number of ways of arranging r_0 objects of one kind and r_1 objects of another so that no two adjacent objects are of the same kind. Since there are $\binom{n}{n_0}$ possible arrangements of the 0s and 1s, the distribution of the r_{ji} is

$$P(r_{ji}) = \frac{N(r_{ji}) F(r_0, r_1)}{\binom{n}{n_0}} \tag{5.15}$$

The deviation of the actually observed distribution from equation (5.15) can be used as a statistical parameter of texture.

Table 5.3 Values of the special function $F(r_0, r_1)$ used in equations (5.13) and (5.15)

$|r_0 - r_1|$	$F(r_0, r_1)$
> 1	0
1	1
0	2

Genetic Sequences

One very important physical instantiation of one-dimensional texture is the sequences of nucleic acids that encode proteins and also constitute the binding sites (promoters) for transcription factors—proteins that bind to DNA as a prerequisite for the transcription of the DNA into RNA that precedes the translation of the RNA into amino acid sequences (proteins). The nucleic acids have a variety of four (the bases A, U, C, G in natural RNA, and A, T, C, G in DNA). Many statistical investigations of intragenome DNA correlations group the four into purines (A and G) and pyrimidines (U or T and C)—rather like analyzing texts. Mere enumeration of the bases is of course inadequate to characterize texture.

Algorithmic Information Content

Irregularity can be quantified by estimating the algorithmic information content (AIC, also called algorithmic or Kolmogorov complexity), which is essentially a formalization of the notion of estimating the complexity of an object from the length of a description of it. The first task is to encode the measured variegation in some standard form. The choice of rules used to accomplish the encoding will depend on the particular problem at hand. For the one-dimensional nucleic acid textures referred to in the preceding paragraph, encoding purines as 1 and pyrimidines as 0 may be sufficient, for example. The formal definition of the AIC of a symbolic string s (encoding the object being described) is "the length of the smallest (shortest) program P that will cause the standard universal computer (a Turing machine T) to print out the symbolic string and then halt". Symbolically (but only exact for infinite strings), denoting the AIC by K,

$$K(s) = \min\{|P| : s = C_T(P)\} \tag{5.16}$$

where |P| is the length of the program (in bits) and C_T is the result of running the program on a Turing machine. Any regularity present within the string will enable the description to be shortened. The determination of AIC is therefore essentially one of pattern recognition, which works by comparing the unknown object with known prototypes (there is no algorithm for discovering patterns *de novo*, although clusters may be established according to the distances between features). The maximum value of the AIC (the unconditional complexity) is equal to the length of the string in the

absence of any internal correlations; that is, considering the string as random, viz.,

$$K_{\max} = |s| \tag{5.17}$$

Any regularities, i.e. constraints in the choice of successive symbols, will diminish the value of K from $K_{\max}$.

Effective Complexity

The concept of effective complexity (EC) [61] was introduced in an effort to overcome the problem of AIC increasing monotonically with increasing randomness. EC is defined as the length of a concise description (which can be computed in the same way as the AIC) of the set of regularities of the description. A very regular symbolic sequence will have only a small number of different regularities, and therefore a short description of them; a random sequence will have no regularities, and therefore an even shorter description. There will be some intermediate descriptions with many different regularities, which will yield a large EC. Essentially,

$$EC = AIC - RIC \tag{5.18}$$

where RIC is the random information content. In a certain sense, EC is actually a measure of our knowledge about the object being described, for it quantifies the extent to which the object is regular (nonrandom), and hence predictable. It presents the same technical difficulty as AIC: that of finding the regularities, both in compiling an initial list of them, and then in finding the regularities of the regularities.

5.4.3 **Two-Dimensional Texture: Lacunarity**

Whereas AIC and EC can be straightforwardly applied to linear texture, it is not generally obvious how the two-dimensional pattern should be encoded. Of course it could be mapped in raster fashion (as is actually done in SPM and SEM), and patterns extending over many lines should appear as regularities.

Another approach to capture information about the spacial correlations of arbitrarily heterogeneous real surfaces is to extend the fractional dimension or fractal representation of roughness to encompass the quantification of voids in rough objects (the lacunarity Λ). Consider an image constructed from binary (black or white, corresponding to values of 0 and 1) pixels, and let the numbers of boxes of side r containing s white pixels have the distribution $n(s,r)$. $\Lambda(r)$ is defined as

$$\Lambda(r) = M_2/M_1^2 \tag{5.19}$$

where M_1 and M_2 are the first and second moments of the distribution,

$$M_1(r) = \sum_{s=1}^{r^2} s(r)n(s,r)/N(r) = \langle s(r)\rangle \tag{5.20}$$

and

$$M_2(r) = \sum_{s=1}^{r^2} s^2(r)n(s,r)/N(r) = \sigma_s^2(r) + \langle s\rangle^2(r) \tag{5.21}$$

where $\langle s \rangle$ and σ_s^2 are, respectively, the mean and variance of the distribution and the total number $N(r) = (M - r + 1)^2$ for a square pattern of size M of boxes of size r (i.e., a type of variance-to-mean ratio). The lacunarity can be thought of as the deviation of a fractal from translational invariance [60], or a scale-dependent measure of heterogeneity (i.e., "texture") of objects in general [4]. Its lowest possible value is 1, corresponding to a translationally invariant pattern (including the special case $\Lambda(M) = 1$). Practically, comparison of the lacunarity plots (i.e., a log-log plot of the function $\Lambda(r)$) of artificially generated patterns with the experimental lacunarity may be used to analyze the texture of the sample.

5.5 METROLOGY OF THE NANO/BIO INTERFACE

Here we come to the most difficult challenge for the nanometrologist. Whereas the techniques discussed in the preceding sections are closely related to well-established ones—indeed metrology was ably handling measurements in the nanoscale long before the term "nanotechnology" was introduced and even before Feynman's lecture [56]—when living organisms are introduced into the system it is moved into unfamiliar territory. Particular problems are associated with: the very narrow range of ambient conditions in which life can be maintained (otherwise the life must be terminated and the sample fixed, but then one may not be measuring what one wishes to know); the fact that life is mainly composed of soft, irregular structures that are constantly evolving on different timescales; and the fact that any living organism, even a single cell, is able to modify its environment (e.g., by secreting macromolecular substances).

One of the most important techniques yielding nanometer resolution is electron microscopy (EM, Section 5.1.2). In principle, however, this is ruled out for investigating the nano/bio interface because of the damage to delicate biostructures caused by the necessary sample preparation, typically including dehydration and impregnation with heavy metals [74]. Given that water is the dominant molecular species in nearly every biological system, the ability to operate in an aqueous environment is an essential prerequisite.

Some measurement problems associated with nanotechnology may not require nanometrology. For example, the toxicologist investigating the effect of nanoparticles on a living organism may merely administer the nanoparticles intravenously and observe the subsequent behavior of the animal.

5.5.1 Determining the Nanostructure of Protein Coronas

The protein corona is the collection of proteins that accumulate on an artificial surface introduced into a biofluid such as blood; that it, the process of opsonization. Since there are many hundreds of different proteins (far more if one includes all the IgG antibodies of different affinities as "different proteins") it is very likely that

the surface becomes coated with a complex structure comprised of many different proteins and, moreover, constantly changing with time.

A major goal is to be able to design surfaces that can determine the nature of the corona. To achieve this, one must understand how the corona is assembled. Unlike other self-assembly problems, the corona is constantly changing its composition and structure. The main challenges are identifying the proteins and characterizing the structure. The first of these is very difficult. If one exposes the corona-decorated nano-objects to a series of different monoclonal antibodies, this would allow one to at least identify the outermost proteins, but the procedure becomes impracticable for more than a few proteins; apart from the expense of procuring the monoclonal antibodies, the initial ones to bind will block access by the subsequent ones. It must, moreover, be repeated over a long succession of intervals. On the other hand, the reflectance techniques described in Section 5.3 are well applicable to measuring the nanostructural evolution of the corona *in situ*. OWLS in particular is well able to handle anisotropy [79] and refractive index profile determination [115], as well as overall thickness and mean refractive index of the corona.

5.5.2 Measuring Cell Adhesion: the Interaction of an Evanescent Field with a Cell

The evanescent field decays exponentially perpendicular to the plane of the waveguide (see, e.g., [162]), and the characteristic decay length $1/s$ (in which the field intensity decays to $1/e$ of its value at the waveguide surface), also called the penetration depth, depends on the refractive indices of the waveguide, the waveguide support S, the cover medium C (including the cells), and the film thickness d_F (see Figure 5.5). For the transverse electric (TE) guided lightmodes the inverse penetration depth is given by

$$s = k(N^2 - n_C^2)^{1/2} \tag{5.22}$$

where k is the wave vector of the light (numerically equal to $2\pi/\lambda$, λ being the vacuum wavelength), N is the effective refractive index of the guided mode, and n_C is the refractive index of the composite cover medium (culture fluid and cells) as sensed by the guided lightmode. This is for decay into the cover medium; the inverse decay length into the substratum is

$$s = k(N^2 - n_S^2)^{1/2} \tag{5.23}$$

In using these equations one should bear in mind that N also depends on n_S as well as all the other waveguide parameters.

To a first approximation, we can suppose that at any distance z from the waveguide surface, the cover medium refractive index is the arithmetic mean of that of the cells and that of the culture fluid:

$$n_C(z) = n_\kappa cA(z) + n_M(1 - cA(z)) \tag{5.24}$$

where n_κ is the refractive index of the cell material (which may itself depend on z), c is the number of cells per unit area of the waveguide, $A(z)$ is the cross-sectional area of the cell parallel to the waveguide at a distance z from the waveguide surface, and n_M is the refractive index of the pure culture fluid. Hence, to determine the total contribution of the cells to n_C, we need to integrate the product $A(z)e^{-sz}$ from $z = 0$ to ∞; this is just the Laplace transform of $A(z)$, i.e.,

$$v' = \int_0^\infty A(z)e^{-sz}\mathrm{d}z = \mathcal{L}(A(z)) \tag{5.25}$$

v' is called the effective volume of the cell [144].

By taking the Laplace transform of the right-hand side of equation (5.24) and dividing by the total volume effectively sensed by the waveguide, equal to $\int_0^\infty e^{-sz}\mathrm{d}z = 1/s$, we obtain the final expression for the cover refractive index:

$$n_C = n_\kappa s c v' + n_M(1 - s c v') \tag{5.26}$$

where we have assumed that n_κ is the refractive index throughout the cell. If this is considered to be inadmissible, for example when taking into account the optically denser nucleus, the term $n_\kappa c A(z)$ in eqn (5.24) can be replaced by $cA(z)[n_D D(z)/A(z) + n_\kappa(1 - D(z)/A(z))]$, where n_D is the refractive index of the nucleus, and $D(z)$ is its cross-sectional area, obviously assumed to be less than $A(z)$. Additional terms could in principle be added for the endoplasmic reticulum, Golgi body, etc., although probably only the nucleus warrants special treatment, its refractive index being significantly higher because of the relatively high concentration of phosphorus-containing nucleic acids.

The fundamental property that is measured with total internal reflexion is the phase shift Φ; the complex Fresnel reflexion coefficient can be written out as

$$\hat{R} = |\hat{R}|e^{i\Phi} \tag{5.27}$$

The optical waveguide is constructed by sandwiching a slab of the high refractive index material F between lower refractive index materials, the mechanically strong support S and the medium C containing the cells. Both these materials are much thicker than $1/s$, and hence can be considered semi-infinite. The condition for the guided waves to propagate is for the different contributions to Φ to sum to zero or an integral multiple of 2π. This condition can be written as a set of mode equations linking the effective refractive index N of a particular mode with n_S, n_F, d_F and n_C. Only discrete values of N are allowed, corresponding to the mode numbers $m = 0, 1, 2, \ldots$, each value of m being represented by two orthogonal polarizations, transverse magnetic (TM) and transverse electric (TE). The sensitivities $\partial N/\partial n_C$ depend on the polarization, mode number and effective waveguide thickness: below a certain minimum *cutoff thickness* no propagation can take place and as the thickness increases the sensitivity rapidly reaches a maximum and then slowly diminishes; the sensitivities decrease for higher order modes for conventional waveguides [162].

For reverse waveguides ($n_S < n_C$—called "reverse symmetry" waveguides in some earlier literature—(see Section 5.5.4) the sensitivity at the cutoff thickness is 1

for all of the modes and slowly decreases above the cutoff thickness [162, 81]. If n_S, n_F and d_F are known, n_C can be determined from measurement of a single value of N. According to equation (5.26), any change in n_C can be interpreted as a change in cell concentration (number per unit area of substratum) c, or in cell size (characterized by the radius R), or in cell shape (if the shape is assumed to be a segment, then the contact area a will suffice as a characteristic parameter, if it is further assumed that the volume does not change upon spreading), or in cell refractive index (distribution).

In summary, then, the fundamental measurable quantity output by the optical measurement setup is one (or more) effective refractive indices N. Further interpretation of N depends on the model chosen, with which the distribution of polarizability within the evanescent field can be linked to changes in the cover medium refractive index n_C.

5.5.3 **Optical Measurement Schemes**

The principal approaches to measure the effective refractive indices N with high precision are grating coupling and interferometry. In the former, a diffraction grating created in the waveguide is used to couple an external light beam incident on the grating with an angle α to the grating normal; measurement of α yields N according to

$$N = n_{\text{air}} \sin\alpha + 2\pi\ell/\Lambda \tag{5.28}$$

where $\ell = 0, \pm 1, \pm 2, \ldots$ is the diffraction order and Λ is the grating constant. The diffraction grating can be created either by modulating the topography or the refractive index of the waveguide material. α can be determined by mounting the waveguide on a high precision goniometer with the grating positioned at its axis of rotation. Photodiodes placed in contact with the ends of the waveguide enable the incoupling peaks to be determined while α is varied. It is customary to measure propagation in both possible directions and take the average, thereby correcting for any irregularity in the grating position and yielding the absolute incoupling angle. Alternatively the grating can be used to couple light already propagating in the waveguide out of it. However, this requires a more elaborate arrangement for introducing the light into the waveguide, and it is more difficult to measure the absolute incoupling angle. Grating coupling is also called optical waveguide lightmode spectroscopy (OWLS), because the spectrum of modes can be obtained by scanning α. As well as the high precision and ease with which absolute effective refractive indices of as many modes as the waveguide will support can be determined, OWLS also enables the incoupling peak shape to be determined; any broadening beyond the theoretical lower limit of width ($\approx \lambda/L_x$ in radians, where L_x is the illuminated grating length [162]) is highly informative regarding the distribution of objects on the waveguide surface [37].

Variants of the classical incoupling scheme described above include the use of chirped gratings (i.e., with a spacially-varying Λ) and varying λ to determine N. Such schemes have been devised to eliminate moving parts, especially the expensive and relatively slow high precision goniometer. However, they are generally less precise and accurate. Nevertheless, it might be worth sacrificing some precision in return

for another benefit, such as high throughput. Hence, incoupling of broadband light and spectral monitoring of the outcoupled light by the same grating provides a very compact way of measuring relative effective refractive index changes (this scheme is also called the resonant waveguide grating (RWG) [51]).

The three main interferometric techniques that *might* be used (no measurements with living cells using interferometry have apparently been reported—often the waveguide geometry is unfavorable; e.g., too narrow to underlie the entire cell, and excessive scattering engendered by the long waveguides used to increase sensitivity vitiates useful information being obtained) are:

1. Allowing the TE and TM modes to interfere with one another; this is the simplest approach because a planar waveguide without any structuring can be used (although the optical measuring arrangement is considerably more elaborate than with grating coupling).
2. The Mach–Zehnder interferometer, the integrated optical analog of the Rayleigh interferometer—typically a ridge or channel waveguide is split into two channels, one of which is the cell substratum, and then recombined.
3. The "dual polarization interferometer" (DPI) in which a sandwich structure consisting of support, reference waveguide, intermediate layer (may be the same material as the support) and finally the waveguide acting as cell substratum is illuminated at one end, and interference between the two beams emerging at the other end is recorded in the far field.

The main advantage of an interferometer is that the length can be arbitrarily extended to increase the sensitivity (provided the waveguide can be made sufficiently uniform and the temperature maintained constant over the entire length). On the other hand it is usually only practical to measure with one mode, and it is very difficult to measure absolute values of the effective refractive index. Moreover, in interferometry any lateral inhomogeneity information is simply averaged. This is in contrast to grating coupling, in which energy is injected continuously under the cells in the entire measuring zone, not just at the end of the waveguide. Therefore, the resonant peak shape carries information about the optical microstructure of the cover medium [37], information inaccessible to an interferometric measurement.

5.5.4 Reverse Waveguides

In the conventional asymmetric waveguide, the support refractive index (e.g., optical glass) is greater than that of the cover, and this continues to hold when the cover consists of an aqueous medium and living cells. The penetration depth into the cover is limited to a distance of around 200 nm, implying that only a small part of the cell body is sensed by the evanescent field. The maximum probe depth is achieved at the cutoff point ($N = n_S$), when the sensitivity of such a waveguide is zero [81, 162]; when the sensitivity is maximal the probe depth is even shorter, around 100 nm. While this is completely sufficient for monitoring the spreading transition from sphere to segment (Figure 4.3), because the change in distribution of polarizability overwhelmingly
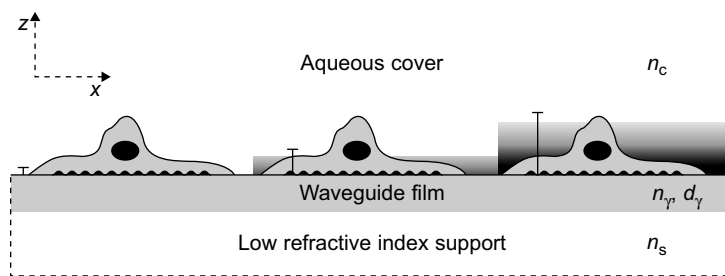
FIGURE 5.6

Cross-section of a cell on a reverse waveguide with (from left to right) successively increasing penetration depths.

takes place in the immediate vicinity of the waveguide surface (and this is expected to hold even if the actual cell–substratum contact is mediated by filopodia), the monitored optical signal will be insensitive to changes of polarizability distribution occurring within the bulk of the cell body (which may arise through a reconfiguration of the cytoskeleton, or repositioning of the nucleus). If, however, $n_S < n_C$, the penetration depth into the cover medium can be readily increased and tuned without any upper limit [80, 81, 162] (note that the condition $\max(n_S, n_C) < N < n_F$ must always be met). This is called the reverse waveguide configuration (Figure 5.6). Typical practical values of waveguide parameters are constrained by $n_M \sim 1.34$ for culture medium, and $n_\kappa \sim 1.4$, making $n_C \sim 1.35$. This means that reverse waveguiding cannot be achieved with the usual waveguide support materials (optical glass or quartz, $n_S \sim 1.5$). Sufficiently mechanically strong supports with a very low refractive index can be fabricated by introducing a large volume fraction of air-filled nanoscale pores into the support material.

5.5.5 The Interpretation of Effective Refractive Index Changes

Contact Area and Detachment Force. The basic interpretation is encapsulated in equation (5.26). If only spherical cells are being investigated, then the number of cells c per unit area under a certain set of environmental conditions (e.g., a constant weak force of detachment) provides a basic quantification of adhesion. If the waveguide is positioned vertically, then the (gravitational) force acting to detach the cells is uniform; the evident complexity involved in placing an OWLS device in a centrifuge seems to have precluded any experimental attempt. If the waveguide is positioned horizontally, then the detaching force can be supplied by fluid flow through the cuvette. An approximate formula for the detachment force F imposed on a cell of (unspread) radius R by a wall shear stress γ is [82]

$$F = 32R^2\gamma \tag{5.29}$$

This formula is insensitive to the exact shape. The actual situation is more complicated (see, e.g., [26]) because the cell can deform under the influence of shear stress before detachment.

Spreading Kinetics. The dynamics of change of some parameter (such as contact area) may be characteristic of the cell. It may be indicative of adhesiveness (and its dependence upon the presence of calcium or other ions); the occurrence of spreading may distinguish between different cell types; it may reveal the importance of nutrient; it may reveal the presence of extracellular matrix molecules coating a mineral substratum; it may reveal the adverse response of cells to drugs or toxic chemicals.

Bulk Versus Surface Effects. In order to separate bulk and surface effects, the differing depth sensitivity curves of the TE and TM modes can also be exploited to derive a weighted difference D_γ of sensitivities, and a *crossover depth* at which $D_\gamma = 0$ (below it is negative and above it is positive [80]). Writing the sensitivities as S, the sign of the linear combination

$$D_\gamma = \gamma S_{\text{TM}} - S_{\text{TE}} \qquad (5.30)$$

depends on the magnitude of the (always positive) tuning parameter γ, selected according to the waveguide parameters and depth zone of interest. Thus, if $D_\gamma < 0$, then the predominant refractive index changes take place near the waveguide surface; if $D_\gamma > 0$, then the changes are predominantly in the bulk of the cell body.

5.5.6 The Interpretation of Coupling Peak Width Changes

The introduction of binary (i.e., cell–no cell) spacial inhomogeneity in the cover medium C broadens the incoupling peaks (Section 5.5.3). The greatest inhomogeneity must necessarily occur when exactly half the surface is covered with objects (i.e., cells) whose refractive index differs from that of the bathing medium. If the locally engendered effective refractive indices are sufficiently different two peaks may be observed. If the penetration depth is rather small, as the cell–substratum contact area a increases from zero (corresponding to the perfectly spherical initially attached state of the cell), some measure of peak width w (such as full width at half maximum) should first increase *pari passu* with effective refractive index N, up to the point at which the cells cover half the detected area, and then decline; the highest value attained by w depends on the optical contrast between the two materials (i.e., cells and culture medium) constituting the covering medium.

Microexudate. Very commonly, living eukaryotic cells respond to the solid part of their environment (e.g., their substratum) by excreting extracellular matrix (ECM) molecules (mostly large glycoproteins), which adsorb onto the substratum, modifying it to render it more congenial. Even the most carefully prepared nanostructure may thereby rapidly lose its initial characteristics. The prokaryotic equivalent is the biofilm, which constitutes a kind of microbial fortress defending the population against external attack. These adsorbed secretions, whose refractive index is greater than that of the culture medium, cause the effective refractive index N to increase additionally to the increase caused by the cell shape changes (spreading).

However, the coupling peak widths w will be negligibly affected because the discrete objects (glycoproteins) comprising the secreted microexudate are much smaller than the wavelength of light, and can in general be assumed to be adsorbed uniformly over the entire waveguide surface. Hence, by simultaneously comparing the evolutions of N and w, two processes contributing independently to the increase of N, namely the increase of a and the accumulation of ECM glycoproteins at the substratum surface, can be separately accounted for.

5.6 SUMMARY

Metrology at the nanoscale imposes stringent requirements of accuracy and precision. A particular challenge is the fact that the size of the measuring instrument and the feature being measured become comparable. Should the nanoscale feature be absolutely small in the quantum sense, then we have the potential problem that its state is destroyed by the measurement. Given the identity of quantum objects, however, this does not usually pose a real practical difficulty—one merely needs to sample (and destroy) one of many. Regardless of the absolute size of the object being measured, if the measurement instrument has the same relative size as the object being measured, the measurement is subject to distortion. Previously this could always be resolved by shrinking the relative features of the measurement instrument. If, however, the object being measured is already at the lower size limit of fabrication (i.e., in the nanoscale) no further shrinkage of the measurement instrument is possible. Nevertheless, this does not pose an insuperable difficulty. Indeed, it has already been encountered and largely overcome whenever light has been used to measure features of an object smaller than the wavelength of the light.

The ultimate goal of nanometrology is to provide a list of the coordinates and identities of all the constituent atoms of a nanoscale structure, and all practical metrology is aimed at that goal. Approaches can be categorized as imaging or nonimaging, each of which category contains methods that can in turn be categorized as contact or noncontact. As one approaches the nanoscale, however, the distinction becomes harder to make. In another dimension, one can distinguish between topographical and chemical features; some techniques yield both. In particular, commercial scanning transmission electron microscopy is now able to yield the atomic coordinates of arbitrary three-dimensional structures.

Techniques able to make time-resolved measurements *in situ* are very useful for monitoring actual processes.

The representation of structure by a list of the atomic coordinates and identities will usually lead to vast, unmanageable data sets. Therefore, there is considerable interest in capturing the salient features of topography and chemistry by identifying regularities, possibly statistically.

The nano/bio interface presents perhaps the greatest challenges to nanometrology, not least because the introduction of living components into the system under scrutiny

moves the problem into territory unfamiliar to most metrologists. However, important advances are being made in this domain, which has already allowed it to assume a far more quantitative nature than hitherto.

Perhaps the greatest current challenge is to devise instrumentation able to combine two or more techniques, enabling the simultaneous observation of multiple parameters *in situ* on the same sample.

5.7 FURTHER READING

R. Leach, Fundamental Principles of Engineering Nanometrology, Elsevier, Amsterdam, 2009.

R.E. Kunz, Miniature integrated optical modules for chemical and biochemical sensing, Sens. Actuators B 38–39 (1997) 13–28.

J.J. Ramsden, High resolution molecular microscopy, in: Ph. Dejardin (Ed.), Proteins at Solid-Liquid Interfaces, Springer-Verlag, Heidelberg, 2006, pp. 23–49.

J.J. Ramsden, OWLS—a versatile technique for drug discovery, Front. Drug Des. Discovery 2 (2006) 211–223.

Nanomaterials and their Production

CHAPTER CONTENTS

A nanomaterial is a material having one or more external dimensions in the nanoscale or having internal or surface structure in the nanoscale (Figure 6.1). If the description is to be kept deliberately vague, one could use the word "nanosubstance" merely to signify that some nanoscale features are involved. Nano-objects are categorized according to the number of their external dimensions in the nanoscale (Figure 6.2). The concept of nano-object includes additional entities not shown in the figure but sometimes referred to, such as nanohorns, whose external dimension in the nanoscale tapers from large to small—that is, a hollow nanocone, except that the principal axis may not be straight. The concept of nanoplate includes all ultrathin coatings, even if their substrate is not planar, as long as the substrate's radius of curvature is much greater than the thickness of the coating.

    A nanostructured material is defined as possessing internal or surface structure in the nanoscale. This structure may arise due to the presence of contiguous elements with one or more dimensions in the nanoscale. Primary atomic or molecular (in the chemical sense) structure is excluded (otherwise all materials would necessarily be

Nanomaterial

Nano-object   Nanostructured material
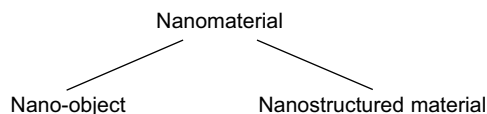
FIGURE 6.1

Fragment of a concept system for nanotechnology (cf. Figure 1.1).

Nano-object (one or more
external dimensions in the
nanoscale)

Nanoparticle
(3 external dimensions in
the nanoscale)

Nanofiber
(2 external dimensions in
the nanoscale)

Nanoplate
(1 external dimension in
the nanoscale)

Nanowire
(electrically conducting
nanofiber)

Nanotube
(hollow nanofiber)

Nanorod
(rigid nanofiber)
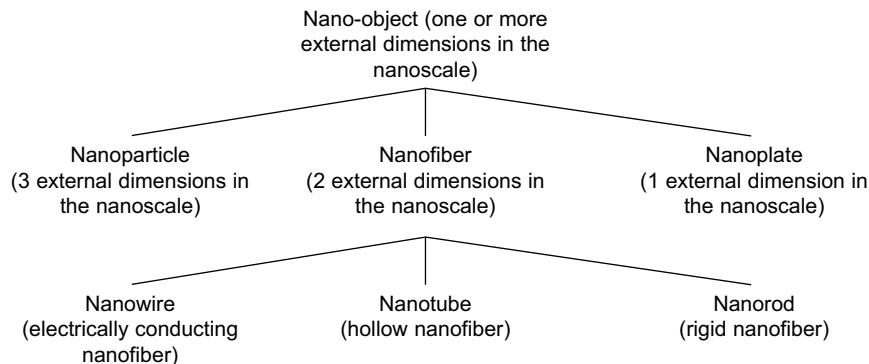
FIGURE 6.2

Concept system for nano-objects (cf. Figures 1.1 and 6.1). See text for further explanation.

categorizable as nanostructured). Ultimately nanostructured materials should be produced using bottom-to-bottom technology (i.e., mechanosynthesis, see Section 8.3), but currently this is only available for producing minute quantities of matter. At present, the largest exemplar of nanostructured materials is composites comprising nano-objects embedded in a matrix. In order to distinguish them from materials in which the embedded objects are larger than nanoscale, nanostructured composites may be called nanocomposites. In such materials, the overall structure is of a statistical nature. It would not, however, be appropriate to call such materials "randomly nanostructured", since almost certainly spacial correlations exist.

Since the above definition refers to "contiguous elements", it would appear that a heap of nano-objects would also fall into the category of nanostructured material. If the attractive energy of interaction between the objects is insignificant compared with thermal energy (i.e., $mgd > k_B T$, where m and d are, respectively, the mass and diameter of the object), one may refer to a powder (or an ultrafine powder or a nanopowder if the smallness of the objects needs to be explicitly stated, or indeed "nanoscale granular material"). If the objects attract each other weakly, one may refer to an agglomerate; typically the external surface area of an agglomerate is similar to the sum of the surface areas of the individual objects comprising it. If the objects are strongly bonded to one another one may refer to an aggregate (typically the external

Table 6.1 The ontology of nanodispersions

Matrix	State of the nano-objects		
	Solid	Liquid	Gaseous
Solid	Nanocomposite^a	–	Nanofoam^b
Liquid	Nanosuspension^c	Nanoemulsion	Nanofoam^b
Gaseous	Aerosol	Aerosol	–

^{a}Other materials falling into this category are nano-alloys, metal–matrix composites, etc.
^{b}These materials may be described with the adjective "nanoporous" (or, in French, using the noun "nanoporeux").
^{c}Nanofluids are included here (Section 6.5.4).

surface area of an aggregate is significantly smaller than the sum of the surface areas of the individual objects prior to aggregation). In terms of interaction energies (Figure 4.4), the particles in an agglomerate are separated by a distance ℓ_1, and those in an aggregate by a distance ℓ_0.

If a material comprises two or more distinct phases, if one or more dimensions of the zones of the phases are in the nanoscale one may refer to a nanostructured material. Certain alloys would fall into this category but not, presumably, solid solutions.

If nano-objects are dispersed into a continuous medium (matrix), one may describe the result as a nanodispersion. Depending on the states of the objects and the matrix, different terms are used to describe them (see Table 6.1).

The remainder of this chapter firstly covers the fabrication of the various kinds of nano-objects mentioned above, except for carbon-based materials, which are dealt with in a separate chapter (9) because of their unique importance and versatility. Secondly, the fabrication of composites is described.

6.1 NANOPARTICLES

Currently one can use either a top–down (comminution and dispersion, see Section 6.1.1) or a bottom–up (nucleation and growth, see Section 6.1.2) approach. The decision which to adopt depends, of course, on which can deliver the specified properties, and on cost. A third approach is bottom-to-bottom mechanosynthesis; that is, "true" nanofacture, according to which every atom of the material is placed in a predefined position. At present this has no commercial importance because only minute quantities can be prepared. Nevertheless, it is the only approach that would forever solve the imperfections associated with top–down and bottom–up processes (the presence of defects, polydispersity, etc.).

Despite being the largest current commercial segment of nanotechnology, nanoparticles have as yet relatively few direct large-scale commercial uses (excluding

photographic emulsions and carbon black, both of which long precede the nano-era); mostly their applications are in composites (i.e., a mixture of component A added to a matrix of component B, the latter usually being the majority component)—a nanocomposite differs from a conventional composite only insofar as the additive is nanosized and better dispersed in the matrix. Applications such as reagents for the remediation of contaminated soils (e.g., iron nanoparticles for decomposing chlorinated hydrocarbons) are being investigated, although the present very imperfect knowledge about the effects of such nanoparticles on soil ecosystems is preventing rapid exploitation in this area.

6.1.1 Comminution and Dispersion

This top–down approach involves taking bulk material and fragmenting it. Crushing and grinding have typically been treated as low-technology operations; theoretical scientists seeking to formalize the process beginning with the formulation of mechanistic phenomenological rules (e.g., the random sequential fragmentation formalism) have hitherto had little industrial impact. The main advantages are universality (i.e., applicability to virtually any material) and low cost. Even soft organic matter (e.g., leaves of grass) can be ground by first freezing it in liquid nitrogen to make it brittle.

The main disadvantages are polydispersity of the final particles, the introduction of many defects (including contamination by the material used to make the grinding machinery—the smaller the particles the worse the contamination because it is introduced at the surface of the particles) and the impossibility of achieving nanoscale comminution, depending on the material: as a compressed brittle body is made smaller, its fracture strength increases until at a certain critical size crack propagation becomes impossible [90], cf. Section 2.7. This explains the well-known size limit to the crushing of materials—typically above the nanoscale (e.g., around $0.8\ \mu$m for calcium carbonate).

Crushing and grinding are venerable industrial processes in the form of hammering, rolling or ball-milling, but the advent of nanotechnology has given rise to novel, very well-controlled methods of achieving monodisperse nanoparticle generation by comminution and dispersion. One such process is electroerosion dispersion (EED) [120], in which granulated metal is ground into a fine powder by electrical discharges—typically a few hundred volts are discharged in a microsecond. The plasma temperature in the discharge filament is 10 000 to 15 000 K, sufficient to melt any metal.

6.1.2 Nucleation and Growth

This process involves a first order phase transition from an atomically dispersed phase to a solid condensed phase. During the first stage of the transition fluctuations in the homogeneous, metastable parent phase result in the appearance of small quantities of the new phase. The unfavorable process of creating an interface opposes the gain in energy through the reduction in supersaturation of the parent phase, leading to a
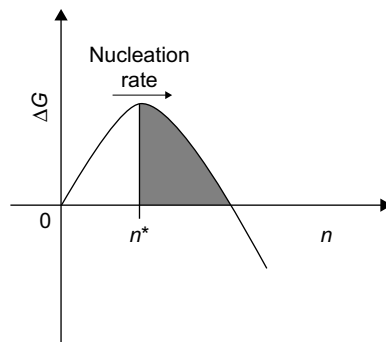
FIGURE 6.3

Sketch of the variation of free energy of a cluster containing n atoms (cf. Figure 2.2). The maximum corresponds to the critical nucleus size. Clusters that have managed through fluctuations to climb up the free energy slope to reach the critical nucleus size have an equal probability to shrink back and vanish, or to grow up to microscopic size.

critical size of nucleus, n^*, above which the nucleic develop rapidly and irreversibly into the new "bulk" (albeit nanoscale) phase. Many syntheses were reported in the 19th century (e.g., [50, 63]).

When atoms cluster together to form the new phase, they begin to create an interface between themselves and their surrounding medium, which costs energy. Denoting the interfacial tension by γ, and using subscripts 1 and 2 to denote the new phase and surrounding medium, respectively (see Section 3.2), the energy cost is $A\gamma_{12}$, where A is the area of the cluster's surface, equal to $(4\pi)^{1/3}(3nv)^{2/3}$, where n is the number of atoms in the cluster, and v the volume of one atom. At the same time each atom contributes to the cohesive energy of the new phase. Summing these two contributions, at first the energy will increase with increasing n, but ultimately the (negative) cohesive energy of the bulk will win (Figure 6.3).

In order to synthesize nanoparticles via nucleation and growth, firstly the atoms are dispersed (dissolved) in a medium under conditions such that the dispersion is stable. Then, one or more of the external parameters is changed such that the bulk phase of the material now dispersed is stable. This could be accomplished, for example, by cooling the vapor of the material. The formation of the new bulk phase is a first order phase transition involving nucleation. Chance fluctuations will generate critical nuclei (see Figure 6.3).

Compound particles can be synthesized by chemical reaction [135]. Suppose the formula of the desired substance is MX, where M represents a metal such as silver or cadmium, and X a metalloid such as sulfur or selenium. One then prepares two solutions of soluble compounds of M and X (for example, silver nitrate and sodium sulfide), which are then mixed together. Figures 6.4 and 6.5 show some examples.

Two key challenges in this process are: (i) to obtain particles that are as uniform (monodisperse) as possible; and (ii) to be able to control the mean size. In the case of
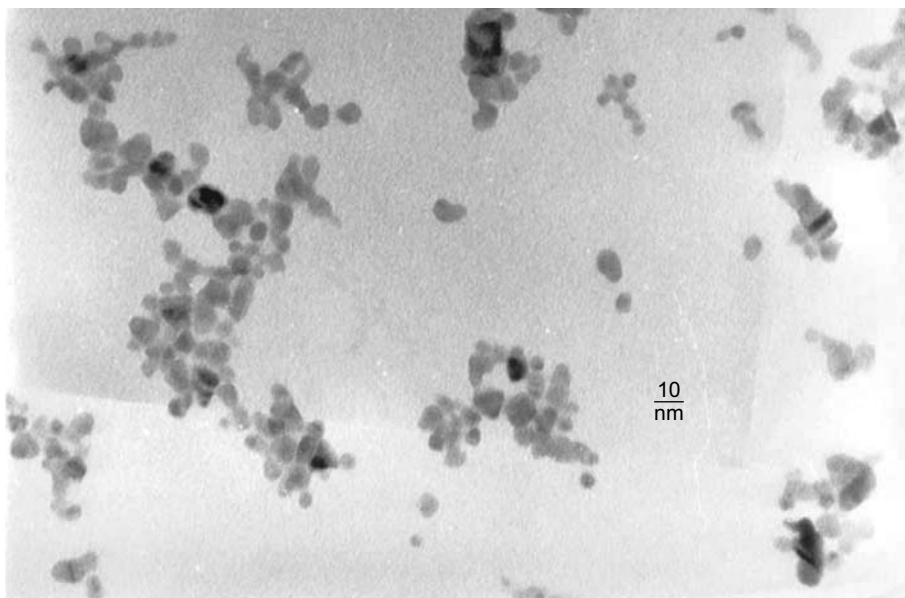
FIGURE 6.4

Transmission electron micrograph of cadmium sulfide nanoparticles prepared in the author's laboratory by nucleation and growth in aqueous solution [135]. The apparent agglomeration is an artifact of the sample preparation, in which a drop of the particle suspension was placed on a carbon-coated copper grid and the water allowed to evaporate. This final stage of the evaporation process can be observed *in situ* in the electron ultramicroscope: the water forms a receding thin film and capillary forces (Section 3.3) at its edge drag the particles together.

synthesis by chemical reaction, the key parameter is the rate of mixing. Two extreme situations yield the desired monodispersity: ultrarapid mixing of very concentrated solutions, and ultraslow mixing of very dilute solutions. In the former case, a very large number of critical nuclei are formed almost simultaneously (the rate of creation of critical nuclei is proportional to the supersaturation; that is, the ratio of the actual concentration to the solubility product of MX); growth of material onto the initially formed nuclei is too slow to compete with fresh nucleation in sinking the added mass. Conditions should be chosen such that the nuclei are just able to grow sufficiently large to be effectively irreversibly stable before all the free M and X ions have been scavenged by the formation of nuclei. Further growth to any desired size can then be achieved in a separate, subsequent, stage by adding fresh material at a rate just sufficient to allow all the nuclei to grow without creating any new ones.

In the latter case, nuclei are formed extremely rarely and are unable to grow beyond the size of minimum stability because of the lack of material; diffusion of fresh material to the few nuclei formed initially is too slow to prevent new nuclei
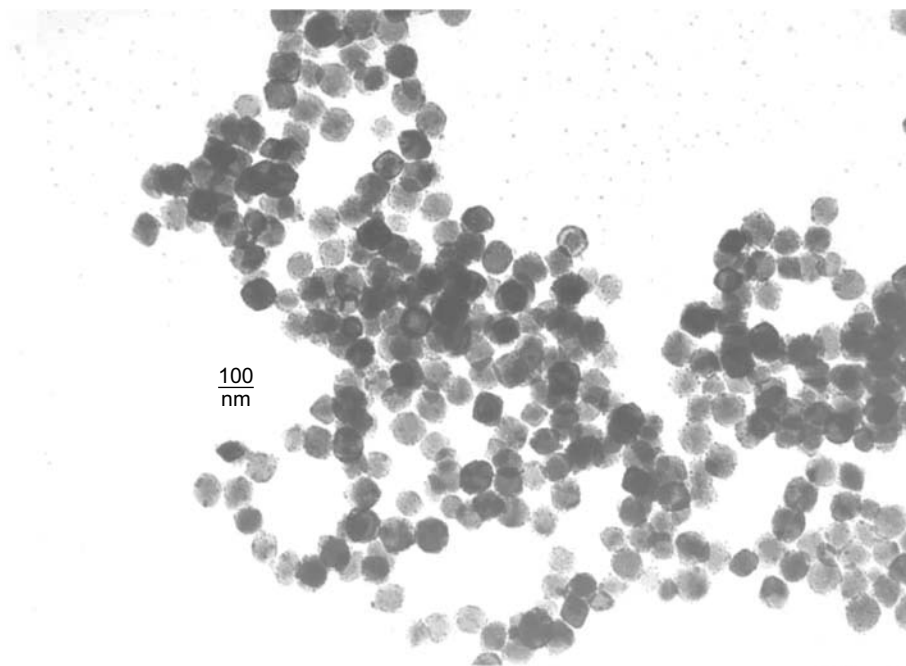
FIGURE 6.5

Transmission electron micrograph of hematite nanoparticles prepared in the author's laboratory by nucleation and growth in aqueous solution (cf. Figure 6.4).

been formed in order to sink the added reagents. Once a sufficient number of nuclei has been synthesized, they can be grown up to the desired size as in the previous case. This approach is very effective for synthesizing monodisperse noble metal particles (e.g., gold) by very slowly reducing the solution of a salt of the metal [50].

Because of the Kelvin relation (equation 2.7), larger particles will have a slightly lower solubility than smaller ones. Therefore, there will be a slight tendency for the smaller ones to dissolve, and for their material to be deposited onto the bigger ones. This process is known as Ostwald ripening and under certain conditions may permit the size distribution of a collection of particles to be narrowed, albeit at the price of increasing the mean size.

Once a collection of nuclei has been synthesized, it is very easy to grow shells of different materials around them; one simply needs to ensure that the new material is added at a sufficient rate to allow all the particles to grow uniformly, and not so rapidly that fresh nuclei are formed.

The interfacial free energy for aggregation of particles made from material 1 in the presence of medium 2 is given by (see Section 3.2.1):

$$\Delta G_{121} = \Delta G_{11} + \Delta G_{22} - 2\Delta G_{12} \tag{6.1}$$

where ΔG_{11} and ΔG_{22} are the cohesive energies of materials 1 and 2, and ΔG_{12} is the solvation energy. Note that water has a very large cohesive energy. Therefore, particles of almost any insoluble material synthesized in water are likely to aggregate, unless appropriate measures to ensure their hydration are taken. A useful strategy is to synthesize the particles in the presence of a very hydrophilic material such as polyethylene glycol or a polyion such as hexametaphosphate, which is able to adsorb on the surface of the particles and effectively hydrate them. Michael Faraday's famous synthesis of gold nanoparticles used citrate ions to hydrate their surface [50]. Crystals of silver chloride, silver bromide and silver iodide ranging in size from tens of nanometers to micrometers, which form the basis of conventional silver halide-based photography, are stabilized in the emulsions used to coat glass or polymer substrates by the natural biopolymer gelatin.

Micelles and superspheres are dealt with in Section 8.2.9.

6.2 NANOFIBERS

"Nanofiber" is the generic term describing nano-objects with two external dimensions in the nanoscale. A nanorod is a rigid nanofiber, a nanotube is a hollow nanofiber, and a nanowire is an electrically conducting nanofiber (Figure 6.2).

Two approaches are presently mainly used to synthesize nanofibers. For some substances, under certain conditions, the natural growth habit is acicular. Therefore, the methods described in Section 6.1.2 can be used to generate nuclei, followed by a growth stage to elongate them. Heterogeneous nucleation can be induced at the solid/gas interface by predepositing small catalytic clusters. Upon addition of vapor, condensation on the clusters and growth perpendicular to the solid substrate takes place. This is used as an efficient way of synthesizing carbon nanotubes (see Section 9.2). If uniform nanopores can be formed in a membrane (e.g., by laser drilling or by self-assembly) they can be used as templates for nanofiber formation. The material for the fiber should be deposited as a shell on the inner surface of the pores (if the goal is to make nanotubes), or else should completely fill the pores (for nanorods). Nanofibers, especially nanorods, formed by either of the two previous methods can also be used as templates for making nanotubes of a different material. Lieber has reported the general synthesis of semiconductor nanowires with control of diameter [44, 66].

6.3 NANOPLATES AND ULTRATHIN COATINGS

Many of the traditional engineering methods of fabricating thin coatings on a substrate have not produced objects in the nanoscale because typically they have been more than 100 nm thick. Nevertheless, the trend is to develop thinner functional surfaces by coating or otherwise modifying bulk material, and insofar as the coating or

modification is engineered with atomic precision, it belongs to nanotechnology, and if it is nanostructured either laterally or perpendicular to the substrate, it will rank as a nanomaterial even if its overall thickness exceeds 100 nm.

The surface treatment of bulk material, especially metals, is an ancient technology. In the cases where nanoscale structural features were apparently necessary to ensure having the required attributes (e.g., in Damascus swords [146]), although the structures were created deliberately the nanoscale aspect might be considered as essentially inadvertent since the technologist is unlikely to have explicitly envisioned the nanoscale structuring. This is in contrast to, for example, the medicinal use of nanoparticulate gold (introduced by Paracelsus in the 16th century), when it was realized that a metal could only be assimilated by a living human organism if it was sufficiently finely divided.

Completely in the spirit of nanotechnology are the monomolecular layers, now called Langmuir films, transferred to solid substrata using the Langmuir–Blodgett and Langmuir–Schaefer techniques; these films might only be a few nanometers thick. Their preparation is a "top–down" approach.

Many physical vapor deposition techniques (such as an evaporation and magnetron sputtering) create films thicker than the nanoscale and hence are out of the scope of this book. Molecular beam epitaxy (MBE), a technique of great importance in the semiconductor processing industry, is briefly covered in Section 6.3.1 (see also Section 8.1.1).

6.3.1 Molecular Beam Epitaxy (MBE)

MBE can be considered as a precise form of physical vapor deposition (PVD). Solid source materials are placed in evaporation cells around a centrally placed, heated, substrate. Pressures less than 10^{-5} torr ensure that the mean free path of the vapor exceeds practical vacuum chamber dimensions ($\sim 1\,\text{m}$), the molecular beam condition. Ultrahigh vacuum (UHV) conditions are needed to ensure the absence of contamination from residual gases (from the chamber walls, etc.). A few seconds are typically required to grow one monolayer. The technique has been developed very successfully using a practical, empirical approach—thermodynamic analysis is difficult because the various parts of the system (sources, substrate, chamber wall) are at different temperatures.

6.3.2 Langmuir Films

Langmuir films, first reported by Pockels at the end of the 19th century, consist of a monomolecular layer of amphiphiles (molecules consisting of an apolar nanoblock conjoined to a polar block of roughly the same size) floating on water. The polar "heads" dip into the water and the apolar "tails" stick up into the air. In other words, the film precursors are molecules of general formula XP, where X is (typically) an apolar chain (e.g., an alkyl chain), called the "tail", and P is a polar "head" group such as oligoethylene oxide, or phosphatidylcholine. When spread on water they mostly

FIGURE 6.6

Langmuir–Blodgett films. (a) A monolayer; (b) a bilayer; (c) a Y-type multilayer. The circles represent the polar heads and the squares the apolar tails of the amphiphilic molecule. (d) A micelle, which can form spontaneously upon dispersal in water if the amphiphilic molecules have a smaller tail than the head (see Section 8.2.9).

remain at the water/air interface, where they can be compressed (e.g., using movable barriers) until the molecules are closely packed to form two-dimensional liquid-like and solid-like arrays. Slowly withdrawing a hydrophilic plate perpendicular to and through the floating monolayer from below will coat the plate with a packed mono-layer (Figure 6.6(a)), as was extensively investigated by Blodgett; these supported layers are called Langmuir–Blodgett (LB) films. The Langmuir–Blodgett technique refers to the transfer of the floating monomolecular films to solid substrata by verti-cally dipping them into and out of the bath. In the Langmuir–Schaefer (LS) technique, the substratum is pushed horizontally through the floating monolayer (Figure 6.7). Very stable multilayer films can be assembled by making P a chelator for multiva-lent metal ions, which bridge lateral neighbors and/or successive layers (assembled head–head and tail–tail). Lateral stability can be increased by UV-irradiation of films with an unsaturated alkyl chain (photopolymerization). The process of perpendicular traversal of the floating monolayer can be repeated many times to build up multilayer films, provided the close packing of the monolayer is maintained, e.g., by adding fresh material *pari passu* in an appropriate fashion. Exceptionally strongly laterally cohesive and rigid Langmuir films can be manipulated as free-standing objects.

According to the nanocriterion of "absence of bulk" (Section 2.2), a few layers (Figure 6.6(b)) would be nanoscale. But even though the multilayer (Figure 6.6(c)) could in principle be of microscopic or even macroscopic thickness, it has no bulk equivalent. Simply throwing amphiphilic molecules into a large vessel and stirring would lead to a highly disordered material and it would take an inordinately long time to thermally anneal it into any kind of ordered structure. Consider a Langmuir–Blodgett film made with two types of fluorescent polar head, let us call them A and B (suppose that B is represented by the black circles in Figure 6.6), with differ-ing optical absorption and emission spectra, such that A's emission can be absorbed
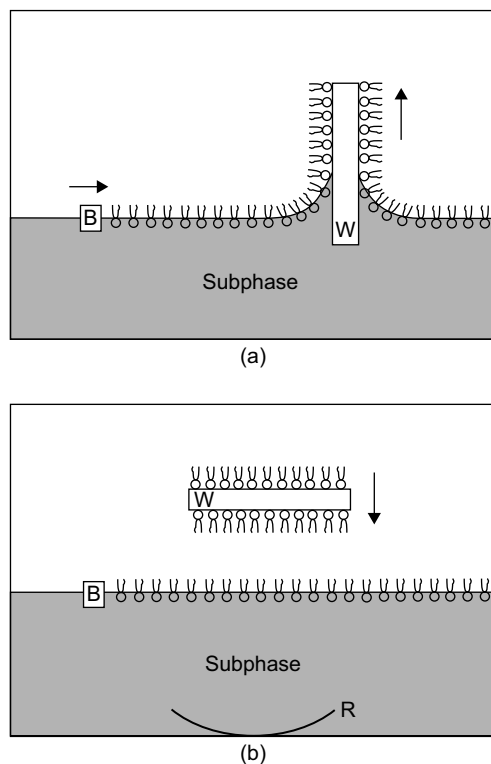
FIGURE 6.7

The Langmuir–Schaefer (LS) technique. (a) The polar substrate is slowly drawn upwards through the Langmuir film to deposit a monolayer. (b) The coated substrate, now apolar, is pushed horizontally through the Langmuir film to deposit a second monolayer (and is caught by the receptacle R). This operation is equivalent to a very rapid vertical descent. The rapidity is needed to ensure that the mensicus curves downwards during deposition.

by B (the process is called Förster energy transfer), which in turn would emit light at wavelength $\tilde{\lambda}_B$. Then, exciting A (with light of wavelength λ_A) would result in emission of wavelength $\tilde{\lambda}_B$. On the other hand, if A and B were simply mixed at random, irradiating the mixture with λ_A would mostly result in emission of wavelength $\tilde{\lambda}_A$, since A and B have to be very close to each other in order for Förster energy transfer to occur: *nanoscale structure* results in qualitatively different behavior. Amphiphilic molecules can also form a bicontinuous cubic phase, which is pervasively nanostructured, although it can be of indefinite (i.e., macroscopic) extent.

If two different immiscible amphiphiles are mixed on the Langmuir trough, the resulting variegated structures can be imaged using scanning probe microscopy after
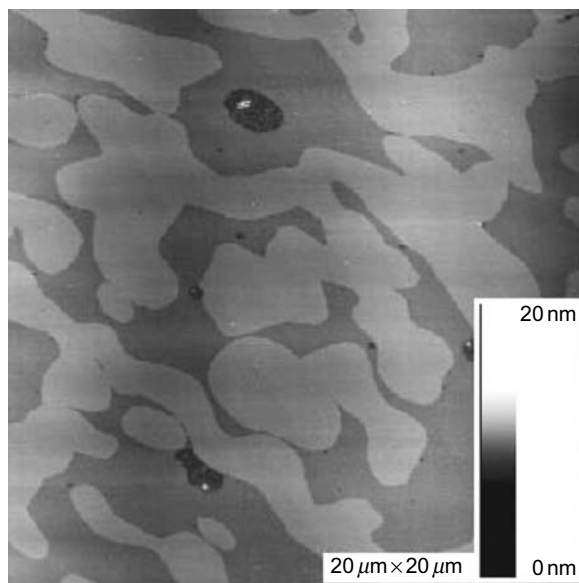
FIGURE 6.8

Scanning force micrograph (contact mode) of a mixed behenic acid/pentadecanoic acid film transferred horizontally onto muscovite mica at a surface pressure of 15 mN/m. Image size: $20 \times 20 \, \mu$m (reproduced with permission from [3]).

transferring the Langmuir film to a solid support using the Langmuir–Schaefer technique (Figure 6.8). These experiments have shown (a) how difficult it is to find convenient mixtures, and (b) how difficult it is to predict theoretically what patterns result. There are moreover experimental difficulties in imaging texture (see Chapter 5) at scales below a few tens of nanometers. The lack of metrology techniques suited for this purpose is in fact one of the main current hindrances to progress in the field.

Although there has been recurrent industrial interest in Langmuir–Blodgett (LB) films for numerous applications in microelectronics and integrated optics, the cumbersome nature of the fabrication methodology, well suited to laboratory research but difficult to automate, and the difficulty of creating defect-free films, have militated against its adoption. The problem of defects (e.g., pinholes in the insulating layer of a field-effect transistor (FET), Figure 7.6) can, however, be overcome by reducing the component area (Section 10.4), suggesting that nanoscale electronics might see a revival of interest. One attractive feature is the versatile chemistry allowing environmentally responsive films to be synthesized; bio-inspired films in a chem-FET configuration can be responsive to both total captured particle number and their state of aggregation (see, e.g., [138]). Bio-inspired LB films are typically in the

liquid crystalline phase, in which defects are self-annealed. Moreover, for electronics applications only very thin films are needed, whereas an optical waveguide requires ∼ 100 monolayers, the defect-free assembly of which is cumbersome.

Post-processing LB films expands the fabrication possibilities. Blodgett herself explored "skeleton" films prepared by selectively dissolving away one component of a mixed film. Any multivalent metal or metal compound nanoplate can be prepared from a metal–fatty acid LB film (e.g., dysprosium behenate) which is then pyrolysed to remove the organic moiety.

Under certain conditions, regular striated patterns can be formed during LB deposition [156]. The cause is dynamic instability of substrate wetting.

6.3.3 Self-Assembled Monolayers (SAMs)

If particles randomly and sequentially added to the solid/liquid interface are asymmetrical, i.e. elongated, and having affinity for the solid substratum at only one end (but the ability to move laterally at the interface), and with the "tail" (the rest) poorly solvated by the liquid, they will tend to adsorb in a compact fashion, by strong lateral interaction between the tails (Figure 6.9). This is a practical procedure of some importance for modifying the surfaces of objects fabricated by other means. The precursors are molecules of general formula XR, where X is (typically) an apolar chain (e.g. alkyl), and R is a ligand capable of binding to a substratum. Addition of XR to the metal surface results in a closely packed array of XR. The film is stabilized by hydrogen or chemical bonds to the substrate, and lateral LW forces between the X.

FIGURE 6.9

A (fragment of a) self-assembled monolayer. The component molecules have the general formula LXR, where X is an apolar chain (e.g., alkyl), and R is a reactive group capable of binding to the substratum S. X can be functionalized at the end opposite from R with a group L to form molecules L–XR; the nature of L can profoundly change the wetting properties of the SAM.

Self-assembled monolayers were discovered by Bigelow et al. [18]. Currently the two main types of R are –SH (thiol or mercaptan), which binds strongly to Au, Ag, Pt, Cu, Hg, etc.), and organosilanes, which bind strongly (bond covalently) to silica. These chemical requirements are the main constraints limiting the versatility of the technology. The original example was eicosyl alcohol ($C_{20}H_{41}OH$) dissolved in hexadecane ($C_{16}H_{34}$) adsorbing on silica glass. Later, molecules with R = –SH and the organosilanes were investigated. If the tail moiety X is poorly solvated by the liquid, its flexibility may enable it to be compactified while the molecule is in the bulk solution, tending to prevent self-aggregation, and only unfurling itself after R is attached to the solid surface—a rudimentary form of programming (cf. programmable self-assembly, Section 8.2.8). SAMs provide a very convenient way to change the wettability of a solid surface. Bigelow et al.'s monolayers were both hydrophobic and oleophobic. An octadecanethiol (R = –SH) film adsorbed on gold would be both oil and water repellent; if L = –OH it will be hydrophilic (see Section 3.2.1).

X can be functionalized at the end opposite from R with reactive groups to form molecules RXL. These can profoundly change the wetting properties of the assembled monolayer. For example, whereas octadecanethiol (L = –H) films are both oil and water repellent, if L = –OH then oil and water will spread. If they are bulky, the functionalized molecules should be mixed with unfunctionalized ones to avoid packing defects. Mixtures with different chain lengths (e.g., X = C_{12} and C_{22}) give liquid-like SAMs. The biologically ubiquitous lipid bilayer membrane could be considered to belong to this category. The component molecules are of type XR, where X is an alkyl chain as before, and R is a rather polar "head group", the volume of which is typically roughly equal to that of X. Placing XR molecules in water and gently agitating the mixture will spontaneously lead to the formation of spherical bilayer shells RXXR called vesicles. The vesicles will coat a planar hydrophilic substratum with a lipid bilayer when brought into contact with it [38].

SAMs can be patterned using photolithography, or "stamping" (microletterpress), to create patterns on substrata (e.g., gold and/or silica) to which the SAM precursor molecules will bind, leaving other zones free. In this procedure, the required pattern is the first created in relief on a silicon substrate, which is then used as a mould for the elastomeric polymer PDMS (polydimethylsiloxane). The SAM molecules can be used directly as ink to coat the projecting parts of the relief pattern, which is then stamped onto the substratum, or else the ink is some substance that passivates the substratum with respect to the SAM molecules, which then selectively bind as required.

6.4 CRYSTALLIZATION AND SUPRAMOLECULAR CHEMISTRY

It is a remarkable fact that many or even most organic molecules, despite their usually very complicated shapes, are able to spontaneously form close-packed crystals.

Perhaps only familiarity with the process prevents it from occupying a more prominent place in the world of self-assembly. In seeking to understand this remarkable phenomenon better, Kitaigorodskii formulated an Aufbau principle [92], according to which the self-assembly (cf. Section 8.2.1) of complicated structures takes place in a hierarchical fashion in the following sequence:

Stage 0 a single molecule (or a finite number of independent molecules).
Stage 1 single molecules join up to form linear chains.
Stage 2 the linear chains are bundled to form two-dimensional monolayers.
Stage 3 the two-dimensional monolayers are stacked to form the final three-dimensional crystal.

Many natural structures are evidently hierarchically assembled. Wood, for example, derives its structural strength from glucose polymerized to form cellulose chains, which are bundled to form fibrils, which are in turn glued together using lignin to form robust fibers. Note that the interactions between the molecular components of the crystal may be significantly weaker than the covalent chemical bonds holding the atomic constituents of the molecules together. Such weakness enables defects to be annealed by locally melting the partially formed structure. There is an obvious analogy between "crystallization" in two dimensions and tiling a plane. Since tiling is connected to computation, self-assembly, which can perhaps be regarded as a kind of generalized crystallization, has in turn been linked to computation.

Just as quantum dots containing several hundred atoms can in some sense (e.g., with regard to their discrete energy levels) be regarded as "superatoms" (Section 2.5), so can supramolecular assemblies (typically made up from very large and elaborate molecules) be considered as "supermolecules". The obtainable hierarchically assembled structures—an enormous literature has accumulated—provide powerful demonstrations of the Kitaigorodskii Aufbau principle. The use of metal ions as organizing centers in these assemblies has been a particularly significant practical development.

6.5 COMPOSITES

Most of the materials around us are composites. Natural materials such as wood are highly structured and built upon sophisticated principles. The basic structural unit is cellulose, which is a polymer of the sugar glucose, but cellulose on its own makes a floppy fabric (think of cotton or rayon), hence to give it strength and rigidity it must be glued together into a rigid matrix. This is accomplished by the complex multi-ring aromatic molecule lignin. Another example is the shell of marine molluscs such as the abalone. They are composed of about 95% by volume of thin platelets, small enough to rank as nano-objects, of aragonite, a form of calcium carbonate and about 5% of protein. The toughness of the composite seashell exceeds by more than tenfold the toughness of pure calcium carbonate.

A striking thing about these natural composites is the enormous increase in desirable properties, such as stiffness or toughness, through small (by volume or mass) inclusions of another material that would, by its nature, normally rank as the matrix in artificial composites. In contrast, the matrix in artificial composites is usually the majority component (in mass or volume or both). The principle of combining two or more pure substances with distinctly different properties (which might be mechanical, electrical, magnetic, optical, thermal, chemical, and so forth) in order to create a composite material that combines the desirable properties of each to create a multifunctional substance has been refined by humans over millennia, presumably mostly by trial and error. Typically, the results are, at least to a first approximation, merely additive. Thus we might write a sum of materials and their properties like

	cellulose	high tensile strength	self-repellent
+	lignin	weak	sticky
=	wood	strong	cohesive

Empirical knowledge is used to choose useful combinations, in which the desirable properties dominate—in principle one might have ended up with a weak and repellent material. The vast and always growing accumulation of empirical knowledge usually allows appropriate combinations to be chosen. Indeed, the *motif* of strong fibers embedded in a sticky matrix is very widely exploited, other examples being glass fiber- and carbon fiber-reinforced polymers.

There are two principal motivations for creating composites. The first is to increase the toughness of the majority component (as in the seashell). This occurs by dividing up the monolithic material into nano-objects and gluing the objects together. The increase in toughness comes partly through the increased ductility of the comminuted objects (Sections 2.7 and 6.1.1) and partly through the fact that any fracture that does occur can only extend to the object boundary or, if it occurs in the matrix, to the nearest object lying in its path, a distance which might well be shorter than the object size. Particularly in this kind of composite the wetting of the principal phase by the glue is of extreme importance. Empirical knowledge is now backed up and extended by fundamental knowledge of the molecular-scale forces involved, but natural materials still far surpass artificial ones in this respect. Proteins in particular are very versatile glues because of the variety of chemical functionalities possessed by amino acids. The second motivation is the combination of properties—these composites might well be called composites to distinguish them from the other kind; for example ultrahard nanoparticles might be added to a relatively soft polymer matrix to create a hard, plastic material.

6.5.1 Polymer–Nano-Object Blends

In fact, most of the recognized successes in nanomaterials so far have been not in the creation of totally new materials through mechanosynthesis (see Section 8.3), which is still an unrealized goal, but in the more prosaic world of blending, which

is the simplest form of combination. For example, one adds hard particles to a soft polymer matrix to create a hard, abrasion-resistant coating. As with atomically-based mechanosynthesis, the results are, to a first approximation, additive. Thus we might again write a sum like

	polypropylene	flexible	transparent
+	titanium dioxide	rigid	opaque
=	thin film coating (paint)	flexible	opaque

This is not actually very new. Paint, a blend of pigment particles in a matrix (the binder), has been manufactured for millennia. What is new is the detailed attention paid to the nanoparticulate additive. Its properties can now be carefully tailored for the desired application. If one of components is a recognized nanosubstance—a nanoparticle or nanofiber, for example—it is acceptable to describe the blend as a nanostructured material. Other applications of this kind include ultralow permeability materials for food packaging, for which it is often undesirable to restrict the ingress of oxygen. The main principle here is to blend particles of a very high aspect ratio (fragments of nanoplates rather than nanoparticles) into the polymer such that the principal plane of the objects is perpendicular to the route of ingress. The tortuosity of gas diffusion is vastly increased in such materials. Very often it is sufficient to apply an ultrathin coating onto a conventional flexible polymer substrate to achieve the desired diminution of gas permeability. The design of friction damping composites requires careful consideration of the viscoelastic properties of the material. "Self-cleaning" and anti-grafitti coatings rely on the active, typically photo-induced, chemistry taking place at the surface of the nanoparticles included in the matrix.

The Paradigm of Paint. The biggest range of applications for such nanocomposites is in thin film coatings—in other words paint. This is a much older composite than most other nanomaterials. Paint consists of a pigment (quite possibly made of nanoparticles) dispersed in a matrix of varnish. Paint can be said to combine the opacity of the pigment with the film-forming capability of the varnish.

Traditional pigments may comprise granules in the micrometer size range; grinding them a little bit more finely (if possible—see Section 6.1.1) turns them into nano-objects. Compared with transparent varnish, paint combines the attribute of protection from the environment with the attribute of color. The principle can obviously be (and has been) extended practically *ad libitum*: by adding very hard particles to confer abrasion resistance; metallic particles to confer electrical conductivity; tabular particles to confer low gas permeability, and so on.

The purpose of adding materials to a polymer matrix is, then, to enhance properties such as stiffness, heat resistance, fire resistance, electrical conductivity, gas permeability, and so forth; the object of any composite is to achieve an advantageous combination of properties. If the matrix is a metal, then we have a metal–matrix composite (MMC, Section 6.5.2). A landmark was Toyota's demonstration in 1991 that the incorporation of a few weight percent of a nanosized clay into a polyamide matrix greatly improved the thermal, mechanical (e.g., doubled the tensile modulus) and gas

permeability (barrier) properties of the polymer compared with the pure polymer or the conventional composite made with micrometer-sized additive. This was the first commercial nanocomposite.

Another mineral–polymer composite is the material from which many natural seashells are constructed: platelets of aragonite dispersed in a protein matrix (see also Section 6.5.5). In this case, however, the "matrix" only constitutes a few percent of the volume of the composite.

There is no general theory suggesting that the advantage scales inversely with additive size; whether a nanocomposite is commercially viable depends on all the parameters involved. There is such a huge variety of materials that it is perhaps futile to attempt a generalization. However, the very small size of individual nanoparticles would make it feasible to incorporate a greater variety of materials within the matrix for a given additive weight percent. Very often, ensuring wetting of the particle by the matrix presents a significant technological hurdle. Most successful composites require the additive to be completely wetted by the matrix. Wetting behavior can be predicted using the Young–Dupré approach (see Section 3.2); if, however, the particle becomes very small, the surface tension will exhibit a curvature-dependent deviation from the bulk value appropriate for a planar interface.

The three main fabrication routes for polymeric nanocomposites are:

1. Blending preformed nanoparticles with the matrix, which is typically in the molten state.
2. Dispersing preformed nanoparticles in monomer and polymerizing the matrix around the nanoparticles (the Toyota composite mentioned above followed this route: modified clay was swollen in the presence of caprolactam followed by polymerization).
3. Synthesizing the nanoparticles *in situ* within the matrix.

In each case the goal is to disperse the composites uniformly in the matrix. Hence the chemistry of the particle–composite interface is very important.

Nanocomposites can in principle be substituted for pure polymer in most applications. Nevertheless, there is persistent reluctance to use them in structurally vital components since extensive data on their long-time performance is still lacking. A general difficulty is that the improved strength, stiffness, etc. of the composites inevitably results in their being subjected to increased designed stress—otherwise there would be no advantage in using them—and hence diminished tolerance to damage.

Polymers with Added Electro-Active Materials. This is a typical functionality-enhancing technology. Antimony tin oxide (ATO) has become a popular additive for applications such as protection against electromagnetic interference (EMI) and electrostatic discharge (ESD). Antistatic coatings incorporate additives with both hydrophobic and hydrophilic radicals, ideally concentrated in the interfacial zone between the polymer and the air: the hydrophobic radicals are oriented towards the polymer and the hydrophilic radicals towards the air, whence they attract moisture, whose conductivity allows static charge accumulation to be dissipated.

Polymers made conductive by the addition of electroactive materials are of growing importance in the automotive industry. If used for body panels, for example, electrostatic spray painting is enabled, thereby reducing the need for primers, and which in turn improves painting reliability, reduces costs and lessens the impact on the environment by using less volatile organic solvents. Conductive plastics have also been used to make fuel pipes for automobiles, eliminating the build-up of static electricity and the danger of inadvertent ignition from a spark.

Much effort is currently being devoted to creating photovoltaic solar cells from semiconductor nanoparticles embedded in a matrix (the dielectric constant of which must be less than that of the particles if the electrons are to be properly routed). The main advantage is that if the nanoparticles are small enough to show size-dependent optical absorption (see Section 2.5), particles of different sizes can be blended to optimally cover the solar spectrum. In addition, the fabrication of a nanoparticle–polymer composite should be significantly cheaper than the physical vapor deposition of a thin semiconductor film on a substrate.

6.5.2 Metal–Matrix Composites

Conventional metal–matrix composites use two types of reinforcement: microparticles—usually silicon carbide—dispersed in an aluminium or magnesium matrix to about 10–15 vol%; or continuous fibers—the most studied system is titanium with SiC reinforcement. The existing materials show enhancements of a few % in stiffness, about 20% in strength and in superior wear resistance and fatigue initiation behavior.

Pure magnesium reinforced with 30 nm alumina particles at up to 1 vol% gives Young's modulus improvements of about 20%, yield strength increases of 45% and some ductility improvement as well. The wear and creep performance is also improved significantly in both aluminium and magnesium alloys. Silicon carbide nanoparticles used to reinforce magnesium result in a 100% increase in microhardness for 5 vol% SiC nanoparticles. SiC nanowire reinforcement of ceramics improves toughness and strength by up to 100% for a few vol%.

Although extensive empirical work has been conducted, the parameter space is so vast (especially if one considers that *distributions* of sizes and shapes, such as mixing particles and rods of different lengths, may offer better toughness and ductility) that our current knowledge must be considered as being very limited. One general principle that has emerged is that a much smaller volume fraction of nanoparticles is required compared with microparticles of the same chemical substance to achieve equivalent improvements in strength and stiffness. Thus the degradation of toughness and ductility will not be as extreme and these nanomaterials may be more damage tolerant.

It is important that the particles are well dispersed and well bonded to the matrix in order to create good reinforcement. Ultrasonic waves are used for dispersion of nanoparticles in the melt. There are interactive influences of the intrinsic properties of reinforcement and matrix and the size, shape, orientation, volume fraction and

distribution of the reinforcement. Usually the reinforcement causes diminution of the ductility and toughness compared with the matrix alone. In the case of high-strength magnesium and aluminum alloys ductility is already limited; consequently attempts have been made to use ductile pure metals as the matrix.

Principle process routes are:

- **Stir casting:** the nano-objects are distributed and suspended in the molten metal, e.g., via high-energy mixing. Melt sizes as large as 7 tons are practicable. The slurry is then cast as billets or rolling bloom. Products with a volume fraction of reinforcement ranging from 10 to 40% are available. The microstructure of the stir cast has to be finely controlled to develop uniform distributions of the particles. Spray casting followed by hot extrusion is also used.

- **Liquid metal infiltration:** requires a preform consisting of a mat or other assembly of the ceramic fibers and particles to be infiltrated. A volume fraction of between 40 and 70% is required to provide sufficient mechanical stability to withstand the infiltration forces without being crushed. Liquid metal infiltration is extensively used by the automotive industry, and is now a preferred process for the thermal management industry as it has the ability to produce near net shape components. Because of the relatively high volume fraction MMCs produced by this process, applications tend to be those requiring wear resistance and high stiffness, rather than good toughness and ductility. The majority of applications use aluminum as the matrix material.

- **Powder metallurgical routes:** these are used to produce continuous and discontinuous MMC in aluminum and titanium matrices. Sintered preforms can be hot extruded to produce rods suitable for testing.

Current applications in the ground transportation industry account for 62% of the MMC market by mass. Aerospace applications are 5% by mass, and general industrial markets comprise about 6% by mass (including cemented carbide and ceramic–metal composite (cermet) materials for general tooling).

6.5.3 Self-Repairing Composites

The basic principle is the incorporation of hollow fibers and microspheres into conventional polymer matrix composites. These are presently typically microsized (20–50 μm in diameter) rather than nanosized. Fibers and microspheres contain fluorescent dyes and polymer adhesives. If the composite is damaged by fiber or resin cracking, the cracks intersect the hollow fibers and microspheres, which liberates the dyes and adhesives, providing more visible indications of damage than the cracks themselves, and self-repairs the matrix cracks. Catalysts for the cure of the adhesive can be placed in the resin separated from the fibers and capsule contents until the moment of release by damage. The concept has also been applied to self-repair concrete structures.

Manufacture of such materials is relatively easily accomplished on a laboratory scale. Hollow glass fibers of 20–50 μm diameter can be easily made, and hollow

carbon fibers may also be made. Microcapsules containing adhesive can be created using *in situ* polymerization, followed by incorporation into a conventional structural epoxy matrix material (Section 6.5.1). The required infiltration of the repair adhesive into the hollow fibers is often a difficult manufacturing step.

The self-repair action repairs only cracks in the polymer matrix; there are no suggested routes to the repair of continuous fibers. It is important that the repair material has sufficient time to flow from the containment vessels into the damaged region and has the time, which is very possibly long, of the order of one day, to cure. In the case of microspheres, it is important that the crack intersects and fractures them rather than propagating between them; this will depend on local elastic conditions at the crack tip.

If successful repair can be demonstrated to have high structural integrity, it is envisaged that these composites will be used in aircraft structures and in other applications where high structural integrity in safety-critical structures is required. It is difficult to envisage how these concepts can be extended into metallic structures, but they could be used in civil concrete structures. Presently the technology is in its infancy. The approaches used so far are very primitive, and there needs to be substantial further development work before they could enter service. In particular, nondestructive techniques for evaluating the strength of the repair are required, together with engineering data on the delamination resistance, the fatigue resistance and the environmental resistance of the repair.

The concept of dye bleeding for damage detection works reasonably well, but an obvious disadvantage is the lack of capability to distinguish damage needing attention from damage that can be left in service. Current synthetic realizations are a very long way from the type of self-repair schemes used in biological structures, which show vastly better flaw tolerance (equation 2.25).

6.5.4 Nanofluids for Thermal Transport

Nanofluids are suspensions of nanoparticles or nanofibers in a liquid medium. The addition of nanoparticles should substantially enhance the heat capacity of a fluid, and hence nanofluids have been proposed for heat transfer fluids. The reported enhancements of critical heat flux are especially interesting in boiling heat transfer. Applications are found wherever cooling is important: in microelectronics, transportation, solid-state lighting, and manufacturing. Thermal interface materials (TIM) are essentially very viscous nanofluids used to eliminate asperity between a surface and a heat sink. Some reports of the degree of thermal conductivity enhancement are at variance with established theoretical predictions, however: both the experiments and the theory need to be scrutinized carefully. Current theory cannot explain the strong measured temperature dependence of the thermal conductivity of the nanofluids [167].

The nanoparticles may be fabricated in dry form and then dispersed in the fluid, or they may be produced *in situ* in the liquid. The latter route was established in the Soviet Union some decades ago. About 1% by volume of copper nanoparticles or

carbon nanotubes dispersed in ethylene glycol or oil increases the thermal conductivity by 40% (copper) and 150% (carbon nanotubes). Similar enhancements using ordinary particles would require more than 10% by volume. The nanofluids are also much more stable than suspensions of conventional particles.

6.5.5 **Alternating Polyelectrolyte Deposition**

If the substratum is electrified (via the Gouy–Chapman mechanism) and the dissolved molecule is a polyion with an electrostatic charge of opposite sign, then it will adsorb on the surface and invert the charge; the strong correlations within the polymeric ion render the Gouy–Chapman mechanism invalid [65]. The polyion-coated substratum can then be exposed to a different polyion of opposite sign, which will in turn be adsorbed and again invert the charge; the process can be repeated *ad libitum* to assemble thick films [143].

This method of *alternating polyelectrolyte deposition* (APED), invented by Iler [84], appears to have immense potential as a simple, robust method of surface modification. It requires the substrate to be electrostatically charged when immersed in water. It is then dipped into an aqueous solution of a polyelectrolyte of opposite charge, with which it rapidly decomes coated. Any excess is then washed off, and the coated substrate is dipped into a polyelectrolyte of the opposite charge, with which it now becomes coated, and whose excess is again washed off, and so on (Figure 6.10).

There are few restrictions on the choices of polyelectrolytes. Much early work was done with polyallylamine as the polycation and polystyrene sulfonate as the polyanion. The essential feature of the technique is that at each dipping stage the substrate charge is not only neutralized but reversed ("overcharging"), hence allowing the deposition to be repeated indefinitely. This phenomenon contradicts the predictions of the mean-field theories—Gouy–Chapman and Debye–Hückel—of the distribution of ions in the vicinity of charged surfaces ("electrified interfaces"). The discrepancy arises because the charges of the polyions are correlated. Imagine a polyion approaching a surface already covered with its congeners. The new arrival will repel the already adsorbed ones, creating a correlation hole (i.e., a negative image) permitting attraction (Figure 6.11) [65]. Since the gaps between already adsorbed polyions may be smaller than the size of one polyion, adsorption results in charged tails, ensuring overcharging (Figure 6.12).

(Monovalent) counterions screen the polyions in the usual Debye–Hückel fashion, diminishing the charging energy of the polyion more than its correlation energy, enhancing the charge inversion. (If the monovalent counterion concentration is very high the correlation disappears and APED is no longer possible.) Multivalent counterions are more difficult to treat theoretically and APED in their presence would appear to be a fruitful area of investigation. The hydrogen ion may play a special role; for example, it has been found that the porosity of built layers can be reversibly controlled by varying ion concentration (e.g., pH).

Instead of polymeric polyions, nanoparticles composed of materials with ionizable surface groups can be used. In this case, although the electrostatic charge of
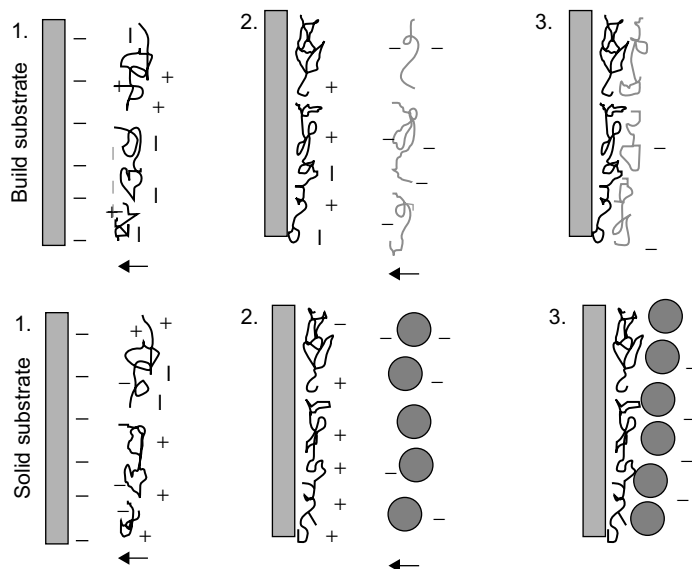
FIGURE 6.10

Upper panel: deposition of a polycation onto a negatively charged substrate followed by a polyanion. Lower panel: deposition of a polycation followed by a negatively charged nanoparticle onto a negatively charged substrate [112].
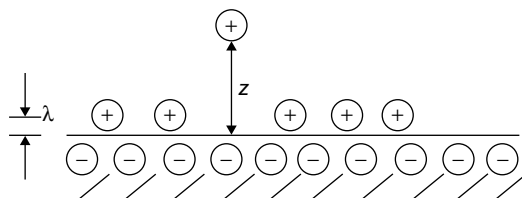
FIGURE 6.11

A polyion approaching a surface already covered with its congeners (see text).
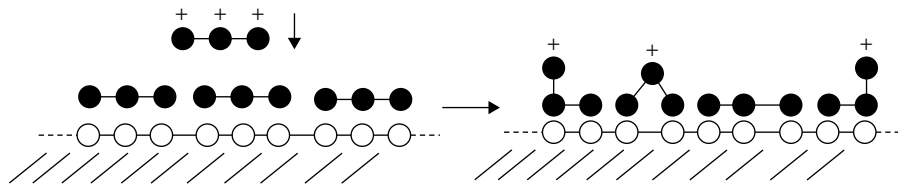
FIGURE 6.12

Overcharging resulting from adsorbed polyion tails (see text).

the surface of the coating is always reversed, typically not all the inner charges are compensated because of steric hindrances, and hence electrostatic charges build up, and the long range electrostatic force ultimately prevents further particles from being deposited [116]. This is an example of a two-dimensional supersphere (Section 8.2.9).

If polymeric polyions are used as the polyelectrolyte of one sign, and ionizable particles as the polyelectrolyte of the opposite sign, the particles act as stress concentrators, thus greatly increasing the toughness of the built material. Large aspect ratio nanoparticles are very useful for diminishing the deleterious effects of defects (pinholes) in multilayer films. In this way sophisticated coatings can be built up. It has been proposed that the shells of many marine organisms, such as the abalone, are assembled using this principle, producing materials that are both robust (cf. equation 2.25) and beautiful: anisotropic nanoparticles are dispersed in a biopolymer matrix, which only occupies a few volume percent of the total mass. Natural biopolymers, which are nearly all heteropolymers, primarily based on amino acids as monomers, but also possibly incorporating polysaccharides and nucleic acids, can incorporate enormous functional variety in ways that we can only dream about at present in synthetic systems. Nevertheless, developments are moving in this direction, as exemplified by the APED variant known as the surface sol-gel process [99].

6.6 SUMMARY

The terminological framework for nanomaterials is given. The two main divisions are nano-objects and nanostructured materials. The various kinds of nano-objects (particles, fibers and plates) are described, along with their methods of manufacture. The most important nanostructured materials are currently nanocomposites. Fabrication and applications are discussed. It must be emphasized that there is a vast body of detailed knowledge concerning this area, part of which is published but unpublished work undertaken in commercial confidence probably preponderates. Wetting of the embedded nano-objects by their matrix is of crucial importance, regarding which present synthetic capabilities reveal themselves as far inferior to what nature can do. Composites with enhanced mechanical properties require careful consideration of how load is transferred between matrix and particle or fiber, and how many cycles of repeated loading and unloading fatigue the material.

6.7 FURTHER READING

J.D. Swalen, et al., Molecular monolayers and films, Langmuir 3 (1987) 932–950.

Nanodevices

A device, fundamentally, is a synonym for an information processor or transducer. Other synonyms are machine or automaton. The most advanced devices are able to adapt to their environment. Adaptation implies sensing (gathering information), processing (information), and actuating (transducing information into physical action) functions. In the analysis of living organisms, or in robotics, or more generally in cybernetics these different functions are often considered as separate units connected via information channels. In the ultraminiature realm of nanotechnology separation might not be realizable, nor indeed is it necessary, except possibly as an aid to

conceptual thinking. Analogous situations are encountered in integrated circuits. A traditional a.c. to d.c. converter is a little circuit with a number of clearly separable components. In an integrated device, a piece of pyroelectric material suffices to fulfill the same function. Sensorial nanomaterials [103] already exist (sometimes they are called "smart" materials), combining the sensing, processing and actuating functions without clearly separating them physically. An excellent example of such an adaptive material, actually belonging to the realm of chemistry rather than of nanotechnology, is viscostatic motor oil, containing some randomly coiled long chain hydrocarbons. If the temperature rises the chains unfurl, thereby increasing the viscosity so as to compensate for the generic viscosity-decreasing effects of temperature. Adaptation, of course, confers on the device a certain independence of being, hence it deserves to be called an entity rather than a mere object.

Apart from this ultimate integration of the macroscopically inseparable functions involved in adaptation, it is especially appropriate to miniaturize the hardware associated with information processing, because there is no definite lower bound to the physical embodiment of one bit of information, considered as its ultimate, irreducible quantum. A single electron or photon can perfectly well carry one bit, even though it lacks extension. This is the basis of the special strength of the connexion between nanotechnology and information technology.

No clue as to the extent of a device is given by the fundamental definition. A single neuron or logic gate is as much a device as a brain or a computer. In this chapter we shall, however, interpret "device" as essentially a rather basic part of an information processor, up to the complexity of a logic gate. A nanodevice is a device with at least one overall dimension in the nanoscale, or comprising one or more nanoscale components essential to its operation. A nanosystem (cf. Chapter 10) is a system of nanodevices, or any system, the nanoscale features of which are essential to its function. Figure 7.1 summarizes the main functional categories of devices.

In this chapter, logic gates and data storage devices are also roughly classified according to the nature of their internal information carriers (e.g., electrons or photons). Their applications are mostly either "pure" information processing (e.g., a logic gate) or sensory (e.g., detection of a magnetic field, or a chemical). "Smart" materials are covered in Section 4.2 (e.g., for drug delivery) or in Section 6.5 (e.g., self-repairing structures).
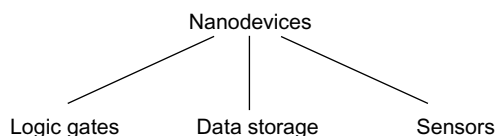
FIGURE 7.1

The main functional categories of devices. Energy transducers might also have been included as a separate category.

A further possible classification is according to the number of terminals: typically either two (e.g., a diode) or three (e.g., a transistor), the latter corresponding to a machine with input in cybernetic terminology. Another one introduces the role of internal states and device history: level 1 devices have no internal states and the output only depends on the input (e.g., a resistor); the output of a level 2 device depends on both input and its internal state (e.g., a condenser); the output of a level 3 device also depends on the history of its inputs (e.g., a memristor).

One may note that most current activity involving nanodevices is taking place in research laboratories, with the potential for showing dramatic industrial growth in the future.

The first section of this chapter briefly surveys some general consequences of miniaturization. This is given special consideration due to the particular importance of digital information processing as an application of nanotechnology. The main alternative is quantum computation, which is summarized in the next section. Specific devices able to execute information processing operations, based on electron charge or spin or on photons are then considered. Finally nanomechanical (relays and sensors) and fluidic (mixers) devices are described. The concluding sections of the chapter deal with sensors and energy-transducing devices.

7.1 ISSUES OF MINIATURIZATION

This section may be read in conjunction with Section 1.5.2.

Surface Relative to the Bulk Interior. An object is delineated by its boundary. Making an object small has an effect on purely physical processes in which it is involved. For example, suppose a spherical object of radius r is heated by internal processes, and the amount of heat is proportional to the volume $V = 4\pi r^3/3$. The loss of heat to the environment will be proportional to the surface area, $A = 4\pi r^2$. Now let the object be divided into n particles. The total surface area is now $n^{1/3}4\pi r^2$, hence more heat is lost. This is the basic reason why small mammals have a higher metabolic rate than larger ones—they need to produce more heat to compensate for its relatively greater loss through the skin in order to keep their bodies at the same steady temperature. It explains why few small mammals are found in the cold regions of the Earth.

Catalysis and Heterogeneous Chemistry. Reactions take place at surfaces (i.e., the interface between the catalyst and the reaction medium, the latter considered to be three-dimensional). Indeed in some cases the main catalytic effect is due to the reduction in dimensionality imposed on the system [140]. In general, the greater the degree of nanostructuring the greater the preponderance of surface per unit mass.

Ballistic Transport (cf. Section 7.4.1). Usually, carriers cannot move through a medium without encountering some resistance, which is caused by the carrier particles colliding with (being scattered by) obstacles (which might be their congeners). The characteristic length associated with this scattering is the mean free path ℓ. If

some characteristic size l of the device is less than ℓ, transport will be ballistic and the carrier can move from one end of the device to the other without encountering any resistance. Similar reasoning can be applied to heat conduction, with resistance normally due to phonon scattering. Since mean free paths in condensed matter are usually of the order of nanometers long, they are likely to be useful candidates for defining the nanoscale (cf. Chapter 2).

How Performance Scales with Size (cf. Section 10.8). Analysis of device performance begins by noting how key parameters scale with device length: area (hence power and thermal losses) as length squared, volume and mass as length cubed, natural frequency as inverse length, and so forth. Relationships such as these are used to derive the way a device's performance scales as it is made smaller [73]. This consideration is apart from qualitatively new phenomena that may intervene at a certain degree of smallness (cf. Chapter 2, Figure 3.1).

Noise in Detectors (Sensors and Dosimeters). Natural phenomena involving discrete noninteracting entities (e.g., the spacial distribution of photons) can be approximated by the Poisson distribution (equation 10.15). A fundamental property of this distribution is that its variance equals its mean. The uncertainty (e.g., of the magnitude of a certain exposure of an object to light) expressed as a standard deviation therefore equals the square root (of exposure).

When objects become very small, the number of information carriers necessarily also becomes small. Small signals are more vulnerable to noise (i.e., the noise is an increasing proportion of the signal). Repetition of a message (e.g., sending many electrons) is the simplest way of overcoming noise. A nanoscale device using only one entity (e.g., an electron) to convey one bit of information would, in most circumstances, be associated with an unacceptably high equivocation in the transmission of information.

An accelerometer (that transduces force into electricity) depends on the inertia of a lump of matter for its function and if the lump becomes too small the output becomes unreliable. Similarly with photodetectors (that transduce photons into electrons): due to the statistical and quantum nature of light, the smallest difference between two levels of irradiance that can be detected increases with diminishing size. On the other hand, for many purposes there is no intrinsic lower limit to the physical embodiment of one bit of information. One bit could be embodied by the presence of a neutron, for example—indeed distinguishable isotopes have been considered as the basis for ultrahigh density information storage [15]. Information processing and storage is therefore the ideal field of application for nanotechnology. The lower limit of miniaturization is only dependent on practical considerations of "writing" and "reading" the information. Hence, nanotechnology is particularly suited to information processors.

Utility. Considering the motor car as a transducer of human desire into translational motion, it is obvious that a nanoautomobile would be useless for transporting anything other than nano-objects. The main contribution of nanotechnology to the

automotive industry is in providing miniature sensors for process monitoring in various parts of the engine and air quality monitoring in the saloon; additives in paint giving good abrasion resistance, possibly self-cleaning functionality, and perhaps novel aesthetic effects; new ultrastrong and ultralightweight composites incorporating carbon nanotubes for structural parts; sensors embedded in the chassis and bodywork to monitor structural health; and so forth.

7.2 DIGITAL INFORMATION PROCESSING

As already mentioned, a device (or component) is fundamentally an information processor, in other words a transducer that encodes (or decodes) information. As noted by Szilard, the elementary "quantum" of information is the binary digit or bit. Binary or Boolean logic, based on zero or one, true or false, presence or absence, and so forth, has very modest physical requirements. There is essentially no intrinsic lower limit to the size of the physical embodiment of "1". Of all technologies, information technology is the one most suited to miniaturizing down to the nanoscale. The main residual problem is heat dissipation (see Section 7.5).

The fundamental component of a digital information processor is the switch, or relay (Figure 7.2). Several relays can be connected together to create logic gates, for example a not–and (NAND) gate, a fundamental component of a binary logic processor. Its characteristics can be summarized in the following truth table:

input 1	input 2	output
0	0	1
1	0	1
0	1	1
1	1	0

Figure 7.3 shows a transistor inverter, the simplest logic gate (NOT), and Figure 7.4 a NOR gate as examples of physical realization.

The relay has the input–output relationship shown in Figure 7.5. The earliest digital computers used electromechanical relays. They are large, slow, expensive, energy-hungry (and hence expensive to run), and unreliable. Frequent operational errors during the execution of programs run with such devices provided Hamming with the inspirational motivation for developing error-correcting codes. Thermionic valves (vacuum tubes) are faster and more reliable, but even more expensive and energy-hungry. The first significant step towards miniaturization was taken with the replacement of relays and valves by solid-state transistors (Figure 7.6). Provided the fabrication process does not pose new difficulties (a clock is usually cheaper to make
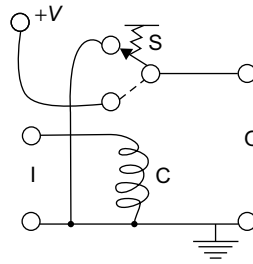
FIGURE 7.2

A relay or switch. When the coil C is energized by applying a voltage across the input terminals I, it pulls the movable contact arm above it to link voltage $+V$ to the output terminals O. If the restoring spring S is present, setting the input I to zero will also cause the output to return to zero. Alternatively, a second appropriately placed coil could be used to move the contact arm in the opposite direction. The device is then bistable and would be suitable for use as a memory element.
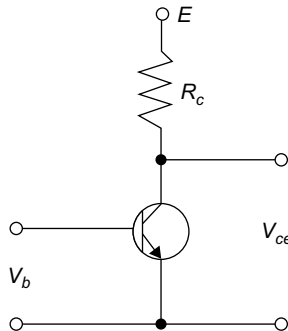
FIGURE 7.3

An inverter. If the voltage V_b on the base is below a certain threshold, corresponding to logic level 0 (typically 0.7 V), the transistor remains in the off state, the current is small and the voltage V_{ce} on the collector is approximately equal to the supply voltage E, typically 5 V, corresponding to logic level 1. If V_b exceeds the threshold then the transistor enters the active region of its characteristic but as V_b increases further (corresponding to logic level 1) V_{ce} saturates at typically around 0.2 V, corresponding to logic level 0.

than a watch), miniaturization uses less material in fabrication and less energy in operation (cf. Section 7.1). At a stroke, the devices became smaller, faster (the electrons carrying the information had less distance to travel), cheaper (not only because the volume of material required was lower, but also because efficient massively parallel fabrication procedures were devised), used less energy, and like all solid-state devices were more reliable (the thermionic valve was more reliable than the relay
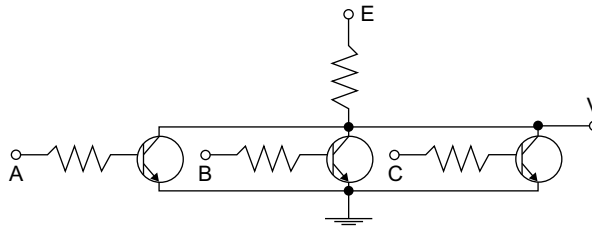
FIGURE 7.4

A NOR logic circuit with inputs A, B and C built according to resistor–transistor logic
(RTL). Note how it is constructed from the basic inverter (Figure 7.3).

FIGURE 7.5

Input–output relationships approaching the ideal of a step (Heaviside) function. The input
might be a voltage (e.g., applied to the coil of a relay, Figure 7.2, or the gate of a transistor,
Figure 7.6) and the output might be the current flowing through the rest of the circuit. Note
that in this example the response characterized by a thick solid line will give an output of
one for any input (voltage) exceeding about 0.6. For the response characterized by the
dashed line, the input would have to exceed about 0.8; i.e., it is less tolerant to deviations
from the ideal (the input of one yielding an output of one). The dotted line marks a possible
hysteretic response when decreasing the input from 1 to 0 and beyond to negative values,
the existence of which opens up the possibility of using the device as a memory element.

because it had no mechanical moving parts, but the vacuum could leak and incandes-
cent electron-emitting filaments could break). A major step in fabrication technology
was the introduction of integration. Miniaturization and, concomitantly, parallel fab-
rication now permits millions of integrated transistors to be fabricated on a single
"chip" of silicon, with additional gains in operation speed because the electrons have
less far to travel, both within and between components.
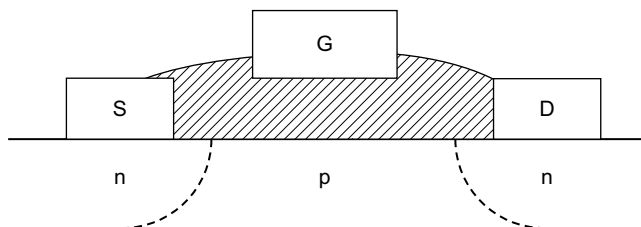
FIGURE 7.6

A field-effect transistor (FET). Regions marked "n" and "p" are n-type and p-type semiconductors (e.g., appropriately doped silicon). The hatched region is an insulator (e.g., silica). Conductors S, G and D are, respectively, the source, gate and drain. Application of a voltage to G (which plays the role of the coil in the relay) increases the concentration of conduction electrons in the p region and allows current to flow from S to D.

A related device is an information store, or memory. A relay or transistor having the property of bistability could function as an information store (memory), with the disadvantage that it would need to be constantly supplied with electrical power. A more elaborate relay, with two separate coils for switching the current "on" and "off", would be better in this regard, since once its position had been flipped, power could be cut off (see Figure 7.5). Read-only memories do not even require the flipping to be reversible: an early type of read-only memory was paper tape in which holes were punched. Reading was carried out by passing the tape between a pair of electrically conducting rollers. In the absence of a hole, there would be no electrical contact between the rollers. A later development was the use of ferromagnets, which could be poled "up" or "down". Since ferromagnetism cannot exist below a certain volume (cf. Section 2.6), this technology is not suitable for ultimate nanoscale miniaturization, but this limit is still far from being reached—the current limitation is the sensitivity of the magnetic field detector (reading head). Memories based on electrical resistance can be fabricated from materials (e.g., NiO) that can be switched from a conducting to an insulating state by applying a voltage pulse. Other materials can have their phase changed from amorphous to crystalline by exposure to light or by passing an electric current, with corresponding changes in reflectance and resistance, but these materials are not especially "nano".

Apart from the logic gates acting as the components of information processors, the other main types of device to be considered are sensors and actuators. A sensor has a clear transduction function. Examples are a magnetic sensor that registers whether the spin in a memory cell is "up" or "down"; a light sensor such as a photodiode that converts light into electricity; a chemical sensor that converts the presence of a certain chemical compound into electricity or light. The main issue in miniaturization is whether the signal exceeds the noise level. An example of an actuator is the coil in the relay (Figure 7.2).
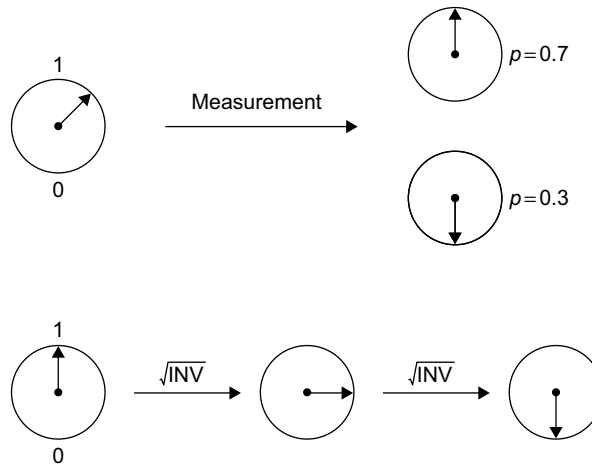
FIGURE 7.7

Qubits. Top: qubits can be represented by spheres (here simplified as disks), a point on the circumference (indicated by an arrow from the center) indicating the net result of the superpositions of all its states. A measurement forces the state into either 1 (with a probability of 0.7 in this example) or 0 (with the complementary probability of 0.3). Bottom: the quantum logic gate. To create an inverter (the NOT operation, cf. Figure 7.3) the arrow must be rotated by $180°$. On the other hand, what might be called $\sqrt{NOT}$ corresponds to rotation by $90°$. This operation, the square root of inversion, has no classical equivalent. Two successive applications of $\sqrt{NOT}$ correspond to classical inversion, NOT.

7.3 QUANTUM COMPUTING

Extrapolation of Moore's law to about the year 2020 indicates that component's sizes will be sufficiently small for the behavior of electrons within them to be perturbed by quantum effects. This implies a profound perturbation of the proper functioning of the technology, and no solution to this problem within the current framework is in view. Quantum computing can be thought of as "making a virtue out of necessity", creating computational devices based on the principles of quantum logic. The essence of quantum computing is to use not binary logic bits as the representation of the elementary unit of information, but qubits, which can exist in an infinite number of superpositions of the binary units zero and one (Figure 7.7).

The key features of quantum objects of interest for computational purposes are superposition—an object can be in several different states simultaneously [42]—and entanglement [6]. Operations can be carried out internally, maintaining superposition, which is only destroyed at the very end of the computation when a single output is required. The system must, however, be kept isolated from its environment during
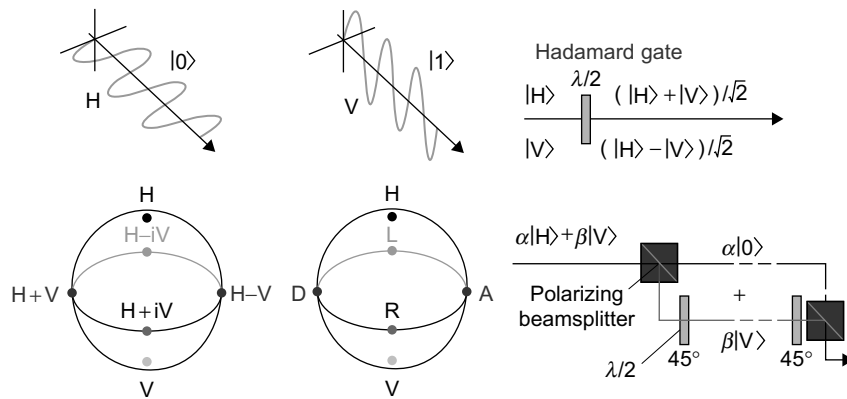
FIGURE 7.8

(Left) optically-encoded qubits. Various combinations of H(orizontal) and V(ertical) polarizations—D(iagonal), A(ntidiagonal), L(eft-circular diagonal), R(ight-circular diagonal)—are represented on the Poincaré sphere. (Right) a birefringent waveplate is used to construct a logic gate, and birefringent waveplates and polarizing beamsplitters can spacially separate photons according to the H ($\equiv 0$) and V ($\equiv 1$) polarizations. Reproduced with permission from [134].

operations. Entanglement with the environment implies decoherence and loss of useful information within the computer. It is easiest to avoid in an isolated small system, hence the interest in realizing quantum computers using nanotechnology.

The architecture of computers needs to be wholly reconceived in order to exploit the peculiarities of quantum mechanics, which means in particular that a particle can exist in two states simultaneously, whereas a cluster of electrons (physically instantiating a bit) in a conventional computer represents either zero or one. The value of a qubit, on the other hand, which might be physically instantiated as a single electron localized on a quantum dot (cf. Section 7.4.5), depends on its position relative to other electrons. For example, two electrons can exist in four different states—00, 01, 10, and 11—depending on their relative positions. If the electrons interact (are entangled) with each other, then any operation carried out on one electron will simultaneously be carried out on the other—implying that one operation is carried out on four different states at the same time. Hence a computer with just 32 bits could perform more than a thousand million operations simultaneously.

The physical embodiment of a bit of information—called a qubit in quantum computation—can be any absolutely small object capable of possessing the two logic states 0 and 1 in superposition, e.g., an electron, a photon or an atom. Electron spin is an attractive attribute for quantum computation. Qubits have also been installed in the energy states of an ion, or in the nuclear spins of atoms. A single photon polarized horizontally (H) could encode the state $|0\rangle$ (using the Dirac notation) and polarized vertically (V) could encode the state $|1\rangle$ (Figure 7.8, upper

left). The photon can exist in an arbitrary superposition of these two states, represented as $\alpha \left| H \right\rangle + \beta \left| V \right\rangle$, with $|\alpha|^2 + |\beta|^2 = 1$. Any polarization can be represented on a Poincaré sphere (Figure 7.8, lower left). The states can be manipulated using birefringent waveplates (Figure 7.8, upper right), and polarizing beamsplitters are available for converting polarization to spacial location (Figure 7.8, lower right). With such common optical components, logic gates can be constructed.

One of the biggest problems with current supercomputers is energy dissipation. They require tens of kilowatts of energy to run and generate vast amounts of heat. Landauer showed in 1961 that almost all operations required in computation can be performed reversibly, thus dissipating no heat. Reversible computation is possible on a quantum computer.

7.4 **ELECTRONIC DEVICES**

Electronic devices (such as the archetypical field effect transistor, Figure 7.6), depend on the movement of elementary packets of charge (i.e., electrons), carrying their electrostatic charge; that is, current flows.

7.4.1 **Ballistic Transport**

Even a simple electrical wire ranks as a device; a transistor (from several of which a logic gate can be constructed, and hence a digital computer) can be thought of as based on a wire whose resistance depends on an external input. In an ideal binary device, the resistance is either infinite (i.e., the conductance is zero) or has some value R; for a piece of material of length l and cross-sectional area A it is given by

$$R = l\rho/A \tag{7.1}$$

where ρ is the resistivity, which depends on the material properties:

$$\rho = 12\pi^3 \hbar/(\ell e^2 S_F) \tag{7.2}$$

where S_F is the Fermi surface area. Note in particular the inverse dependence on mean free path, ℓ. As is well known, an electron moving in a perfect lattice experiences no resistance whatsoever; but lattice defects (e.g., impurities) and the thermal fluctuations inevitably present at any temperature above absolute zero result in an effective mean free path of several tens of nanometers. But if a wire is shorter than a critical length $l_c = \ell$ the resistance becomes ballistic, with a value of $h/(2e^2) = 25.75$ kΩ per sub-band, independent of material parameters. Therefore, any device with a characteristic length smaller than ℓ could justifiably be called a nanodevice.

7.4.2 **Depletion Layers**

The operation of an electronic device typically depends on junctions, including internal junctions between different kinds of materials and junctions between the

device material and its contact with the external world. Some kinds of transistors depend on the field-dependent barrier arising through the generation of a space charge (depletion layer) depending on the difference V_b between the electron Fermi levels of the two materials juxtaposed at the interface and on the sparse presence of carrier-donating impurities (e.g., electron donor centers). The depletion width may be approximated by

$$W = \sqrt{\frac{2\varepsilon_0 \varepsilon V_b}{eN}} \qquad (7.3)$$

where N is the impurity concentration, a typical value of which is $10^{16}/cm^3$. Therefore, for silicon, the relative dielectric constant ε of which is 11.9, W would be about 35 nm for $V_b = 10$ mV. Rectification cannot occur if the device is significantly smaller than W, hence the depletion width provides a demarcating length scale below which different behavior occurs.

7.4.3 Single-Electron Devices

Even if the device becomes very small, and at any moment only has one electron passing through it, it does not qualify as a single electron device unless the electron is isolated by substantial tunneling barriers to localize its otherwise extended wavefunction, as in Figure 7.9.
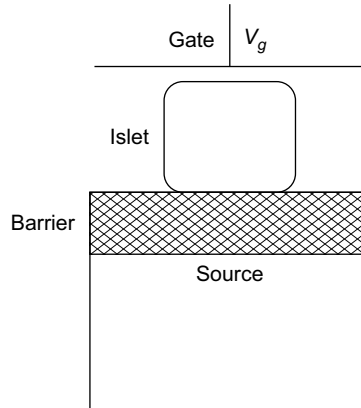
FIGURE 7.9

Single-electron box, consisting of one small semiconductor island separated from a larger electrode ("source") by a tunnel barrier. An external electric field may be applied to the island using another electrode (the gate) separated from the small island by a thick insulator preventing tunneling; the field changes the electrochemical potential of the islet, thereby determining the tunneling conditions.

A key concept is the energy E_a required to add an electron (e.g., to the islet shown in Figure 7.9);

$$E_a = E_c + E_k \tag{7.4}$$

where E_c is the charging energy required to add an electron to an island of capacitance C ($= 4\pi\varepsilon_0 r$ for a sphere of radius r)

$$E_c = e^2/C \tag{7.5}$$

and E_k is the quantum kinetic energy (electron quantization energy); for a degenerate electron gas

$$E_k = 1/(gV) \tag{7.6}$$

where g is the density of states on the Fermi surface and V is the island volume. One would expect thermal fluctuations to swamp single-electron effects (SEE) unless $E_a \geq 10k_BT$. If the island size is somewhat less than about 10 nm, however, E_a is sufficiently large for SEE to be observable at normal room temperature. Under this condition, however, E_k becomes comparable to or greater than E_c (providing the real justification for calling such small islands quantum dots). Not only are such small objects generally difficult to make, but they have to be made with almost exact control of size and shape, to which the transport properties are very sensitive. It is, therefore, preferable to make them large enough for E_a not to be dominated by E_k.

For devices in which information is represented as electrostatic charge, a scalar quantity, the lower limit of its magnitude is the charge e of a single electron. Neglecting noise and equivocation issues, single electron devices can essentially be achieved by downscaling the components of a conventional transistor. Developments in fabrication technologies (Chapter 8) have led to nanoscale devices with the same architecture as their microscopic counterparts. Truly nanoscale devices using electrons involve single-charge transport in minute tunnel junctions.

The simple but effective "orthodox" theory [108] of electron tunneling assumes no electron energy quantization inside the conductors. Strictly speaking this is valid only if the quantum kinetic energy E_k of the added electron is much less than k_BT, but seems to adequately describe observations provided $E_k \ll E_c$, the charging energy e^2/C, where C is the capacitance of the object (island) to which the electron is added (equation 7.5). The second major assumption of the orthodox theory is that the tunneling time is assumed to be negligibly small—typically it is of the order of 1 fs—in comparison with other relevant time scales (such as the interval between successive tunneling events). The third assumption is to ignore coherent quantum processes (simultaneous tunneling events), valid if the barrier resistance R is much greater than the quantum unit of resistance

$$R_Q = h/4e^2 \tag{7.7}$$

which is about 6.5 kΩ. Equation (7.7) ensures the localization of each electron at any particular instant—hence sufficiently nontransparent tunnel barriers effectively suppress the quantum-mechanical uncertainty of electron location. This suppression

makes single-electron manipulation possible, and also guarantees that single-electron devices are not quantum devices in the sense of Section 7.3. Tunneling of a single electron is a random event with a rate (probability per unit time) Γ depending solely on the reduction ΔW of the free (electrostatic) energy of the system as a result of the event and expressed as

$$\Gamma = (I/e)[1 - \exp\{-\Delta W/k_B T\}]^{-1} \tag{7.8}$$

where I is the d.c. current (often satisfactorily given by Ohm's law) in the absence of single-electron charging effects. Useful general expressions for ΔW are [108]

$$\Delta W = e(V_i + V_f)/2 \tag{7.9}$$

where V_i and V_f are the voltage drops across the barrier before and after the tunneling event, or

$$\Delta W = e(V_i - V_t) \tag{7.10}$$

where

$$V_t = e/C_{kl} - e[1/C_{kk} + 1/C_{ll}]/2 \tag{7.11}$$

where C is the capacitance matrix of the system, k and l being the numbers of islands separated by the barrier. Equation (7.8) implies that at low temperatures (i.e., $k_B T \ll \Delta W$) only tunneling events increasing the electrostatic energy and dissipating the difference are possible; their rate is proportional to ΔW, since an increase in the applied voltage increases the number of electron states in the source electrode, which may provide an electron capable of tunneling into an empty state of the drain electrode (Figure 7.10, upper diagram). Note that the simplicity of the above equations does not imply that it is necessarily simple to calculate the properties of single-electron systems—several tunneling events may be possible simultaneously, but we can only get the *probability* of a particular outcome, implying the need for a statistical treatment to yield average values and fluctuations of variables of interest. For large systems with very many possible charge states a Monte Carlo approach to simulate the random dynamics may be the only practical method.

In a nanosized device, electron confinement engenders significant energy quantization (cf. Section 2.5). Experimental work has revealed features not accounted for by the orthodox theory, in particular cotunneling (the simultaneous tunneling of more than one electron through different barriers at the same time as a single coherent quantum-mechanical process) and discrete energy levels (Figure 7.10, lower diagram), resulting in a different energy dependence of the tunneling rate (cf. (7.8)):

$$\Gamma = \Gamma_0[1 + \exp\{-\Delta W/k_B T\}]^{-1} \tag{7.12}$$

where Γ_0 is a constant; the orthodox rate (7.8) is the sum of the rates (7.12) over all the levels of a continuous spectrum of the island. Hence if $\Delta W \gg k_B$, in the miniature device the tunneling rate is a constant.
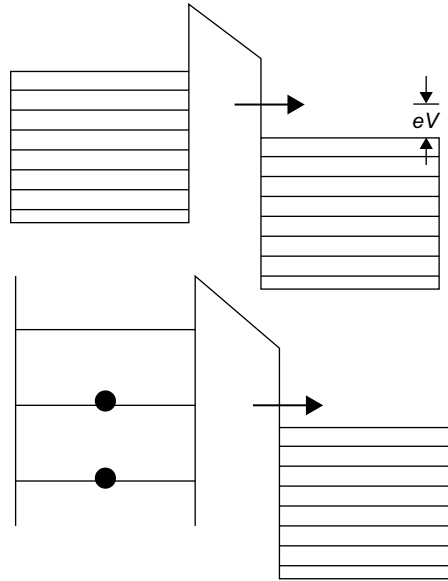
FIGURE 7.10

Top: energy diagram of a tunnel junction with quasicontinuous spectra; bottom: energy diagram showing discrete levels of the source electrode.

The "single-electron box" (Figure 7.9) is the simplest single-electron device. The Gibbs free energy of the system is

$$W = (ne - Q_e)^2/2C_\Sigma + const \qquad (7.13)$$

where n is the number of uncompensated electrons, C_Σ is the total islet capacitance and the parameter Q_e (the "external charge") is given by

$$Q_e = C_0 V_g \qquad (7.14)$$

where C_0 is the islet–gate capacitance.

The average charge $Q = ne$ of the single-electron box increases stepwise with the gate voltage (i.e., external charge Q_e); this phenomenon is called the Coulomb staircase, with a fixed distance $\Delta Q_e = e$ between steps (for $E_c \gg k_B T$). The reliable addition of single electrons despite the presence of thousands of background electrons is possible due to the great strength of the unscreened Coulomb attraction.

Localized states with Coulomb interactions cannot have a finite density of states at the Fermi level, which has significant implications for electron transport within nanoscale material. By definition, at absolute zero all electronic states of a material below the Fermi level are occupied and all states above it are empty. If an additional electron is introduced, it must settle in the lowest unoccupied state, which is above the Fermi level and has a higher energy than all the other occupied states. If, on the

other hand, an electron is moved from below the Fermi level to the lowest unoccupied state above it, it leaves behind a positively charged hole, and there will be an attractive potential between the hole and the electron. This lowers the energy of the electron by the Coulomb term $-e^2/(\varepsilon r)$ where e is the electron charge, ε the dielectric permittivity, and r the distance between the two sites. If the density of states at the Fermi level is finite, two states separated by but very close to the Fermi level could be chosen, such that the energy difference was less than $e^2/(\varepsilon r)$, which would mean—nonsensically—that the electron in the upper state (above the Fermi level) has a lower energy than the electron located below the Fermi level. The gap in states that must therefore result is called the Coulomb gap and materials with a Coulomb gap are called Coulomb glasses.

The single-electron box (Figure 7.9) is unsuitable as a component of an electronic logic device. It has no internal memory; the number of electrons in it is a unique function of the applied voltage. Furthermore the box cannot carry current; hence a very sensitive electrometer would be needed to measure its charge state. The latter problem can be overcome by creating a single-electron transistor (Figure 7.11) [108]. Its electrostatic energy is (cf. equation 7.13)

$$W = (ne - Q_e)^2/2C_\Sigma - eV[n_1 C_2 + n_2 C_1]/C_\Sigma + const \qquad (7.15)$$

where n_1 and n_2 are the numbers of electrons passed through the two tunnel barriers. This is the most important single electron device.

If the source–drain voltage is small no current flows because tunneling would increase the total energy ($\Delta W < 0$ in equation 7.8)—this effect is called the Coulomb blockade. Simplistically this is because one cannot add less than one electron, hence the flow of current requires a Coulomb energy of E_c. The blockade is overcome by a certain threshold voltage V_t. If one gradually increases the gate voltage V_g the source–drain conductance shows sharp resonances almost periodic in V_g. The quasiperiodic oscillations of the I–V curve are known as Coulomb diamonds and are closely related to the Coulomb staircase.

7.4.4 Molecular Electronic Devices

Another approach to ultraminiaturize electronic components is to use a single molecule as the active medium. Current realizations of such molecular electronic
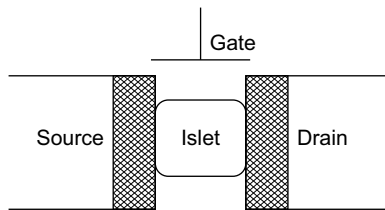
FIGURE 7.11

Single-electron transistor (cf. Figure 7.9). The hatched areas represent tunneling barriers.

devices comprise an organic molecule uniting electron donor (D^+, i.e., a cation) and acceptor (A^-, i.e., an anion) moieties separated by an electron-conducting bridge (typically π, i.e., a π-conjugated (alkene) chain) between a pair of (usually dissimilar) metal electrodes $M^{(1)}$ and $M^{(2)}$, mimicking a semiconductor p–n junction. When a forward bias is applied across the electrodes, chosen for having suitable work functions, the process

$$M^{(1)}/D^+-\pi-A^-/M^{(2)} \rightarrow M^{(1)}/D^0-\pi-A^0/M^{(2)} \tag{7.16}$$
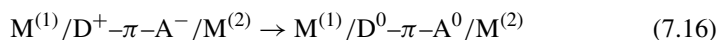
occurs, followed by intramolecular tunneling to regenerate the starting state. Under reverse bias, the energetically unfavorable formation of $D^{2+}-\pi-A^{2-}$ that would be required blocks electron flow, hence we have rectification (Figure 7.12).

7.4.5 Quantum Dot Cellular Automata

The quantum dot cellular automaton (QCA) uses arrays of coupled quantum dots to implement Boolean logic functions [153]. An individual quantum dot can be of the order of 10 nm in diameter, hence QCA devices are very small. Figure 7.13 shows the basic concept. Four quantum dots coupled by tunnel barriers are placed in a square array. This constitutes one QCA cell. Electrons can tunnel between the dots but cannot leave the cell. Two electrons are placed in each cell; Coulomb repulsion ensures that they occupy opposite corners. Hence there are two ground state configurations with the same energy; they can be labeled 0 and 1.
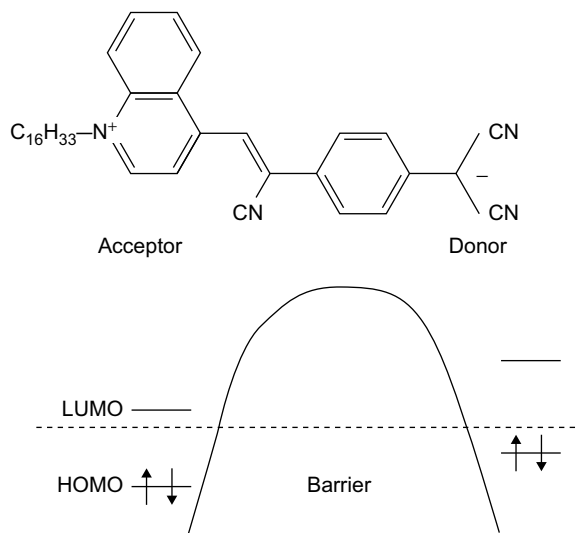
FIGURE 7.12

A typical molecule (Z-β-(1-hexadecyl-4-quinolinium)-α-cyano-4-styryldicyanomethanide) used to create a rectifying junction (see text). The chemical structure is shown at the top. The energetic structure is shown at the bottom.
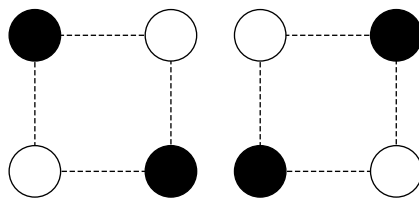
FIGURE 7.13

An illustration of the concept of the quantum cellular automaton. Four quantum dots arranged in a square and occupied by two electrons have two possible ground state polarizations (left and right squares), assigned to represent "0" and "1".
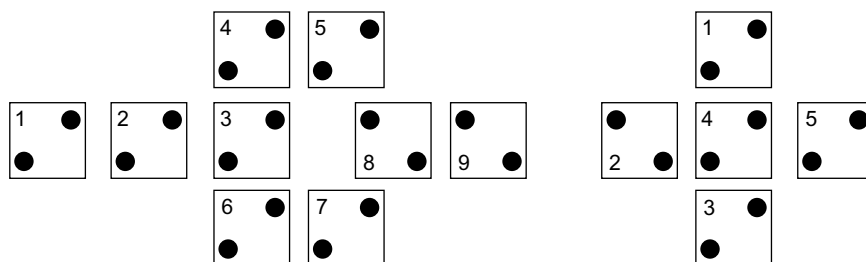
FIGURE 7.14

Some QCA devices. On the left, a wire (cells 1–3) leads input (cell 1) to fanout (cells 3, 4 and 6). When the parallel wires are combined (cells 5 and 7 converging on cell 8), the output (cell 9) is inverted. On the right, we have two inputs (say cells 1 and 2) and a programming input (say cell 3). The central cell 4 cannot minimize all cell–cell Coulomb repulsions but chooses the configuration that minimizes the overall repulsion. Depending on the programming input, the output (cell 5, mimicking the configuration of cell 4) is that of either the AND or the OR logic function.

Figure 7.14 shows several simple QCA devices. On the left we have wires, fanout and an inverter; on the right a majority gate. In each case, the configuration of adjacent cells is determined by minimizing Coulomb repulsion. All logic functions can be implemented using the majority gate and the inverter.

7.5 TRENDS IN THE MINIATURIZATION OF ELECTRONICS

Interest in how device performance scales with size as very small dimensions are approached is often considered to have begun with a paper by Dennard et al. [40]. That paper considered a MOSFET channel length of 1 μm as the ultimate lower limit (reflecting then current manufacturing practice). Since then attention has been

focused on the fundamental lower limit of miniaturization of information processing, for which the starting point is the Shannon–von Neumann–Landauer (SNL) expression for the smallest energy to process one bit of information [174]:

$$E_{bit} = k_B T \ln 2. \tag{7.17}$$

This equals 17 meV at 300 K. Using Heisenberg's uncertainty principle $\Delta x \Delta p \geq \hbar$ and writing the momentum change Δp as $\sqrt{2 m_e E_{bit}}$ one has

$$x_{min} = \hbar / \sqrt{2 m_e k_B \ln 2} \tag{7.18}$$

amounting to 1.5 nm (at 300 K). Similarly with the other uncertainty relation $\Delta E \Delta t \geq \hbar$ and equating ΔE to E_{bit} one finds that the minimal switching time is 40 fs. The power dissipation per unit area is then

$$P = \frac{E_{bit}}{x_{min}^2 t_{min}} \tag{7.19}$$

a colossal several megawatts per square centimeter!

The essential features of classical information processing technology are the physical carrier of information (e.g., the electron), the creation of distinguishable states within the system of carriers, and the capability for a regulated change of state. Distinguishability and regulated state changes are achievable with energy barriers (Figure 7.15). This structure constitutes a switch: when "on", electrons can freely move from one well to the other (the transition probability equals unity); when "off" no movement can take place. If the barrier width b is large enough to prevent tunneling, the probability of the thermal transition from one barrier to the other is $\exp(-E_b/k_B T)$, whence $E_b = k_B T \ln 2$ for a probability of one half, equal to the SNL limit (equation 7.17). Roughly speaking, x_{min} can be equated to b.

The actual switching operation is represented in Figure 7.16. It follows from Poisson's equation linking electrical potential ϕ and charge density ρ

$$\nabla^2 \phi = -\rho / \varepsilon_0 \tag{7.20}$$

that a change of barrier height implies a change of charge q; it also implies charging or discharging a capacitance $C = \Delta q / \Delta \phi$. This applies regardless of the detailed nature of the device (FET, QCA, SET, etc.) and "changing barrier height" means changing height, width or shape; that is, barrier deformation. Inevitably energy is dissipated in the process (e.g., amounting to CV^2 if it is charged from a constant voltage source V) and this energy will be at least as much as the SNL limit (equation 7.17).

When appraising the minimum heat dissipation of the order of 1 MW/cm^2, one should bear in mind that current heat removal technologies can manage at most about three orders of magnitude less than this. Thus, device miniaturization based on the current paradigm of charge transport-based information processing is limited by heat. Spin-based logic is one way to eliminate dissipation (Section 7.6.4). Another way might be to introduce new classes of electron charge-based devices and devise ingenious new ways of carrying out logical operations. The basic variables describing an electronic circuit are current I, voltage V, charge Q and magnetic flux ϕ.
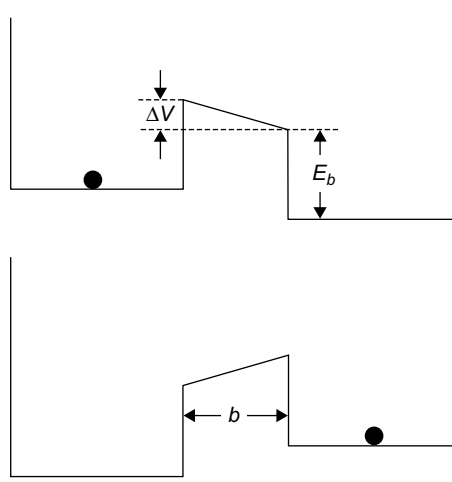
FIGURE 7.15

A very simple model device consisting of two wells separated by a barrier of width b and energy E_b [174]. The electron as the information carrier is represented by the black disk.
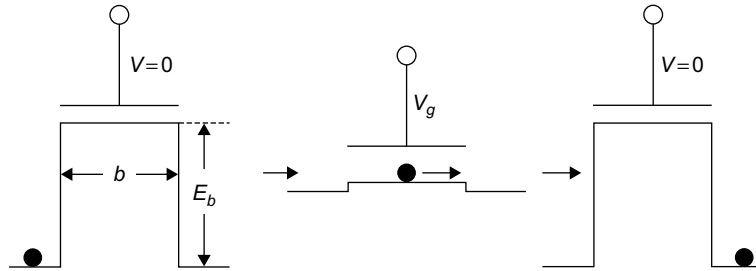
FIGURE 7.16

A model charge transport device with a voltage-controlled barrier of width b. In the absence of an applied voltage the barrier height is E_b, diminishing to practically zero upon application of a voltage V_g [32].

Pairs of these basic variables can be permuted in six different ways: $Q = \int I \mathrm{d}t$ and $\phi \int V \mathrm{d}t$ (definitions of C and ϕ); $R = V/I$ (Ohm's law), $C = Q/V$ ($C = \mathrm{d}q/\mathrm{d}v$ in differential form) and $L = \phi/I$ (properties of the three basic circuit elements, i.e., the resistor, the condenser and the inductor); symmetry demands the existence of a sixth relation, $M = \phi/V$, where M is memristance (memory resistance) which implies the existence of a fourth basic circuit element, the memristor, which shows resistance hysteresis as the applied voltage increases and then diminshes, thus endowing it with memory. Memristive components may allow one to create components fulfilling certain logic functions combining several existing classical components.

Although the memristor is still considered a curiosity among electronics engineers, the phenomenon of memristance is widely encountered in biological systems.

7.6 SPINTRONICS (MAGNETIC DEVICES)

Electrons have spin as well as charge. This is of course the origin of ferromagnetism, and hence magnetic memories, but *their* miniaturization has been limited not by the ultimate size of a ferromagnetic domain but by the sensitivity of magnetic sensors. The influence of spin on electron conductivity was invoked by Nevill Mott in 1936 [122], but remained practically uninvestigated, much less exploited, until the discovery of giant magnetoresistance (GMR) in 1988. Spintronics, sometimes called magnetoelectronics, may be loosely defined as the technology of devices in which electron spin plays a role; it has three main directions now:

- The development of ultrasensitive magnetic sensors for reading magnetic memories.
- The development of spin transistors, in which barrier height is determined by controlling the nature of the spins of the electrons moving across it.
- The development of devices in which logical states are represented by spin.

7.6.1 Ultrasensitive Magnetic Sensors

The discovery of the giant magnetoresistance effect (GMR) in 1988 can be considered as the beginning of the spintronics era. This phenomenon is observed in thin (a few nanometers) alternating layers (superlattices) of ferromagnetic and nonmagnetic metals (e.g., iron and chromium) (Figure 7.17). Depending on the width of the nonmagnetic spacer layer, there can be a ferromagnetic or antiferromagnetic interaction between the magnetic layers, and the antiferromagnetic state of the magnetic layers can be transformed into the ferromagnetic state by an external magnetic field. The spin-dependent scattering of the conduction electrons in the nonmagnetic layer is minimal, causing a small resistance of the material, when the magnetic moments
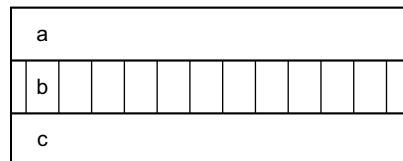
FIGURE 7.17

Diagram for describing spin-controlled electron transport in a thin film multilayer. To exploit the giant magnetoresistance effect, layers a and c are ferromagnetic and layer b is a nonmagnetic metal. In the magnetic tunnel junction layer b is a dielectric.

of the neighboring layers are aligned in parallel, whereas for the antiparallel align-ment the resistance is high. The technology is already used for read-write heads in computer hard drives. It is noteworthy that the discovery of GMR depended on the development of methods for making high-quality ultrathin films (Section 6.3.1). The GMR effect has clearly demonstrated that spin-polarized electrons can carry a mag-netic moment through nonmagnetic materials while maintaining spin coherence: this is the meaning of the term "spin transport" nowadays.

A second type of magnetic sensor is based on the magnetic tunnel junction (MTJ). In this device, a very thin *dielectric* layer separates ferromagnetic (electrode) layers, and electrons tunnel through the nonconducting barrier under the influence of an applied voltage (Figure 7.17). The conductivity depends on the relative orientation of the electrode magnetizations and the tunnel magnetoresistance (TMR): it is low for parallel alignment of electrode magnetization and high in the opposite case. The magnetic field sensitivity is even greater than for GMR. MTJ devices also have high impedance, enabling large signal outputs. In contrast with GMR devices, the elec-trodes are magnetically independent and can have different critical fields for changing the magnetic moment orientation. The first laboratory samples of MTJ structures (NiFe-Al_2O_3-Co) were demonstrated in 1995.

A further direction of development of spintronic devices is based on the develop-ment of multilayer structures of ferromagnetic semiconductors, which demonstrate properties not available for their metal analogs (e.g., the possibility to control by electric field a magnetic state of material) and the giant planar Hall effect, which exceeds by several orders of magnitude the Hall effect in metal ferromagnets. These might be exploited to create memory devices, but they have the disadvantages of poor integratability and the necessity of using additional controlling transistors.

GMR and MTJ are important enablers of size reduction of ferromagnetic memory cells and hence the miniaturization of magnetic random access memory (MRAM), the limiting factor of which is not the ability to make small ferromagnetic islands (which must nevertheless be above the superparamagnetic size boundary, see Section 2.6) but the ability to sense minute magnetic fields.

A very important development for increasing storage density is the ability to control the motion of a series of movable domain walls along magnetic nanowires by spin-polarized current pulses. This provides the basis of magnetic domain-wall "racetrack" memory.

7.6.2 Other Memory Devices

By analogy with ferromagnetic memory, ferroelectric materials are being investigated for nonvolatile storage (nanocapacitor arrays). Using a nanoporous template, ferro-electric ceramic (e.g., lead zirconate titanate) can be deposited as nanoscale islands on a suitable metal (e.g., platinum). They require ultrasensitive electrometers to be read.
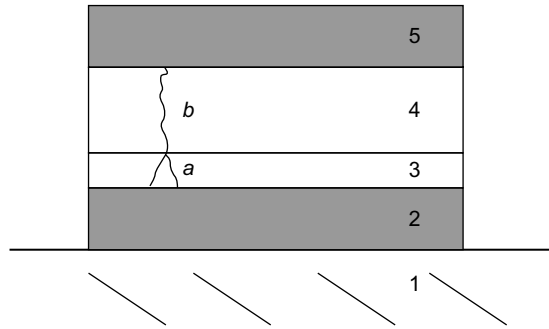
FIGURE 7.18

A resistive random access memory (RRAM) cell. Layer 1 is a silicon substrate, layers 2 and 5 are platinum electrodes, layer 3 is a dielectric oxide (e.g., SiO_2, Gd_2O_3), 2–20 nm thick; layer 4 is a solid electrolyte (ionic conductor), with a resistance about one tenth of that of layer 3, typically a metal-doped nonstoicheometric oxide (with oxygen vacancies) (e.g., MoO_{3-x}:Cu, $0 \leq x \leq 1.5$). See text for explanation of operation.

A more promising direction is resistive random access memory (RRAM). This works as follows (Figure 7.18) [83]. It is necessary to prepare the device by applying a "foaming" voltage across the electrodes 2 and 5 (the polarity is indifferent). The high electric field across layer 3, which has the highest resistance, results in localized electrical breakdown to create a conducting tip a with an area of 10 nm^2 or less. The device is still in a high resistance state because of the layer 4. Now a positive voltage is applied to layer 5. This reduces some of the metal ions in the layer 4 and creates a conducting metal filament b continuing the path initiated by the conducting tip. The device is now in a low resistance state, and remains in it until a positive voltage is applied to layer 5, which oxidizes and hence destroys the metal filament b.

Crosspoint cell arrays can be made according to this principle to create an addressable memory.

7.6.3 Spin-Dependent Transistors

This kind of spintronics device uses control of electron spin in order to gate electrons passing through a transistor. Figure 7.19 shows a proposed example. Note that it is made entirely of nonmagnetic materials.

7.6.4 Single Spin Logic

Spintronic-based devices have the potential to be very fast and operate with very low power. Spin has polarization, and is bistable in a magnetic field, (parallel or anti-parallel, which can represent the two logic states zero and one). Switching between them requires merely flipping spin without physically displacing the electrons. In
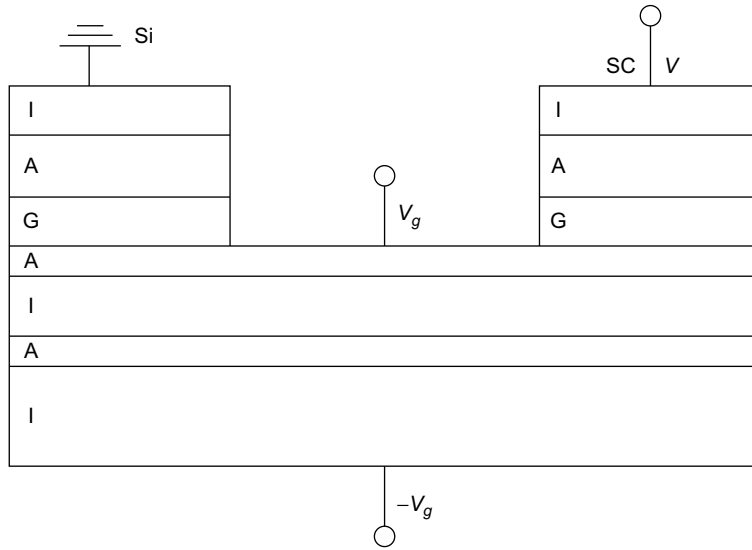
FIGURE 7.19

A spin-dependent transistor. The labels signify: I, InAs; A, AlSb; G, GaSb. The grounded terminal Si is the spin injector and the terminal SC at a potential V is the spin collector. The gate voltage V_g manipulates the spins of the two-dimensional electron gas in the layer I below it. See [67] for more details.

contrast, electronic charge-based devices have the fundamental shortcoming that charge is a scalar quantity possessing only magnitude. Hence the only way to switch between zero and one (for example, see Figure 7.3) is to change the magnitude of the charge, which invariably implies current flow I and its associated dissipation I^2R, where R is a resistance. Charged-based devices are therefore intrinsically dissipative.

Single spin logic (SSL) is the spin analogy of single electron (charge) logic (Section 7.4.3). Figure 7.20 illustrates a NOT–AND gate based on single spin logic [12]. A "true" spintronics device encodes binary information as spin, in contrast to the so-called spin transistor (Section 7.6.3), in which spin merely mediates switching.

7.7 PHOTONIC DEVICES

The "natural" length scale associated with light is its wavelength, λ: a visible photon of a typical green colour has a wavelength of about 500 nm. Macroscopic optics deals with distances $\gg \lambda$ (the far field). There might be some justification for referring to the near field, involving distances $\ll \lambda$, as nanophotonics. The word "photonics" was itself coined in analogy to "electronics", and is particularly used in connexion with
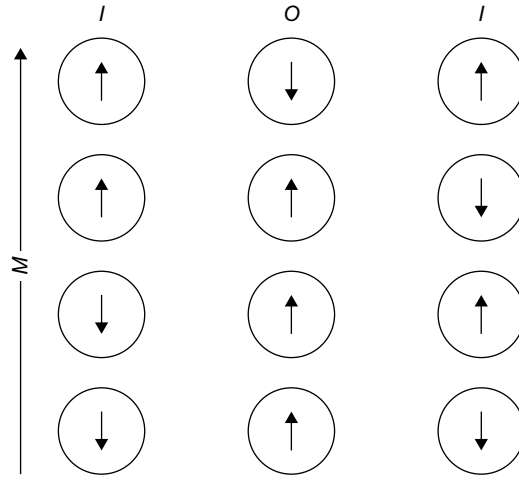
FIGURE 7.20

A spintronic NOT-AND (NAND) gate. The two inputs are the left and right columns, and the output is the central column. The physical entities are quantum dots. The upspin state parallel to the global magnetic field **M** represents 1 and the downspin state represents 0. Exchange interaction causes nearest neighbors to prefer antiparallel spins, but if there is a conflict, the global magnetic field biases the output to 1.

integrated optics, in which light is guided in waveguides with at least one characteristic dimension $< \lambda$. Light can be confined in a channel or plate made from a transparent material having a higher refractive index than that of its environment. Effectively, light propagates in such a structure by successive total internal reflexions at the boundaries. The channel (of fiber) can have a diameter, or the plate thickness, less than the wavelength of the light. Below a certain minimum diameter or thickness (the cut off), however, typically around one-third of the wavelength of the light, propagation is no longer possible. The science and technology of light guided in thin structures is called integrated optics and fiber optics, The main areas of application are in communication (transmission of information) and in sensing (Section 5.3). However, the fact that waveguiding cannot take place below a certain minimum waveguide dimension, typically of the order of 100 nm, means that there can be little justification for using nanophotonics as a synonym for integrated optics. On the other hand, the near field is particularly associated with rapidly decaying evanescent fields, and optical waveguides generate evanescent fields at their surfaces, with a penetration depth (here, into the cover medium) given by (cf. equation 5.22)

$$z_{\mathrm{C}} = (\lambda/2\pi)(N^2 - n_{\mathrm{C}}^2)^{-1/2} \tag{7.21}$$

where N is the effective refractive index of the guided mode and n_{C} is the refractive index of the cover medium. Hence, the term "nanophotonics" could reasonably

be used to describe near-field measurements and processes, especially those associated with evanescent fields generated by optical waveguides. In other words the nanoregime of photonics is the regime of the near zone and concerns processes taking place within the evanescent field. The term can also be used to describe light sources based on nano-objects (quantum wells in the form of plates, wires or particles) and nanoparticle-based photoelectrochemical solar energy converters (cf. Figure 7.24).

Semiconductor lasers, in which a voltage is applied across a semiconductor crystal (which in effect constitutes a Fabry–Perot cavity) to create a nonequilibrium population distribution of electrons and holes, whose luminescent recombination generates photons stimulating further emission, were already in existence when Dingle and Henry showed that using quantum wells as the active lasing medium would result in more efficient lasers with lower threshold currents [41], essentially because quantum confinement of the charge carriers and the optical modes enhances carrier–radiation interaction; moreover the lasing wavelength could be tuned by changing the thickness of the layers: quantum well lasers are made from alternating ultrathin layers of wider and narrower band gap semiconductors (for example, n-AlGaAs and GaAs, respectively). Improvements in the technology of ultrathin film fabrication, especially with the introduction of techniques such as molecular beam epitaxy (MBE) (Section 6.3.1), enabled quantum well lasers to become a reality.

Reduction of the dimensionality from two to one (quantum wires) and to zero should lead to further improvements (even higher gain and lower threshold current than quantum well lasers). The carriers are confined in a very small volume and population inversion occurs more easily, leading to lower threshold currents for lasing. Furthermore, the emission wavelength can be readily tuned by simply varying the dimensions of the dot (or well). The main difficulty is to ensure that the dots comprising a device are uniformly sized. If not, the density of states is smeared out and the behavior reverts to bulk-like, negating the advantage of the zero-dimensional confinement. Early attempts to produce quantum dots a few tens of nanometers in diameter using electron beam lithography followed by the usual semiconductor processing (etching, see Section 8.1.1) were bedevilled by damage and contamination introduced by the processing. An important advance came through the exploitation of frustrated wetting (Stranski–Krastanov growth): lattice mismatch between the deposited layer and the substratum results in strain, which was found to be relieved by the spontaneous formation of monodisperse islands (Section 8.1.2).

Microcavities have the potential for fabricating ultrahigh-Q and ultralow threshold lasers. Cylindrical or spherical microcavities are particularly interesting in that they have extremely low loss "whispering gallery" modes (WGM), which occur when light circumnavigating the cavity is trapped by total internal reflexions and constructively interferes with itself. To construct a laser based on a microcavity it is surrounded by gain medium and pumped by an external light source. The lower limit of miniaturization is exemplified by the laser oscillation demonstrated with a single atom in an optical resonator. Threshold pump energies of the order of 100 μJ have been demonstrated. Quality factors in excess of 10^{10} can be routinely obtained.

7.8 MECHANICAL DEVICES

The spring constant (stiffness) k of a nanocantilever varies with its characteristic linear dimension l, and its mass m as l^3. Hence the resonant frequency of its vibration

$$\omega_0 = \sqrt{k/m} \qquad (7.22)$$

varies as $1/l$. This ensures a fast response—in effect, nanomechanical devices are extremely stiff. Since the figure of merit (quality factor) Q equals ω_0 divided by the drag (friction) coefficient, especially for devices operating in a high vacuum Q can be many orders of magnitude greater than the values encountered in conventional devices. On the other hand, under typical terrestrial operating conditions water vapor and other impurities may condense onto moving parts, increasing drag due to capillary effects (cf. Figure 3.1), and generally degrading performance.

The simplest mechanical device is probably the vibrating cantilever; we may use this to fix ideas (Figure 7.21; see Figure 7.22 for a more picturesque representation). From equation (7.22), the fundamental (resonant) frequency is the square root of the quotient of the spring constant and the mass, i.e.,

$$\omega_0 = \sqrt{Yl/m_{\text{eff}}} \qquad (7.23)$$

where l is an appropriate length of the device, and we specify an *effective mass* that depends on the geometry and anisotropy of the material of the cantilever as well as on its mass density. The very fact that l/m_{eff} scales as l^{-2} suggests that ω_0 becomes very high as l falls to nanometer size. The quality factor seems to scale with device
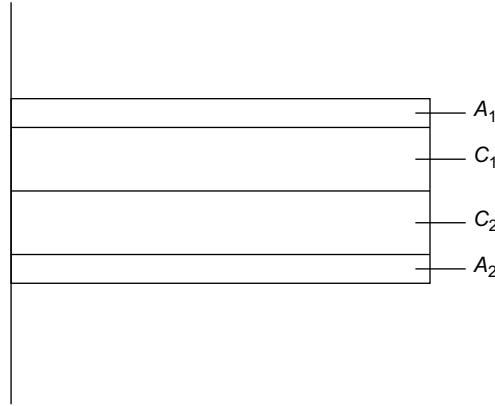
FIGURE 7.21

Schematic diagram of a cantilever. The core structure is represented by layers C and the surface layers are labeled A; the latter may be modified by interaction with the environment. Thus, accrual of mass on C may create A, or very porous pre-existing structure may be stiffened. Any asymmetry between C_1 and C_2 (e.g., if only C_1 can accumulate additional material) may lead to static bending of the cantilever.
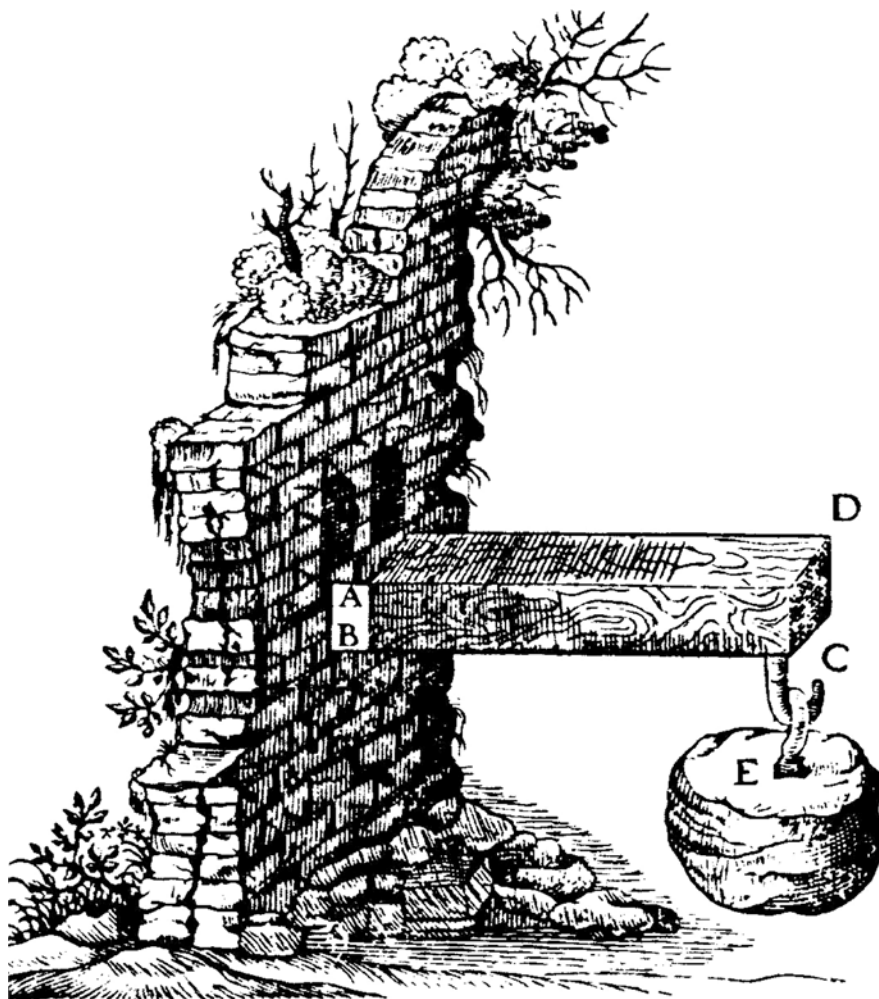
FIGURE 7.22

A cantilever (from G. Galilei, Discorsi e Dimostrazioni Matematiche. Leyden, 1638).

volume $\sim V^{1/5}$, hence the power that must be applied to drive the cantilever to an amplitude equal to the thermal fluctuations becomes extraordinarily low, $\sim k_B T/(Q/\omega_0)$. Nevertheless, individual mechanical quanta ($\hbar\omega_0$) have not yet been observed, although it is now feasible for them to be greater than the thermal energy $k_B T$.

Envisaged applications of nanoscale cantilevers are as logic devices; in effect ultraminiature versions of the interlocking between points and signals developed for

railway systems to prevent a signalman from selecting operationally incompatible combinations; and as sensors.

The advent of ultrastiff carbon-based materials (Chapter 9) has generated renewed interest in mechanical devices that can now be made at the nanoscale (nanoelectromechanical systems, NEMS). Ultrasmall cantilevers (taking the cantilever as the prototypical mechanical device) have extremely high resonant frequencies, effective stiffnesses and figures of merit Q and, evidently, very fast response times $\sim Q/\omega_0$. It therefore becomes conceivable that a new generation of relays, constructed at the nanoscale, could again contend with their solid-state (transistor-based) rivals that have completely displaced them at the microscale and above. Relays have, of course, excellent isolation between input and output, which makes them very attractive as components of logic gates.

Sensing applications can operate in either static or dynamic mode. For cantilevers (Figure 7.21) used as sensors, as the beam becomes very thin, its surface characteristics begin to dominate its properties (cf. Section 2.2). Classically the adsorption of particles onto such a beam would increase its mass, hence lower its resonant frequency. This effect would be countered if the adsorbed adlayer had a higher stiffness than the beam material, thereby *increasing* its resonant frequency. In static sensing mode the adsorption of particles on one face of a beam breaks its symmetry and causes it to bend.

Manufacturing variability may be problematical for ultrasensitive mass sensors; ideally they should be fabricated with atomic precision using a bottom-to-bottom approach (Section 8.3); in practice nowadays they are fabricated using semiconductor processing technology.

7.9 FLUIDIC DEVICES

In gases, the dimensionless Knudsen number associated with the motion of the molecules is the ratio of the mean free path ℓ to some characteristic size l:

$$K_n = \ell/l \qquad (7.24)$$

The mean free path depends on the pressure and the simplest possible derivation assuming the gas molecules to be elastic objects with an effective diameter σ yields

$$\ell = 1/\pi\sigma^2 n \qquad (7.25)$$

where n is the number per unit volume. Following similar reasoning to that of Section 7.4.1, one could take the nanoscale as the scale corresponding to $K_n = 1$; it would therefore be pressure-dependent. At atmospheric pressure, the mean free path of a typical gas is several tens of nm. Given that even a moderate vacuum of say 10^{-3} mbar gives a mean free path of about 0.1 m, a low pressure nanodevice would seemingly have no unique feature.

In fluids, we typically have molecules or particles of interest dissolved or suspended in the solvent. The particles (of radius r) undergo Brownian motion,

i.e., thermal collisions, and the mean free path is effectively the distance between successive collisions with the solvent molecules (formally, it is the average distance from any origin that a particle travels before significantly changing direction) and is related to the diffusion coefficient D and the thermal velocity c of the particle according to

$$\ell = 3D/c \tag{7.26}$$

the diffusivity depends on the friction coefficient according to the Stokes–Einstein law:

$$D = k_B T/(6\pi\eta r) \tag{7.27}$$

where η is the dynamic viscosity of the solvent, and the thermal velocity depends on both the particle size and density ρ:

$$c = \sqrt{6k_B T/(\pi^2 \rho r^3)} \tag{7.28}$$

Substituting these expressions into equation (7.26) gives

$$\ell = \frac{\sqrt{(3/2)k_B T \rho r}}{6\eta} \tag{7.29}$$

from which it follows that under any practical conditions (e.g., water at $20\,^\circ\mathrm{C}$ has a density of 1 g/cm^3 and a viscosity of 10^{-3} Pa s) the particle size exceeds the mean free path, hence there can be no ballistic transport and hence no unique feature at the nanoscale. In other words, although in analogy to the case of electron conduction (Section 7.4.1) a nanofluidic device could be one for which the mean free path of particles moving within it is greater than the characteristic size of the internal channels, this condition could never be realistically met. If some way could be found of working with nanoparticles of material from a white dwarf star (mass density $\sim 10^9$ g/cm^3), this conclusion would no longer be valid!

For devices that depend on thermal conductivity, it is limited by phonon scattering, but analogously to electron conduction (Section 7.4.1), if the device is small enough ballistic thermal transport can be observed, with the quantum of thermal conductance being $\pi^2 k_B^2 T/(3h)$.

Another approach to defining a nanofluidics scale could be based on the fact that fluidics devices are fundamentally mixers. For reactions in minute subcellular compartments (or taking place between highly diluted reagents) $\langle a\rangle\langle b\rangle \neq \langle ab\rangle$ (see Section 2.4 for the notation). The nanoscale in fluidics could be reasonably taken to be that size in which $\Delta^2(\gamma_t)$ differs significantly from zero. The scale depends on the concentration of the active molecules (A and B).

This calculation ignores any effect of turbulence on the mixing, for which the characteristic scale would be the Kolmogorov length ℓ_K:

$$\ell_K = (\kappa^3/\varepsilon)^{1/4} \tag{7.30}$$

where κ is the kinematic viscosity and ε is the average rate of energy dissipation per unit mass. But, this length is of micrometer rather than nanometer size. Furthermore, it is essentially impossible to achieve turbulent flow in nanochannels; even microfluidics is characterized by extremely low Reynolds numbers (below unity).

Fluids are often moved in miniature reactors using electroösmosis, in which an electric field is applied parallel to the desired rate of flow. If the walls of the device are ionizable; for example, most metal oxides are hydroxylated at the interface with water, and hence can be protonated or deprotonated according to

$$|- OH_2^+ + OH^- \rightleftharpoons |- OH + H_2O \rightleftharpoons |- O^- + H_3O^+ \qquad (7.31)$$
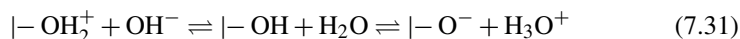
hence, the interface will be charged and the counterions will be dispersed in a diffuse (Gouy–Chapman) double layer, their concentration decaying with a characteristic Debye length $1/\kappa$, given by (for monovalent ions)

$$\kappa^2 = e^2 \sum_i n_i z_i^2 / (\epsilon k_B T) \qquad (7.32)$$

where n_i is the concentration of the ith ion bearing z_i elementary charges. The decay length is typically of the order of nanometers or tens of nanometers; upon application of the electric field the counterions will move, entraining solvent and engendering its flow. If the radius of the channel is less than $1/\kappa$, electroösmosis will be less efficient. This length could therefore be taken as determining the nanoscale in fluidics devices (provided some electrokinetic phenomenon is present), because below this scale bulk behavior is not followed. Electroösmosis is not of course useful in a macroscopic channel, because only a minute fraction of the total volume of solvent will be shifted. Note the analogy between the semiconductor space charge (cf. equation 7.3) and the Gouy–Chapman ion distribution.

The capillary length,

$$l_{\text{cap}} = (k_B T/\gamma)^{1/2} \qquad (7.33)$$

(confusingly, an alternative definition is current in macroscopic fluid systems, in which surface tension is compared with sedimentation)—the surface tension γ is that between the solid wall and the fluid—gives an indication of the predominance of surface effects (cf. Section 2.2). It is typically of the order of nanometers. There is no discontinuity in behavior as the characteristic size of a fluidic device falls to or below l_{cap}.

In summary, as the radius of a fluid channel becomes less than a characteristic length such as $1/\kappa$ or l_{cap}, in effect the properties of the channel wall influence the entire volume of the fluid within the channel. The implications of this are far from having been thoroughly worked out. Especially in the case of aqueous systems, the structure of the water itself (especially the average strength of its hydrogen bonds) is also affected by wall interactions, doubtless with further far-reaching implications. But most discontinuities due to miniaturization of fluidic devices seem to already manifest themselves at the microscale.

7.9.1 **Mixers and Reactors**

Miniaturizing mixers has been very successful at the microscale, as can be deduced from the huge proliferation of lab-on-a-chip devices for analytical and preparative work in chemistry and biochemistry. Particular advantages are the superior control over flow compared with macroscopic mixers (cf. Figure 2.4), one very important benefit of which is much more predictable selection of reaction products, wherever several are possible (in consequence, yields can be raised to 100%), and (in principle) great ease of scale up, simply by having many micromixers in parallel (although this does not yet seem to have been convincingly demonstrated for any industrial-scale production facility). It is, however, by no means clear that even greater success will attend further miniaturization down to the nanoscale. On the contrary, performance may be degraded. This needs further investigation.

7.9.2 **Chemical and Biochemical Sensors**

These devices are included under fluidics since the sample to be analyzed is almost invariably a fluid (cf. Section 11.5). The Holy Grail of clinical sensing is continuous, noninvasive monitoring (cf. Section 4.2). Currently, most tests require a sample of the relevant biofluid (e.g., blood) to be drawn from the patient. For most people this is a somewhat unpleasant procedure, hence the tests are carried out infrequently. It is, however, recognized that much more insight into a patient's pathological state could be obtained by frequent, ideally continuous, monitoring. At present, this is only possible in intensive care stations, where the patient is immobilized, and even there continuous invasive monitoring does not take place (the oxygen content of the blood is monitored noninvasively by analyzing the optical reflectance spectrum of the skin covering blood vessels). It seems to be a very difficult technical problem to extend such noninvasive analysis to the plethora of biomarkers currently under intensive study as symptomatic of disease or incipient disease. An alternative approach is to develop sensors so tiny that they can be semipermanently implanted inside the body, where they can continuously monitor their surroundings.

Because of the large and growing number of afflicted people, diabetes has received overwhelmingly the most attention. The sensing requirement is for glucose in the blood. The glucose sensor follows classic biosensing design: a *recognition element* to capture the analyte (glucose) mounted on a *transducer* that converts the presence of captured analyte into an electrical signal (Figure 7.23). The recognition element is typically a biological molecule, the enzyme glucose oxidase, hence (if small enough) this device can be categorized as both nanobiotechnology and bionanotechnology.

Both components of the biosensor are excellent candidates for the application of nanotechnology. Molecular recognition depends on a certain geometrical and chemical arrangement of atoms in some sense complementary to the analyte molecules,
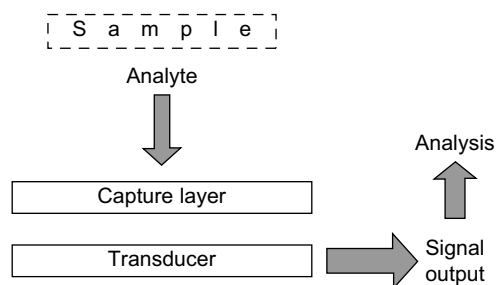
FIGURE 7.23

A prototypical biosensor. The capture layer concentrates the analyte in the vicinity of the transducer, which reports the concentration of analyte in the capture layer, which is directly related to the concentration in the sample.

together with cooperative motions to enhance affinity. Atom-by-atom assembly therefore represents the perfect way to artificially fabricate recognition elements. The ultimate goal of the transducer is to detect a single captured analyte molecule, hence the smaller it can be made the better, provided issues related to noise and detection efficiency (Section 10.8) can be overcome. The goal is to exploit the phenomenon that a minute change in an atomically scaled device (e.g., the displacement of a single atom) can significantly change its properties, including functionality.

7.9.3 Energy Conversion Devices

Originally inspired by the photosynthetic machinery in blue-green algae and plants, in order to overcome the problem that the output of a photovoltaic or thermal solar energy converter (i.e., electricity or heat, respectively) must be subsequently stored, possibly via further conversion and inevitably involving losses, attempts have been made to create devices that convert the solar energy directly into chemical form. The basic principle is that the absorption of light by a semiconductor creates highly reducing and highly oxidizing species (respectively, the conduction band (CB) electrons and the valence band (VB) holes). Instead of separating them in a space charge as in the conventional p–n junction device and collecting the current, if the semiconductor is a nano-object with its characteristic dimension comparable to the Bohr radii of the charge carriers (Section 2.5), they may react with chemical species adsorbed on the object's surface before they have time to recombine (Figure 7.24).

Challenges that still need to be overcome include: matching the semiconductor band gap to the solar spectrum; matching the positions of the band edges to the redox potentials of the desired chemical reactions; finding catalysts to facilitate the desired chemical reactions; preventing photocorrosion (i.e., reduction or oxidation of the semiconductor by its own charge carriers).
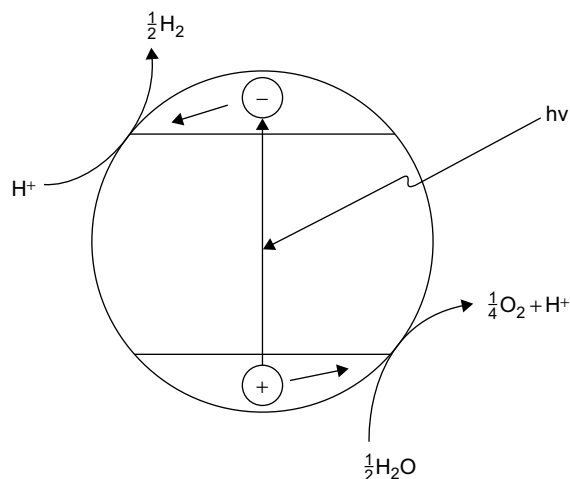
FIGURE 7.24

Radiant energy harvesting using semiconductor nanoparticles. The absorption of a photon raises an electron from the valence band (VB) to the conduction band (CB), leaving a positive hole in the former. The CB electron ⊖ is a strong reducing agent and the VB hole ⊕ a strong oxidizing agent, possibly capable of, respectively, reducing and oxidizing water to hydrogen and oxygen. At present, this concept of nanoparticle photoelectrochemistry is used to sacrificially destroy persistent organic pollutants allowed to adsorb on the particles. This is also the basis of self-cleaning windows and compact water purification devices.

7.10 SUMMARY

A device is fundamentally an information processor or transducer. The types of devices considered in this chapter are mainly logic gates and sensors. The relay is considered as the paradigmatical component of the logic gate, which is the foundation of digital processing hardware. Ultraminiature devices based on electron charge as the carrier of information are considered. These include single-electron transistors, which are scaled-down versions of conventional transistors; molecular electronic devices, in which organic molecules replace the inorganic semiconducting media of conventional transistors; and quantum dot cellular automata. As these devices become very small, movement of electronic charge becomes problematical and using electron spin as the information carrier becomes attractive. This includes both hybrid devices in which spin controls electron current (the main application is for magnetic field sensing), and "all-spin" devices in which spin is also the carrier of information. Photons may also be used as information carriers, although mostly not in the nano-realm, except for molecular photonics, which is, however, covered in Chapter 11 because the only current example uses a biomolecule for the central function. Quantum computation, although not specifically tied to the nanoscale, is briefly described

as a completely different approach to increasing computational density, for which photonic devices are well suited. Nanoscale mechanical devices are considered, both as relays and sensors. A brief survey of fluidic devices in the nanoscale is given, including biosensors. Finally, nanoparticulate systems for the conversion of solar energy into fuels are considered.

7.11 FURTHER READING

L.O. Chua, Memristive devices and systems, Proc. IEEE 64 (1976) 209–223.

D.K. Ferry, L.A. Akers, Scaling theory of modern VLSI, IEEE Circuits Devices Mag. (1997) 41–44.

J.A. Hutchby, V. Zhirnov, R. Cavin, G. Bourianoff, Emerging nanoscale memory and logic devices: a critical assessment, IEEE Comput. Mag. (2008) 78–82.

L. Jacak, Semiconductor quantum dots—towards a new generation of semiconductor devices, Eur. J. Phys. 21 (2000) 487–497.

K.K. Likharev, Single-electron devices and their applications, Proc. IEEE 87 (1999) 606–632.

S.S.P. Parkin, M. Hayashi, L. Thomas, Magnetic domain-wall racetrack memory, Science 320 (2008) 190–194.

A. Politi, J.L. O'Brien, Quantum computation with photons, Nanotechnol. Percept. 4 (2008) 289–294.

T. Shibata, Computing based on the physics of nanodevices—a beyond-CMOS approach to human-like intelligent systems, Solid State Electron. 53 (2009) 1227–1241.

This page intentionally left blank

Nanofacture of Devices

CHAPTER CONTENTS

Whereas Chapter 6 considered the mainly current methods of producing nano-materials with a regular structure (and often with a merely statistical order), in this chapter we look at more sophisticated technologies, both present and supposed future, capable in principle of fabricating functional artifacts of arbitrary complexity. The entire field is encompassed within the three divisions of top–down, bottom–bottom, and bottom–up; Figure 8.1 gives a summary. Top–down requires great ingenuity
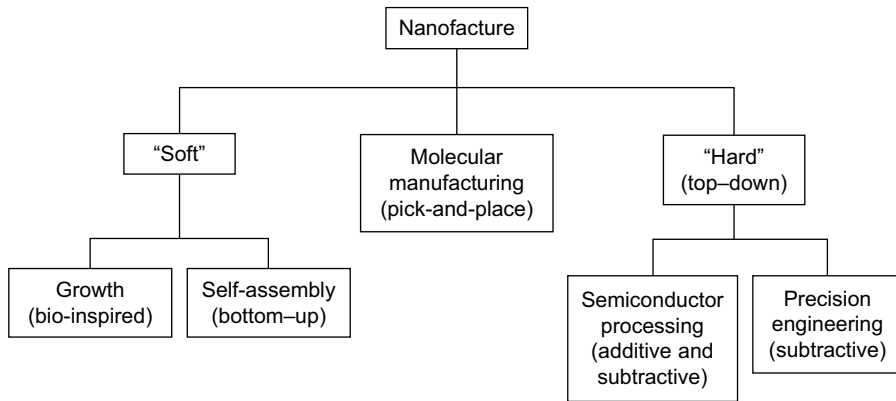
FIGURE 8.1

Different modes of nanomanufacture (nanofacture).

(and expense) in the fabrication machinery; bottom–up requires great ingenuity in the conception of the building blocks. Bottom–bottom, the least developed of the three, requires above all ingenuity in conception.

8.1 TOP–DOWN METHODS

These share the general feature of requiring large (and also expensive, see Section 1.3.2, requiring considerable concentrations of capital) installations. What might be called the traditional route, that of scaling down processes familiar in macro and micro engineering, appears on the extreme right of the diagram, Figure 8.1. Concerted incremental improvement in the entire manufacturing process transforms precision engineering into ultraprecision engineering (Figure 1.2). The stiffness of the parts of mechanical devices used to shape objects is particularly important for achieving precision. These processes are essentially subtractive: material is removed by grinding or etching.

8.1.1 Semiconductor Processing

This well-established industrial technology refers to the operations of sequentially modifying (e.g., oxidizing), depositing (additional layers on), and removing (parts of) a substratum (e.g., silicon) over areas selected by exposing photoresist coating the working surface through a mask and then dissolving away the unexposed resist (or the converse). Remaining resist either protects from etching or from fresh deposition. This works well at the micrometer scale, and is used to fabricate very large-scale integrated circuits. Problems of scaling it down to produce features with lateral sizes in the nanorange runs into the diffraction limit of the light used to create the mask (cf. equation 5.2), partly solved by using light of shorter wavelengths or high energy electrons.

In contrast to the difficulties of nanoscaling lateral features, very high-quality thin films can be deposited with nanometer control perpendicular to the plane of a substratum. These methods are grouped under the heading of physical vapor deposition (PVD). The material to be deposited is evaporated from a reservoir or magnetron sputtered from a target. The most precise control is obtainable with molecular beam epitaxy (MBE), developed at AT&T Bell Laboratories in the late 1960s: evaporated material is beamed onto the substratum under conditions of ultrahigh vacuum. Deposition is typically very slow (several seconds to achieve 1 nm film thickness) and hence can be epitaxial. Ultrathin layers (of the order of a nanometer) with atomically sharp interfaces can be deposited (Section 6.3.1).

Chemical vapor deposition (CVD) is similar to PVD, except that the precursor of the thin layer is a reactive gas or mixture of gases, and the substratum is typically heated to accelerate chemical reaction to form a solid product deposited as a film. The decomposition can be enhanced with a plasma (this typically allows the substratum to be maintained at a lower temperature than otherwise).

Related technologies are used to modify existing surfaces of materials, such as exposure to a plasma, and ion implantation, in which electrostatically charged high-energy (typically 10–100 keV) ions are directed towards the surface, where they arrive with kinetic energies several orders of magnitude higher than the binding energy of the host material, and become implanted in a surface layer that may be tens of nanometers thick.

Physical and chemical vapor deposition processes are often able to yield structure at the nanoscale, or the lower end of the microscale, with some structural control achievable via the deposition parameters. The structures obtained emerge from a particular combination of deposition parameters and must be established experimentally. Theoretical understanding of structure and process is still very rudimentary, although attempts are under way (such as the "structure zone model").

8.1.2 Mismatched Epitaxy

In the epitaxial growth of semiconductors (see Section 8.1.1), when material B is evaporated onto a substrate of a different material A, three analogous situations can arise:

- Frank–van der Merwe (B wets A, and a uniform layer is formed).
- Volmer–Weber (no wetting, hence islets distributed in size of B on A are formed).
- Stranski–Krastanov (B can wet A, but there is a slight lattice mismatch between A and B, and the accumulating strain energy is ultimately sufficient to cause spontaneous dewetting, resulting in rather uniform islets of B, which relieves the strain—this explanation is plausible but there is still much discussion regarding the mechanism).

One can therefore apply equation (3.22). If $S > 0$ we have the Frank–van der Merwe regime; the substratum is wet and layer-by-layer growth takes place. If $S < 0$ we have the Volmer–Weber regime; there is no wetting and three-dimensional islands grow. Yet if $S > 0$ for the first layer at least, but there is a geometric mismatch between the lattices of the substratum and the deposited layer, strain builds up in the latter,

which is subsequently relieved by the spontaneous formation of monodisperse islands (the Stranski–Krastanov regime); it can be thought of as frustrated wetting. Its main application is for fabricating quantum dots for lasers (quantum dot lasers are a development of quantum well lasers; the carriers are confined in a small volume and population inversion occurs more easily than in large volumes, leading to lower threshold currents for lasing, and the emission wavelength can be readily tuned by simply varying the dimensions of the dot (or well), see Section 2.5). The main difficulty is to ensure that the dots comprising a device are uniformly sized. If not, the density of states is smeared out and the behavior reverts to bulk-like. Initially, the quantum dots were prepared by conventional semiconductor processing, but it was found to be very difficult to eliminate defects and impurities, whereas the Stranski–Krastanov self-assembly process does not introduce these problems.

8.1.3 Electrostatic Spray Deposition (ESD)

Ceramic precursors are dissolved in a suitable solvent and mixed immediately prior to forcing through an orifice maintained at a high potential difference with respect to the substratum. The liquid breaks up into electrostatically charged droplets, which are attracted both by gravity and the Coulombic force to the substratum, which is typically heated to accelerate the reaction that forms the final material. For example, calcium nitrate and phosphoric acid dissolved in butyl carbitol and atomized upon leaving the nozzle at a potential of 6–7 kV with respect to a titanium or silicon substratum maintained at a few hundred °C about 30 mm below the nozzle create coatings of calcium phosphate with intricate and intriguing nanostructures [104].

8.1.4 Felting

The fabrication of texture by felting has been known for centuries (in Europe, and much longer in China) in the guise of papermaking, reputed to have been invented in AD 105 by Ts'ai Lun in China. It arrived in Europe, firstly in Italy, by the end of the 13th century, probably via Samarkand, Baghdad, Damascus, Egypt, the Maghreb and Muslim Spain; it appeared in England around 1490. Vegetable fibers (typically based on cellulose) are macerated until each individual filament is a separate unit, mixed with water, and lifted from it in the form of a thin layer by the use of a sieve-like screen, the water draining through its small openings to leave a sheet of felted (matted) fiber upon the screen's surface. The first papermaking machine was invented by Robert in France at around the end of the 18th century, and was later perfected by two Englishmen, the Fourdrinier brothers. The machine poured the fibers out in a stream of water onto a long wire screen looped over rollers; as the screen moved slowly over the rollers, the water drained off and delivered an endless sheet of wet paper.

With the advent of various kinds of nanofibers, it is now possible to make paper-like materials at the nanoscale. This has been attempted, most notably using carbon nanotubes, when it is sometimes called buckypaper. The process is different from the

FIGURE 8.2

Sketch of a two-dimensional random fiber network (RFN) used to model a felted assembly.

shaking and/or stirring used to prepare a mixed powder. Randomly oriented fibers (highly elongated objects) are rapidly placed on top of each other, with their long axes parallel to a substratum (which is, in the case of cellulose-based writing paper, removed later on) to form a random fiber network (RFN) (Figure 8.2). The network coheres because of numerous fiber–fiber contacts, but the structure is different from that formed by the entanglement of very long randomly coiled polymers. Mixtures of fibers would be especially interesting for creating nanotexture.

Simplified models, such as the random fiber network, into which the only input is the actual distribution of fiber lengths and their surface chemical properties, are useful for calculating basic properties of felted materials, for example mass distribution, number of crossings per fiber, fractional contact area, free-fiber length distribution and void structure.

An example of a naturally felted structure is the basement membrane assembled from extracellular matrix proteins such as laminin, used in multicellular organisms to support cells and tissues. The deposition of fibrous proteins such as laminin in the presence of divalent cations such as calcium allows sheets of arbitrary thickness to be assembled. Somewhat different is the giant glycoprotein mucin constituting the mucous films lining the epithelial surfaces of multicellular organisms: the molecules are large and flexible enough to be entangled with one another.

8.1.5 Ultraprecision Engineering

Current ultrahigh-precision ("ultraprecision") engineering is able to achieve surface finishes with a roughness of a few nanometers. According to McKeown et al. it is achievable using 11 principles and techniques [113], which include:

- dynamic stiffness; thermal stability; seismic isolation
- rigid-body kinematic design; three-point support
- measuring system isolated from force paths and machine distortion
- high accuracy bearings
- quasistatic and dynamic error compensation.

Although some of the attributes implied by these principles and techniques would appear to call for large, massive machines, the need for a further, twelfth, principle has become apparent, namely that of miniaturization—ultraprecision machines should be as small as is feasible taking account of the 11 principles. This is because smaller load paths and metrology and thermal loops directly improve dynamic performance and accuracy. There should be, therefore, ultimately convergence between assemblers and ultraprecision machine tools (UPMT).

8.2 BOTTOM–UP METHODS

The increasing difficulty of continuing the miniaturization of classical photolithography and its derivatives in semiconductor processing (Section 8.1.1), and the extreme laboriousness of current mechanosynthesis (Section 8.3), coupled with the observation of self-assembly in nature, generated interest in alternative fabrication technologies. The primitive idea of self-assembly ("shake and bake") is to gather precursors in random positions and orientations and supply energy ("shaking") to allow them to sample configuration space. The hugeness of this space suggests that a convergent pathway is inherent in the process in order to allow it to be completed in a reasonable time, as in protein "folding" (Section 8.2.11). Once the precursors are in position, "baking" is applied to strengthen the bonds connecting them and fix the final object permanently.

8.2.1 Self-Assembly

Assembly means gathering things together and arranging them (fitting them together) to produce an organized structure. A child assembling a mechanism from "Meccano" captures the meaning, as does an assembly line whose workers and machines progressively build a complicated product such as a motor car from simpler components; an editor assembles a newspaper from copy provided by journalists (some of which might be rejected). In all these examples the component parts are subject to constraints—otherwise they would not fit together—and their entropy S must necessarily decrease. Since the free energy $\Delta G = \Delta H - T\Delta S$ must then necessarily increase, in general the process will not, by itself, happen spontaneously. On the contrary, a segregated arrangement will tend to become homogeneous. Hence in order for self-assembly to become a reality, something else must be included. Typically enthalpy H is lost through the formation of connexions (bonds) between the parts, and provided $|\Delta H|$ exceeds $|T\Delta S|$ we have at least the possibility that the process can occur spontaneously, which is presumably what is meant by self-assembly.

The final result is generally considered to be in some sort of equilibrium. Since entropy S is always multiplied by the temperature T in order to compute the free energy, it should be noted that the relevant temperature is not necessarily what one measures with a thermometer, but one should only take the (typically small number of)
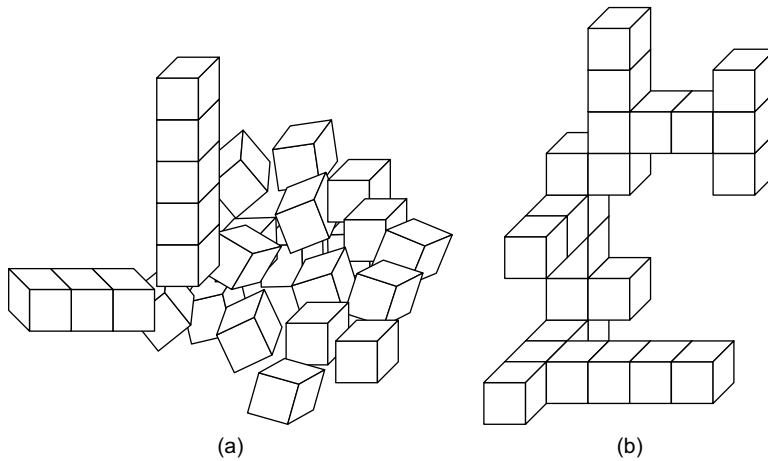
FIGURE 8.3

(a) The result of mixing isotropically sticky cubelets. (b) Putative result of mixing selectively sticky cubelets.

relevant degrees of freedom into account; the temperature is the mean energy per degree of freedom. If slightly sticky granules are thrown into a rotating drum, or stirred in a mixing bowl, typically they will all eventually clump together to form a "random" structure (e.g., Figure 8.3(a)), but one which is evidently less random than the initial collection of freely flowing granules, hence the entropy is lower, but every time this experiment is repeated the result will be different in detail, and one feels that the system is really ergodic if one had enough time to make all the repetitions, and the ergodicity is only broken because one has insufficient time. To make the example more definite, think of the dry ingredients of concrete (sand and cement) thrown into a "cement mixer" to which a small amount of water is then added.

The goal of a self-assembly process suitable for manufacturing devices is for the same structure to be formed each time the constituent particles are mixed together (Figure 8.3(b))—as was imagined by von Foerster [57]. Of course, energy needs to be put into the system. In the "purest" form of self-assembly, the energy is thermal (random), but it could also be provided by an external field (e.g., electric or magnetic). If we are satisfied by the constituent particles being merely joined together in a statistically uniform fashion and, moreover, the process happens spontaneously, then it is more appropriate to speak of self-joining or self-connecting. The mixing of concrete referred to above is, at least at first sight, an example of such a self-connecting process; more generally, one can refer to gelation. Approaching the phenomenon from a practical viewpoint, it is clear that gelation is almost always triggered by a change of external conditions imposed upon the system, such as a change of temperature or dehydration and the spontaneity implied by the prefix "self-" is absent. The only action we can impose upon the system without violating the meaning of "self-" is that of bringing the constituent particles together. Note that here we diverge from

the everyday meaning of the term "assembly", which includes the gathering together of people, either spontaneously as when a group of people wishes to protest against some new measure introduced by an authoritarian government, or by decree.

If the process in which we are interested is deemed to begin at the instant the constituent particles are brought together, then we can indeed put the mixing of concrete in the category of self-joining, because we could (although it is not usually done in that way) bring the wetted particles of sand and cement together, whereupon they would spontaneously join together to form a mass.

The meaning of self-joining (of which self-connecting is a synonym) is then the property possessed by certain particles of spontaneously linking together with their neighbors when they are brought to within a certain separation. One can also imagine there being kinetic barriers to joining, which can be overcome given enough time. Note that the particles each need more than one valency (unit of combining capacity), otherwise dimers would be formed and the process would then stop. A good example is steam condensed to form water. We can suppose that the steam is first supercooled, which brings the constituent particles (H_2O molecules) together; the transition to liquid water is actually a first order phase transition that requires an initial nucleus of water to be formed spontaneously. "Gelation" (cf. Section 3.7) then occurs by the formation of weak hydrogen bonds (a maximum of four per molecule) throughout the system.

Strictly speaking, it is not necessary for all the constituent particles to be brought together instantaneously, as implied in the above. Once the particles are primed to be able to connect themselves to their neighbors, they can be brought together one by one. This is the model of diffusion-limited aggregation (DLA). In nature, this is how biopolymers are formed: monomers (e.g., nucleic acids or amino acids) are joined sequentially by strong covalent bonds to form a gradually elongating linear chain. The actual self-assembly into a compact three-dimensional structure involves additional weak hydrogen bonds between neighbors that may be distant according to their positions along the linear chain (see Section 8.2.11); some of the weak bonds formed early are broken before the final structure is reached.

In the chemical literature, self-assembly is often used as a synonym of self-organization. A recapitulation of the examples we have already discussed shows, however, that the two terms cannot really be considered to be synonymous. The diffusion-limited aggregate is undoubtedly assembled, but can scarcely be considered to be organized, not least because every repetition of the experiment will lead to a result that is different in detail, and only the same when considered statistically. "Organized" is an antonym of "random"; therefore, the entropy of a random arrangement is high; the entropy of an unorganized arrangement is low. It follows that inverse entropy may be taken as a measure of the degree of organization; this notion will be further refined in the next section. The diffusion-limited aggregate differs from the heap of sand only insofar as the constituent particles are connected to each other. An example of organization is shown in Figure 8.3(b). The impossibility of self-organization has been proved by Ashby, as will be described in Section 8.2.3.

Before discussing self-organization, we must first discuss organization, of which self-organization is a part. If the elements in a collection (here we shall not say "system", because that already implies a degree of organization) are organized to some degree, that implies that they are in some way connected to each other, which can be considered as a kind of communication, and are hence subject to certain constraints. In other words, if the state of element B is dependent on the state of A to some extent, then we can say that B's state is conditional on that of A. Likewise, the relation between A and B may be conditional on the state of C. Whenever there is conditionality, there is constraint: B is not as free to adopt states as it would be in a totally unorganized system [10].

8.2.2 Thermodynamics of Self-Organization

Consider a universe U comprising a system S and its environment E; i.e., $U = S \cup E$ (Figure 8.4). Self-organization (of S) implies that its entropy spontaneously diminishes; that is,

$$\delta S_S / \delta t < 0 \tag{8.1}$$

Accepting the second law of thermodynamics, such a spontaneous change can only occur if, concomitantly,

$$\delta S_E / \delta t > 0 \tag{8.2}$$

with some kind of coupling to ensure that the overall change of entropy is greater than or equal to zero. If all processes were reversible, the two changes could exactly balance each other, but since (inevitably, we may suppose) some of the processes involved are irreversible, overall

$$\delta S_U / \delta t > 0 \tag{8.3}$$

Therefore, although the system itself has become more organized, overall it has generated more disorganization than the organization created, and it is more accurate to call it a self-disorganizing system [57]. Hence, the "system" must properly
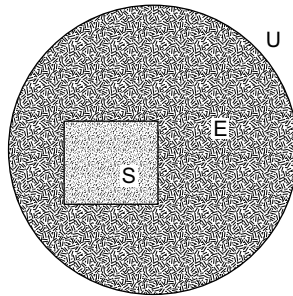
FIGURE 8.4

Universe U comprising system S and its environment E.

be expanded to include its environment—it is evidently intimately connected with it; without it there could be no organization. Despite its true nature as a self-disorganizing system having been revealed, nevertheless we can still speak of a self-organizing *part* S of the overall system that consumes order (and presumably energy) from its environment. It follows that this environment must necessarily have structure itself, otherwise there would be nothing to be usefully assimilated by the self-organizing part.

The link between entropy (i.e., its inverse) and organization can be made explicit with the help of relative entropy R (called redundancy by Shannon), defined by

$$R = 1 - S/S_{\max} \tag{8.4}$$

where $S_{\max}$ is the maximum possible entropy. With this new quantity R, self-organization implies that $\delta R/\delta t > 0$. Differentiating equation (8.4), we obtain

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{S(\mathrm{d}S_{\max}/\mathrm{d}t) - S_{\max}(\mathrm{d}S/\mathrm{d}t)}{S_{\max}^2} \tag{8.5}$$

our criterion for self-organization (namely, that R must spontaneously increase) is plainly

$$S\frac{\mathrm{d}S_{\max}}{\mathrm{d}t} > S_{\max}\frac{\mathrm{d}S}{\mathrm{d}t} \tag{8.6}$$

The implications of this inequality can be seen by considering two special cases [57]:

1. The maximum possible entropy $S_{\max}$ is constant; therefore $\mathrm{d}S_{\max}/\mathrm{d}t = 0$ and $\mathrm{d}S/\mathrm{d}t < 0$. Now, the entropy S depends on the probability distribution of the constituent parts (at least, those that are to be found in certain distinguishable states); this distribution can be changed by rearranging the parts, which von Foerster supposed could be accomplished by an "internal demon".
2. The entropy S is constant; therefore $\mathrm{d}S/\mathrm{d}t = 0$ and the condition that $\mathrm{d}S_{\max}/\mathrm{d}t > 0$ must hold; that is, the maximum possible disorder must increase. This could be accomplished, for example, by increasing the number of elements; however, care must be taken to ensure that S then indeed remains constant, which probably needs an "external" demon.

Looking again at inequality (8.6), we see how the labor is divided among the demons: $\mathrm{d}S/\mathrm{d}t$ represents the internal demon's efforts, and S is the result; $\mathrm{d}S_{\max}/\mathrm{d}t$ represents the external demon's efforts, and $S_{\max}$ is the result. There is therefore an advantage (in the sense that labor may be spared) in cooperating—e.g., if the internal demon has worked hard in the past, the external demon can get away with putting in a bit less effort in the present.

These considerations imply that water is an especially good medium in which self-assembly can take place because, except near its boiling point, it has a great deal of structure (Section 3.8) that it can sacrifice to enable ordering in S. Hence,

biological self-assembly of compact protein and nucleic acid structures takes place in an aqueous environment. Presumably thermophilic microbes that live at temperatures close to $100°C$ have some difficulties on this score.

8.2.3 The "Goodness" of the Organization

Examining again Figure 8.3, it can be asserted that both putative results of mixing slightly sticky cubelets together are organized, but most people would not hesitate to call the structure in (b) better organized than that in (a). Evidently there is some meaning in the notion of "good organization", even though it seems difficult to formulate an unambiguous definition. Can a system spontaneously (automatically) change from a bad to a good organization? This would be a reasonable interpretation of "self-organization", but has been proved to be formally impossible [10]. Consider a device that can be in any one of three states, A, B or C, and the device's operation is represented by some transformation, e.g.,

$$\downarrow \begin{matrix} A & B & C \\ B & C & A \end{matrix}.$$

Now suppose that we can give an input f into the device, and that the output is determined by the value of f, e.g.,

$$\begin{array}{c|ccc} \downarrow & A & B & C \\ \hline f_A & B & C & A \\ f_B & A & A & A \\ f_C & A & B & C \end{array}.$$

Spontaneous (automatic) operation means that the device is able to autonomously select its input. The different possible input values are here represented by a subscript indicating the state of the device on which the input now depends. However, this places severe constraints on the actual operation, because $f_A(B)$ (for example) is impossible; only $f_A(A)$, $f_B(B)$ and $f_C(C)$ are possible, hence the operation necessarily reduces to the simple transform, lacking any autonomy:

$$\downarrow \begin{matrix} A & B & C \\ B & A & C \end{matrix}.$$

Any change in f must therefore come from an external agent.

8.2.4 Particle Mixtures

Consider the shaking and stirring of a powder—a collection of small (with respect to the characteristic length scale of the final texture) particles of type A and type B initially randomly mixed up. The interaction energy v is given by the well-known Bragg–Williams expression

$$v = v_{AA} + v_{BB} - 2v_{AB} \tag{8.7}$$

where the three terms on the right are the interaction energies (enthalpies) for A with itself, B with itself, and A with B. One would guess that the mixture is miscible if $v < 0$, and immiscible otherwise. If the interaction energies are all zero, there will still be an entropic drive towards mixing (the case of the perfect solution). Edwards and Oakeshott [46] introduced the *compactivity X*, analogous to the temperature in thermodynamics, defined as

$$X = \partial V / \partial S \qquad (8.8)$$

where V is the volume (playing the role of energy) and S the entropy, defined in analogy to Boltzmann's equation as $S = k_B \ln \Omega$, where Ω is the number of configurations. If the number of particles of type A at a certain point r_i is $m_A^{(i)}$ (equal to either zero or one), and since necessarily $m_A^{(i)} + m_B^{(i)} = 1$, by introducing the new variable $m_i = 2m_A^{(i)} - 1 = 1 - 2m_B^{(i)}$ we have $m_i = \pm 1$. Defining $\phi = \langle m_i \rangle$, which Bragg and Williams give as $\tanh \phi v / k_B X$ [46], three regimes are identified depending on the interaction parameter $v/k_B X$ [46]:

- miscible: $v/k_B X < 1$, $\phi = 0$;
- domains of unequal concentrations: $v/k_B X > 1$, ϕ small ($X = v/k_B$ emerges as a kind of critical point)
- domains of pure A and pure B: $v/k_B X \gg 1$, $\phi = \pm 1$.

8.2.5 Mixed Polymers

The entropy of mixing entities is

$$S_{\text{mix}} = -k_B \sum_i \phi_i \ln \phi_i \qquad (8.9)$$

where ϕ_i is the volume fraction of the entities of the ith kind. The per-site (where each site is a monomer unit), free energy $\mathcal{M}$ of mixing two polymers A and B is

$$\mathcal{M} = \phi_A \ln \phi_A / N_A + \phi_B \ln \phi_B / N_B + \chi \phi_A \phi_B \qquad (8.10)$$

where N_B and N_B are the degrees of polymerization, and χ is the Flory–Huggins interaction parameter, given by (cf. equation 8.7)

$$\chi = (v_{AA} + v_{BB} - 2v_{AB})z/(2k_B T) \qquad (8.11)$$

where z is defined (for a polymer on a lattice) as the number of lattice directions. The first two terms on the right-hand side of equation (8.10), corresponding to the entropy of mixing (equation 8.9), are very small due to the large denominators, hence the free energy is dominated by the third term, giving the interaction energy. If $\chi > 0$, then phase separation is inevitable. For a well-mixed blend, however, the separation may take place exceedingly slowly on laboratory timescales, and therefore for some purposes nanotexture might be achievable by blending two immiscible polymers. However, even if such a blend is kinetically stable in the bulk, when prepared as a thin film on the surface, effects such as spinodal decomposition may be favored

due to the symmetry breaking effect of the surface (for example, by attracting either A or B). Separation can be permanently prevented by linking the two functionalities as in a block copolymer (Section 8.2.6).

8.2.6 Block Copolymers

One of the problems with mixing weakly interacting particles of two or more different varieties is that under nearly all conditions complete segregation occurs, at least if the system is allowed to reach equilibrium. This segregation is, however, frustrated if the different varieties are covalently linked together, as in a block copolymer. A rich variety of nanotexture results from this procedure. If A and B, assumed to be immiscible ($\chi > 0$) are copolymerized to form a molecule of the type AAA$\cdots$AAABBB$\cdots$BBB (a diblock copolymer), then of course the A and B phases cannot separate in the bulk, hence microseparation results, with the formation of domains with size ℓ of the order of the block sizes (that is, in the nanoscale), minimizing the interfacial energy between the incompatible A and B regions. One could say that the entropy gain arising from diminishing the A–B contacts exceeds the entropic penalty of demixing (and stretching the otherwise random coils at the phase boundaries). The block copolymer can also be thought of as a type of supersphere [136]. If χN is fairly small (less than 10) we are in the weak segregation regime and the blocks tend to mix; but in the strong segregation regime ($\chi N \gg 10$) the microdomains are almost pure and have narrow interfaces. As the volume fraction of one of the components (say A) of the commonly encountered coil–coil diblock copolymers increases from zero to one, the bulk morphologies pass through a well characterized sequence of body-centered cubic spheres of A in B, hexagonally packed cylinders of A in B, a bicontinuous cubic phase of A in a continuous matrix of B, a lamellar phase, a bicontinuous cubic phase of B in a continuous matrix of A, hexagonally packed cylinders of B in A, and body-centered cubic spheres of B in A.

When block copolymers are prepared as thin films (thickness d less than 100 nm) on a substratum (e.g., by spin-coating or dip-coating), the symmetry of the bulk system is broken, especially if one of the blocks of the copolymer is preferentially attracted to or repelled from the surface of the substratum. If $d < \ell$, the surface may be considered to have a strong effect on the structure of the thin film. For example, poly-2-vinylpyridine does not wet mica, and a polystyrene–polyvinylpyridine block copolymer thin film on mica has a structure different from that of the copolymer in the bulk [155]: structures such as antisymmetric surface-parallel lamellae, antisymmetric hybrid structures (cf. Stranski–Krastanov film growth, Section 8.1.2), and surface-perpendicular lamellae or columns are typically formed. There is at present considerable interest in such processes for fabricating photolithography masks in the nanoscale range more conveniently than by electron beam writing. Reticulated structures seem to have been investigated the most extensively: block copolymer micelles can be formed by dissolving the polymer in a fluid that is a solvent for only one of the components, and then used to coat surfaces, yielding a more or less regular array. This process has attracted interest as a route to making nanoporous membranes. For

example, polystyrene–polymethylmethacrylate (PMMA) copolymers, prepared with a volume fraction, such that there are cylindrical microdomains of the PMMA in the bulk, can be coated on a suitable substratum (silicon or silica) such that the cylinders are oriented normal to the surface. Exposure to ultraviolet light crosslinks the polystyrene but degrades the PMMA, which can then be selectively dissolved out of the film, leaving a nanoporous polystyrene membrane with a pore size controllable by varying the molecular weight of the copolymer [169].

One advantage of these polymer-based processes is the tremendous variety of starting materials available (by the same token, the systematic experimental investigation of the effects of compositional variation across the whole range of possibilities represents a huge undertaking). As well as changing the chemical nature of the monomers and the degrees of polymerization, the block copolymers have also been mixed with homopolymers as a way of modifying the characteristic scale of the texture [69].

8.2.7 The Addition of Particles to the Solid/Liquid Interface

Many self-assembly processes are based on the addition of nano-objects to a substrate. As noted in Section 8.2.5, a surface can have a symmetry-breaking effect. Consider a chemically and morphologically unstructured surface (medium 1) brought into contact with a fluid (medium 2) in which objects (medium 3) are suspended (i.e., their buoyancy is such that they move purely by diffusion (Brownian motion) and are not influenced by gravity). Suppose that the materials are chosen such that the interfacial free energy is negative (Section 3.2). On occasion, those in the vicinity of the interface will strike it. The rate of arrival of the particles at the substratum is proportional to the product of particle concentration c_b in the suspending medium and the diffusion coefficient D of a particle, the constant of proportionality depending on the hydrodynamic regime (e.g., convective diffusion); this rate will be reduced by a factor $1/\int_{\ell_0}^{\infty}[\exp(\Delta G_{123}(z)/k_B T) - 1]dz$ in the presence of an energy barrier; the reduction factor could easily be a hundred or a thousandfold. Once a particle of radius r adheres to the substratum, evidently the closest the center of a second particle can be placed to the first one is at a distance $2r$ from the center of the first; in effect the first particle creates an *exclusion zone* around it (Figure 8.5).

A corollary of the existence of exclusion zones is that the interface will be jammed (i.e., unable to accept a further particle) at a surface coverage of substantially less than 100%. The actual value of the jamming limit depends on the shape of the particle; for spheres it is about 54% of complete surface coverage. This process is known as random sequential addition (RSA). Although a random dispersion of particles in three dimensions is thereby reduced to a two-dimensional layer, the positions of the particles remain random: the radial distribution function is totally unstructured. Even if the particles can move laterally, allowing the interface to equilibrate in a certain sense, it is still jammed at a surface coverage of well below 100%.

Numerically Simulating RSA. The process is exceptionally easy to simulate: for each addition attempt one selects a point at random: if it is further than $2r$ from the center
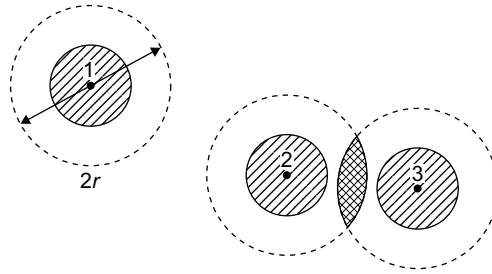
FIGURE 8.5

The concept of exclusion zone. The particles' projected area is hatched. The area enclosed by the dashed lines is the exclusion zone and has twice the radius of the actual particle. The exclusion zone is defined as that area within which no center of any particle can be placed without violating the condition of no overlap of hard bodies. The cross-hatched area marks the overlap of the exclusion zones of particles numbered 2 and 3. If the particles interact with each other with longer range forces than the hard body (Born) repulsion, then the radius is increased to an effective radius equal to the distance at which the particle–particle interaction energy $\Delta G_{323}(z)$ equals the thermal energy $k_B T$.

of any existing particle a new particle is added (the available area for less symmetrical shapes may have to be computed explicitly) and if it is closer then the attempt is abandoned. The success of this simple algorithm is due to the fortuitous cancellation of two opposing processes: correlation and randomization. In reality, if a particle cannot be added at a selected position because of the presence of a previously added one, it will make another attempt in the vicinity of the first one, because of the Rabinowitch ("cage") effect [111]; successive attempts are strongly positionally correlated. One other hand, as a particle approaches the interface through the bulk fluid, it experiences hydrodynamic friction, which exerts a randomizing effect; the two effects happen to cancel out each other [11].

Functions for Characterizing Nano-Object Addition. We have the fraction of occupied surface θ, equal to the product of the number of objects ν per unit area and the area a occupied by one object; and the fraction of surface ϕ available for adsorption (sometimes called the available area function). In general, we have for the rate of addition:

$$\mathrm{d}\theta/\mathrm{d}t = k_a c^* \phi(\theta) \tag{8.12}$$

where k_a is the addition rate coefficient (dependent upon the interfacial free energy function $\Delta G_{123}(z)$, see Section 3.2) and c^* is the effective bulk concentration (subsuming hydrodynamic and other factors). Much RSA research concerns the relation of ϕ to θ. An early theory relating them was Langmuir's: if small objects adsorb to discrete substratum sites larger than the particles,

$$\phi = 1 - \theta. \tag{8.13}$$

Substituting this into equation (8.12) and integrating, we see that in Langmuir adsorption the surface is completely filled up ($\theta \to 1$) exponentially in time (for a uniform rate of arrival of particles at the surface). In the absence of discrete sites (or in the presence of such sites that are smaller than the particles), the particles adsorb wherever they happen to arrive (assumed to be random locations). If a particle arrives such that its center would fall within the exclusion zone of a previously adsorbed particle (Figure 8.5) its adsorption attempt is rejected. Since the exclusion zone is four times as large as the particle, we should have

$$\phi = 1 - 4\theta \qquad (8.14)$$

but as θ increases, exclusion zones will overlap (Figure 8.5) and compensating terms have to be added, proportional to θ^2 for two overlapping particles, and so on up to six (the maximum number that can fit; in practice terms up to second or third order will usually be found to be sufficient):

$$\phi = 1 - b_1\theta + b_2\theta^2 + b_3\theta^3 + \mathcal{O}(\theta)^4 \qquad (8.15)$$

with $b_1 = 4$ and the coefficients b_2 and b_3 determined by purely geometrical considerations; $b_3 = 6\sqrt{3}/\pi$ is identical for both irreversible and equilibrium adsorption, whereas the coefficient b_3 varies from about 1.4 for irreversible (random sequential addition, RSA) to about 2.4 for equilibrium (reversible, whether via desorption and readsorption or via lateral movement) adsorption. In this case $\phi \to 0$ for $\theta < 1$; the "jamming limit" at which $\phi = 0$ is $\theta_J \approx 0.55$ for spheres adsorbing irreversibly.

The RSA formalism was developed in the context of particles interacting predominantly via hard body repulsion. The particle radius r is implicitly considered to be the hard body radius r. "Soluble" (stably suspended) particles must have repulsive particle–particle interactions and cannot in fact approach each other to a center-to-center distance of $2r$ but will behave as particles of an effective radius r', where r' is that value of z at which the total interfacial (IF) interaction energy (see Section 3.2) $\Delta G_{(323)}(z) \sim k_B T$, the thermal energy.

Generalized Ballistic Deposition. The ballistic deposition (BD) model describes objects falling onto a surface under the influence of gravity. Whereas in RSA if a particle attempts to land with its center within the exclusion zone around a previously adsorbed particle it is rejected, in BD the particle is not eliminated, since it is not buoyant enough to diffuse away, but rolls along on top of previously adsorbed particles until it finds space to adsorb. The coefficients of equation (8.15) are then different, namely $b_1 = b_2 = 0$ and $b_3 \approx -9.95$. BD and RSA can be combined linearly to give generalized ballistic deposition (GBD), where

$$\phi(\theta) = \phi^{\text{RSA}}(\theta) + j\phi^{\text{BD}}(\theta) \qquad (8.16)$$

with the parameter j defined as

$$j = p'/p \qquad (8.17)$$

where p' is the probability that a particle arriving via correlated diffusion ("rolling") at a space large enough to accomodate it will remain (i.e., will surmount any energy

barrier), and p is the probability that a particle arriving directly at a space large enough to accomodate it will remain. p is clearly related to the lateral interaction ("stickiness") of particles for each other, and as $j \to \infty$ the model describes nanoparticle aggregation at a surface. Essentially, the exclusion zones are thereby annihilated, and ϕ can be simplified to eqn (8.13).

Two-Dimensional Crystallization. If, however, the particles can adhere to each other on the interface, the possibility for organizing arises. This has been very clearly demonstrated when lateral mobility was expressly conferred on the particles by covering the substratum with a liquid-crystalline lipid bilayer and anchoring the particles (large spherical proteins) in the bilayer through a hydrophobic "tail" [142]. The particles structure themselves to form a two-dimensional ordered array (crystal). When such an affinity exists between the particles trapped at the interface, the exclusion zones are annihilated. From this fact alone (which can be very easily deduced from the kinetics of addition [142]) one cannot distinguish between random aggregation forming a diffusion-limited aggregate (DLA) (cf. reaction-limited aggregation, RLA) and two-dimensional crystallization; they can generally be distinguished, however, by the fact that in the latter the crystal unit cell size is significantly bigger than the projected area of the particle (cf. a three-dimensional protein crystal: typically about 70% of the volume of such a crystal is occupied by solvent). The process of two-dimensional crystallization has two characteristic timescales: the interval τ_a between the addition of successive particles to the interface

$$\tau_a = 1/[aF\phi(\theta)] \tag{8.18}$$

where a is the area per particle, F is the flux of particles to an empty surface (proportional to the bulk particle concentration and some power < 1 of the coefficient of diffusion in three dimensions), and ϕ is the fraction of the surface available for addition, which is some function of θ, the fractional surface coverage of the particles at the interface; and the characteristic time τ_D for rearranging the surface by lateral diffusion (with a diffusion coefficient D_2)

$$\tau_D = a/(D_2\theta) \tag{8.19}$$

If $\tau_D \gg \tau_a$ then lateral diffusion is encumbered by the rapid addition of fresh particles before self-organization can occur and the resulting structure is indistinguishable from that of random sequential addition. Conversely, if $\tau_a \gg \tau_D$ there is time for two-dimensional crystallization to occur. Note that some affinity-changing conformational change needs to be induced by the interface, otherwise the particles would already aggregate in the bulk suspension. In the example of the protein with the hydrophobic tail, when the protein is dissolved in water the tail is buried in the interior of the protein, but partitions into the lipid bilayer when the protein arrives at its surface.

Another intriguing example of interfacial organization is the heaping into cones of the antifreeze glycoprotein (AFGP), consisting of repeated alanine–alanine–glycosylated threonine triplets, added to the surface of a solid solution of nanocrystalline $Si_{0.6}Ti_{0.4}O_2$ [102]. Under otherwise identical conditions, on mica the

glycoprotein adsorbs randomly sequentially. Despite the simplicity of the structure of AFGP, its organization at the silica–titania surface appears to be a primitive example of programmable self-assembly (Section 8.2.8).

8.2.8 Programmable Self-Assembly

The addition of monomers to a growing crystal is self-assembly (Section 6.4), but the result is not useful as a nanofacturing process because there is nothing to limit growth, except when the imposed conditions are carefully chosen (Section 6.1.2). Crystallization is an example of passive or nonprogrammable self-assembly. Its ubiquity might well engender a certain pessimism regarding the ultimate possibility of realizing true self-assembly, the goal of which is to produce vast numbers of identical, prespecified arbitrary structures. Yet, in biology, numerous examples are known (see also Section 1.4): the final stages of assembly of bacteriophage viruses, of ribosomes and of microtubules; they occur not only *in vivo* but can also be demonstrated *in vitro* by simply mixing the components together in a test-tube. As apparent examples of what might be called "passive" self-assembly, in which objects possessing certain asymmetric arrangements of surface affinities are randomly mixed and expected to produce ordered structures [57], they seem to contradict the predictions of Sections 8.2.2 and 8.2.3.

It has long been known that biomolecules are constructions: that is, they have a small number of macroscopic (relative to atomic vibrations) degrees of freedom, and can exist in a small number (≥ 2) of stable conformations. Without these properties, the actions of enzymes, active carriers such as hemoglobin, and the motors that power muscle, etc., are not understandable. Switching from one conformation to another is typically triggered by the binding or dissociation of a small molecule; for example the "substrate" of an enzyme, or adenosine triphosphate (ATP). The initial collision of two particles is followed by a conformational change in one or both of them; e.g.,

$$A + B + C \rightarrow A\text{–}B^* + C \rightarrow A\text{–}B^*\text{–}C \tag{8.20}$$
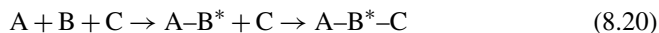
where the asterisk denotes a changed conformation induced by binding to A; C has no affinity for B, but binds to B^*. This process is illustrated in Figure 8.6, and is called programmable self-assembly (PSA). Graph grammar, which can be thought of as a set of rules encapsulating the outcomes of interactions between the particles [93, 94] (cf. stigmergic assembly, Section 8.2.12), is useful for representing the process. The concept of graph grammar has brought a significant advance in the formalization of programmable self-assembly, including the specification of minimal properties that must be possessed by a self-assembling system (e.g., the result implying that no binary grammar can generate a unique stable assembly [95]).

While models of programmably self-assembling robots have been created in the macroscale, artificially synthesizing molecules with the required attribute remains a challenge. Biology, however, is full of examples (e.g., the "induced fit" occurring when an antibody binds an antigen). Microscopically, these are manifestations
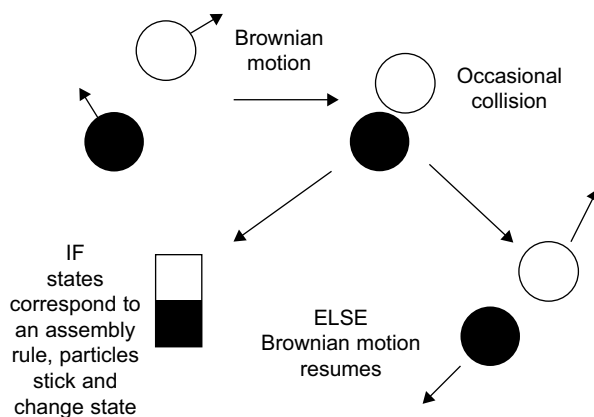
FIGURE 8.6

Illustration of programmable self-assembly, with a primitive local rule.

of cooperativity (Section 3.6). However, the cooperativity is systemic in the sense that entire macromolecules may be acting in concert as a system (cf. Sections 8.2.10 and 8.2.11).

8.2.9 Superspheres

If the competing interactions have different sign and range, ordered structures of definite size can assemble spontaneously. This provides a simple example of programmable self-assembly. Consider nanoparticles suspended in water and weakly ionized such that they all carry the same electrostatic charge. When the suspension is stirred, suppose that the repulsive electrostatic force is too weak to overcome the attractive Lifshitz–van der Waals (LW) force when two particles happen to collide. Therefore, every collision will lead to sticking, and aggregates will slowly form. The LW force is, however, very short range and can only act between nearest neighbors. The electrostatic force, on the other hand, has a much longer range, and can therefore be summed over the entire aggregate. Ultimately the aggregate will become large enough for the summed electrostatic repulsion to exceed the LW nearest neighbor attraction. The result is monodisperse "superspheres" (i.e., aggregates of small (maybe spherical) particles).

Weakly electrostatically charged quantum dots (nanoparticles) suspended in water aggregate to form uniformly sized superspheres containing several hundred nanoparticles. Nearest neighbors interact with weak, short range LW interactions, which easily dominate the slight electrostatic repulsion between them. Because, however, the electrostatic interaction is long range (it can be tuned by varying the ionic strength of the aqueous solution), the overall electrostatic repulsion within a supersphere gradually accumulates, and when a certain number of nanoparticles have been aggregated,

the electrostatic repulsion exceeds the attractive LW force between nearest neighbors [136]. To form superspheres, the attractive interaction should be short range, and the repulsive interaction should be long range.

An interesting kind of nanostructure was shown in Figure 6.6(d). The small spheres (called micelles) have polar heads ionizable in water, resulting in q elementary charges on the surface of the sphere, which exert an expanding pressure

$$\Delta P^{(el)} = (qe)^2/(8\pi\epsilon R^4) \tag{8.21}$$

the size of the micelle adjusts itself to exactly compensate the Laplace contraction (equation 2.2); in consequence such micelles are highly monodisperse because they are at a local energy minimum.

8.2.10 Biological Self-Assembly

It has long been known that many biological systems exhibit remarkable capabilities of assembling themselves starting from a randomly arranged mixture of components. These include the bacteriophage virus (the final stages of assembly), and proteins and ribonucleic acids (RNA), which can be spontaneously transformed from a random coil of the as-synthesized linear polymer to a compact, ordered three-dimensional structure (Section 8.2.11). It is clear that the starting precursors of the final structures have to be very carefully designed—this is a carefully tuned example of programmable self-assembly (PSA) in action (Section 8.2.8).

Although appreciation of self-assembly in biology has played a hugely important inspirational role, the highly specialized chemistry of living systems, the fragility of many of its products, and its inherent variability at many levels have made it unsuitable for mimicking directly and incorporating into our present industrial system (cf. Section 8.2.13). This is particularly so in the case of the food industry. The extreme complexity, both structural and chemical, of its products and the relative ease of letting them grow renders efforts to manufacture food synthetically largely superfluous.

More debatable is whether self-assembly offers a viable route to creating artificial cells for photovoltaic solar energy conversion. The natural system comprises the photosystems embedded within the chloroplast, whose maintenance requires the rest of the machinery of the cell, and whose effective operation requires a macroscopic structure of the plant (stem and branches) to support the leaves in which the chloroplasts are embedded. The classical artificial system is the semiconductor photovoltaic cell. Can its efficiency, as well as ease of manufacture, be enhanced by using nanostructured photoactive components? Most appraisals of the photovoltaic cell as a "renewable" or "sustainable" energy source pay scant regard to the entire manufacturing cycle, and the key question of working lifetime under realistic conditions is scarcely addressed by laboratory trials. Given the history of considerable efforts to more closely mimic the molecular machinery of the natural photosystems in a nanoconstruction, it has been natural to look at extending the mimicry beyond discrete components to systems. Nevertheless, except for the ultimate, and

still hypothetical, stage of molecularly manufactured nanosystems, none of the proposed solutions comes anywhere near the performance (considered as an overall system) of natural photosynthesis, which can simply be left to grow over vast areas (provided water is available).

The observation that preassembled bacteriophage components (head, neck and legs) could be mixed in solution exerted a profound inspiration on the world of "shake and bake" advocates. These components are essentially made up of proteins—heteropolymers made from irregular sequences chosen from the 20 natural amino acids of general formula $H_2N–C^{(\alpha)}HR–COOH$, where R is naturally one of 20 different side chains (residues), ranging from R=H in glycine, the simplest amino acid, to elaborate heterocycles such as $R=CH_2–[C_8NH_6]$ in tryptophan. The conformation and hence affinity of a protein depends on environmental parameters such as the pH and ionic strength of the solution in which it is dissolved, and this is one mechanism for achieving programmability in assembly, since the local pH and ion concentration around a protein molecule depend on the amino acids present at its surface. These factors also determine the conformation of the highly elongated, so-called fibrous proteins such as fibronectin, now known to consist of a large number of rather similar modules strung together (the "daisy chain" or "pearl necklace" model). Some other examples of biological self-assembly have already been mentioned in Section 8.2.8. A further one is provided by the remarkable S-layers with which certain bacteria are coated. One should also mention the oligopeptides found in fungi (e.g., alamethicine) and the stings of bees (mellitin) and wasps (mastoparan) that self-assemble into pores when introduced into a bilayer lipid membrane. But, biological self-assembly and self-organization is by no means limited to the molecular scale (see Section 8.2.12).

8.2.11 **Biopolymer Folding**

Biopolymer "folding" means the transformation of a linear polymer chain, whose monomers are connected only to their two nearest neighbors, and which adopts a random coil in solution, into a complex three-dimensional structure with additional (hydrogen) bonds between distant monomers.

Predicting the final three-dimensional structure is *prima facie* a difficult problem. Energetics are clearly involved, because bonds between distant monomers form spontaneously (if geometric constraints are satisfied), releasing enthalpy and hence lowering the free energy. On the other hand, this raises the entropy because the chain becomes constrained. Finding the free energy minimum by systematically searching configuration space is a practically impossible task for a large molecule with thousands of atoms—it would take longer than the age of the universe. Since the protein molecule can fold within seconds, it seems clear that the solution to the problem lies in determining the pathways. The *Principle of Least Action* (PLA) is useful for this purpose: the most expedient path is found by minimizing the action.

Action is the integral of the Lagrangian $\mathcal{L}(= L - F$ for conservative systems, where L and F are respectively the kinetic and potential energies). Minimization of the

action is an inerrant principle for finding the correct solution of a dynamical problem; the difficulty lies in the fact that there is no general recipe for constructing $\mathcal{L}$.

A solution leading to a successful algorithm has been found for the folding of ribonucleic acid (RNA) [52]. Natural RNA polymers are made up from four different "bases", A, C, G and U (see Section 4.1.4). As with DNA, multiple hydrogen bonding favors the formation of G–C and A–U pairs, which leads to the appearance of certain characteristic structures. Loop closure is considered to be the most important folding event. F (the potential) is identified with the enthalpy; that is, the number n of base pairings (contacts); and L corresponds to the entropy. At each stage in the folding process, as many as possible new favorable intramolecular interactions are formed, while minimizing the loss of conformational freedom (the principle of sequential minimization of entropy loss, SMEL). The entropy loss associated with loop closure is ΔS_{loop} (and the rate of loop closure $\sim \exp(\Delta S_{\text{loop}})$); the function to be minimized is therefore $\exp(-\Delta S_{\text{loop}}/R)/n$, where R is the universal gas constant. A quantitative expression for ΔS_{loop} can be found by noting that the N monomers in an unstrained loop ($N \geq 4$) have essentially two possible conformations, pointing either inwards or outwards. For loops smaller than a critical size N_0, the inward ones are in an apolar environment, since the nano-enclosed water no longer has bulk properties, and the outward ones are in polar bulk water. For $N < N_0$, $\Delta S_{\text{loop}} = -RN \ln 2$ (for $N > N_0$, the Jacobson–Stockmayer approximation based on excluded volume yields $\Delta S_{\text{loop}} \sim R \ln N$).

In summary, SMEL applied to biopolymer folding is a least-action principle that involves sequentially maximizing the number of contacts while minimizing entropy loss.

A similar approach can be applied to proteins [53]. However, in proteins the main intramolecular structural connectors (apart from the covalent bonds between successive amino acid monomers) are the backbone hydrogen bonds, responsible for the appearance of characteristic structures such as the alpha helix but which, being single, are necessarily weaker than the double and triple hydrogen bonds in DNA and RNA. They therefore need to be protected from competition for hydrogen bonding by water, and this can be achieved by bringing amino acids with apolar residues to surround the hydrogen bonds [54]. This additional feature, coupled with the existence of multiple conformational states already referred to (Section 8.2.8) means that proteins are particularly good for engaging in programmable self-assembly, a possibility that is, of course, abundantly made use of in nature.

8.2.12 Biological Growth

The development of an embryo consisting of a single cell into a multicellular organism is perhaps the most striking example of self-organization in the living world. The process of cell differentiation into different types can be very satisfactorily simulated on the basis of purely local rules enacted by the initially uniform cells (see [110] for the modeling of neurogenesis). On a yet larger scale, it is likely that the construction

of nests by social insects such as ants and wasps relies on simple rules held and enacted by individual agents (the insects) according to local conditions; this process has been called stigmergy and is evidently conceptually related to programmable self-assembly (Section 8.2.8).

Reproducibility is interpreted somewhat differently by living processes in comparison with the mass, standardized manufacturing processes of the Industrial Revolution paradigm. Although the basic building blocks (e.g., proteins) of living organisms are identical, templated from a master specification encoded in DNA (see Section 4.1.4), organisms are not identical in the way that very large-scale integrated circuits (VLSIs) are. What is specified (genetically) is at most an algorithm (subject to local environmental influence) for constructing an organism, or maybe just an algorithm for an algorithm. Reproducibility exists at the functional level: every termite's nest is able to protect its denizens from a wide range of hazards under varied local conditions of topography, soil type and vegetation; every dog has a brain able to ensure its survival for a certain period; and so forth. This concept of an algorithm specifying how the construction should take place is what is used for building the nests of social insects, which are constructed stigmergically—each insect is armed with rules specifying what to do under a variety of local circumstances.

There are some hitherto relatively unexploited niches for creating nano-objects via biological growth. For example, the magnetic protein ferritin, which is constituted from an iron oxide core surrounded by protein, can be made on a large scale by low-cost biotechnological manufacturing routes and due to its strict monodispersity can be used, after eliminating the protein matter, in magnetic memory devices [163].

8.2.13 Self-Assembly as a Manufacturing Process

Practical, industrial interest in self-assembly is strongly driven by the increasing difficulty of reducing the feature sizes that can be fabricated by semiconductor processing technology. Pending the ultimate introduction of productive nanosystems based on bottom-to-bottom fabrication (Section 8.3), self-assembly is positioned as a rival to the "top–down" processes that currently constitute the majority of nanofacture. For certain engineering problems, such as membranes for separating valuable resources (such as rare earth ions) from a matrix in which they are present in very diluted form, self-assembly may already be useful to manufacture regular structures such as dots or stripes over an indefinitely large area; passive self-assembly might be able to produce such structures with feature sizes at the molecular scale of a few nanometers.

As well as the mimicry of natural surfaces for biological molecular recognition for sensing and other nanomedical applications (Chapter 4), self-assembly fabrication techniques should be generally applicable to create feature sizes in the nanometer range (i.e., 1–100 nm), which is still relatively difficult to achieve, and certainly very costly (limiting its application to ultrahigh production numbers) using conventional top–down semiconductor processing techniques. Furthermore, self-assembly is more flexible regarding the geometry of the substrata to which it can be applied; e.g.,

there is no restriction to planar surfaces. This is particularly advantageous if it is (for example) desired to engineer the surfaces of intricately curved objects such as vascular stents in order to increase their biocompatibility.

A current challenge of "bottom–up" nanotechnology is the formulation of design rules. Generally we ask: "how to design X to carry out a desired function Y?" *Design* means, essentially, specifying structure, hence the question can be reworded as: "what structure will give function Y?"—"function" being interpreted as properties and performance; and structure being constituted from certain *numbers* of different *types* of entities, *connected* together in a certain way. Micro (and macro) engineering benefits from vast experience; i.e., a look-up table with structure in the left-hand column and function in the right. There is less experience in the nanoworld, but if a nanostructure is simply a microstructure in miniature (often it is not), this experience can be transferred. Undoubtedly one needs to ask whether the properties of matter change at the nanometer scale (Chapter 2); i.e., do we need a new set of structure–property relations? These relations may also affect the fabrication process.

Regarding the formulation of assembly rules, again current experience is exiguous and rules come mainly from knowledge of certain biological processes. Typically they are either very general (such as the Principle of Least Action, see Section 8.2.11), without any specific indication of how to apply them to a particular case, or very specific and possibly quite esoteric (e.g., relying on chemical intuition). Challenges lie in the adaptation of general principles to specific cases, and in the formalization and generalization of known specific heuristic rules (or intuition). Indeed, the main disadvantage of bottom–up is that the process is not well understood theoretically. Hence although we need to be able to, at present we cannot design the starting objects (precursors) to achieve a specified final device.

8.3 BOTTOM-TO-BOTTOM METHODS

Essentially, the contribution of nanotechnology to the effort of ever improving machining accuracy is simply to take it to the ultimate level, in the spirit of "shaping the world atom-by-atom" (the subtitle of a report on nanotechnology prepared under the guidance of the US National Science and Technology Council Committee on Technology in 1999). Rather like the chemist trying to synthesize an elaborate multi-functional molecule (e.g., adding conjugated olefins to provide color, and hydroxyl groups to provide Lewis acid/base interactions), the materials nanotechnologist aims to juxtapose different atoms to achieve multifunctionality. This approach is known as mechanosynthetic chemistry or, in its large-scale industrial realization, as molecular manufacturing. The essential difference from chemistry is the eutactic environment, in which every atom is placed in a precise location.

The famous experiment of Schweizer and Eigler, in which they rearranged xenon atoms on a nickel surface to form the logo "IBM" [151], represented a first step

in this direction. Since then there has been intensive activity in the area, both computational and experimental, but it still remains uncertain to what extent arbitrary combinations of atoms can be assembled disregarding chemical concepts such as affinity and potential energy surfaces, and whether the process can ever be scaled up to provide macroscopic quantities of materials.

Also known as molecular manufacturing or mechanosynthesis or "pick and place" chemistry, bottom-to-bottom methods literally construct things atom-by-atom. In other words it is chemistry with positional control; i.e., taking place in a eutactic environment. This is what is sometimes called "hard" or "true" nanotechnology (in the sense of being uncompromisingly faithful to Feynman's original vision [56]).

A specific realization based on carbon (mainly diamondoid) structures has been elaborated by Drexler and others. This envisages an appropriately functionalized molecular tool driven by mechanical forces (such as the tip of a scanning probe microscope), which abstracts hydrogen from passivated surfaces to form radicals ("dangling bonds"), where other atoms can be added. Hence this approach is also called "tip-based nanofabrication".

8.3.1 Tip-Based Nanofabrication

The use of scanning probe microscopies (SPM) to characterize structure at the nanoscale suggests the reciprocal use of SPM-based methods to generate such structure, that is, by picking up atoms from a store and placing them exactly where required, as was first proposed by Drexler [43] (Figure 8.7).

Regardless of the chemical elements involved, the current molecular manufacturing paradigm involves tips—in fact it is the technology of the scanning probe microscope (Section 5.1.1). Schweizer and Eigler's manipulation of xenon atoms on an ultracold nickel surface to form the letters "IBM" was an iconic demonstration of
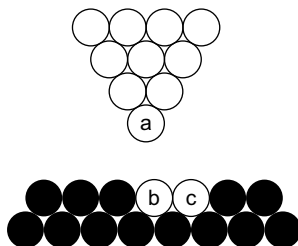
FIGURE 8.7

Illustration of tip-based nanofacture (in cross-section). Tip and substrate can move relative to each other. Each circle or disc corresponds to an atom (or superatom, see Section 8.3.2). The working tip is marked a. It can engage in vertical interchange (of atoms a and b) and lateral interchange (of atoms b and c). Each of these concerted actions involves an intermediate state in which two or more atoms are simultaneously bonded to the tip [157].

the possibilities [151], but also of the extreme laboriousness of the procedure (it took almost 24 hours—and had to be carried out at 4 K).

Curiously, advance has meanwhile been very slow, possibly because of a lack of long-term investment and clear direction, despite the large overall funds spent on nanotechnology in the USA and elsewhere (too much funding has been allocated to incremental work with short-term goals). A more recent landmark was the demonstration of atomic extraction using purely physical forces [130, 131]— 13 years after Schweizer and Eigler's demonstration [151]. Clearly this approach is still very much in its infancy. Even the issue of automatic control of the tip (Schweizer and Eigler manipulated their tip manually), which would enable one to simply specify the desired coordinates of each atom, and which would also undertake the necessary checks to ascertain whether each atom had been correctly placed, has received minimal attention. What has been undertaken is extensive calculations, mainly using density functional theory (DFT), of mechanosynthetic reactions; the slowly advancing experimental work is lagging far behind.

Although nanoscale miniature tools work extremely fast (typically in the GHz range), high throughput of nanofacture can only be achieved by massive parallelization, which in turn is only feasible if the required tools can make themselves—this was also part of Feynman's original vision. The IBM "Millipede" project, in which large numbers of SPM tips work in parallel, still falls many orders of magnitude short of achieving the throughput that would be required for materials fabrication yielding macroscopic quantities.

In summary, the main challenges of mechanosynthesis are:

1. Experimental verification of the calculations of rudimentary processes;
2. Extension of work on C, Si to cover all atoms;
3. Software control;
4. Strategies to increase throughput.

8.3.2 **Nanoblocks**

Significant acceleration of mechanosynthesis could take place if "nanoblocks"—preassembled (possibly by self-assembly, Section 8.2.1) units that may comprise dozens or hundreds (or more) atoms—are manipulated via bottom-to-bottom methods. This immediately increases throughput (in terms of the volume of artifacts produced) by a factor n_b, the number of atoms per block. The principles of software control should remain unchanged. The basic idea of nanoblocks is to retain the quality, specificity and individuality of quantum objects while having the relative ease of manipulation (compared with individual atoms) of nanoparticles. Regarding the choice of blocks, there are three main possibilities:

1. Multi-atom clusters (also known as superatoms) [34]; examples are fullerenes (see Section 9.3), $Al_{13}K$ and K_3As_7. Some of these have properties analogous to an atom (Section 2.5);
2. Nanoparticles or nanorods prepared conventionally (Sections 6.1 and 6.2);
3. Biopolymers, especially globular proteins or RNA.

An advantage of 1. and 3. are strict identity of the blocks, whereas with 2. one will generally have a distribution of sizes and possibly of some other properties too. An advantage of 2. and 3. is that they can readily be prepared in large quantities (this is also true for some kinds of clusters, notably fullerenes (Section 9.3). In all cases doping or modification is generally possible. Some of these blocks spontaneously self-assemble into regular structures (superlattices) [34], but their greatest potential use is in bottom-to-bottom fabrication.

In some sense nanoblocks are the hardware equivalent of finite element analysis.

8.3.3 Dip Pen Nanolithography

At a length scale slightly larger than that of atoms, dip pen nanolithography (DPN) has been developed by Chad Mirkin as a way of placing molecules in finely resolved zones of a substratum. The technique works by coating the scanning probe microscope tip with a weakly adherent molecular ink (e.g., proteins, relatively weakly adhering to a hydrophilic tip). When it is desired to write, the tip is lowered to the vicinity of the substratum, to which the affinity of the ink is much stronger, and the ink's molecules are transferred to the substratum through a water meniscus—the molecules ("ink") are transferred to the substratum by capillary action (Section 3.3) [88]. It follows that writing is strongly dependent on the ambient relative humidity. Although not atomically precise manufacturing, it allows features of the order of 100 nm to be written, and may be considered as featuring an approximately eutactic environment.

DPN was presumably inspired by printed electronic devices. Circuits can be fabricated at extremely low cost by printing onto a suitable substrate. Conventional processes such as screen printing and inkjet are suitable, with inks formulated using "pigments" that are conductive or semiconductive nanoparticles. This technology is especially attractive for radio frequency identification tags (RFID), which are expected to become widely used in packaging, and as security devices on products and even on documents if they can be produced at sufficiently low cost.

8.4 SUMMARY

Top–down methods (exemplified by ultraprecision engineering and semiconductor processing) constitute the bulk of current industrial nanotechnology. Due to the enormous expense of the capital equipment required, however, it is impractical for use other than for very high volume products (such as computer or cellular phone chips) or very unique products for which a high price is affordable (giant astronomical telescopes; long-range spacecraft).

The "real" vision of nanotechnology (especially associated with Feynman and Drexler) is based on mechanosynthesis (chemistry with positional control), possibly facilitated by using pre-constructed nanoblocks as the elementary units of fabrication. A productive nanosystem is based on assemblers, devices that are themselves in the nanoscale, hence the method is also known as bottom-to-bottom. Because of their

minute size, the only practical way to fabricate large or large quantities of entities is for the assemblers to first assemble copies of themselves, which then all work in parallel. The practical realization of this vision is focused on tip-based methods inspired by the scanning probe instruments used in nanometrology; at present single objects comprising of the order of ten atoms can be made in this way.

Originally inspired by biology, a third approach is based on creating objects (which could be nanoblocks) capable of spontaneously assembling into useful structures. This method is known as bottom–up or self-assembly and has been quite successful at creating regular structures (e.g., nanoporous membranes for separating vapors) but the creation of arbitrary geometries requires programmable self-assembly. In other words bottom–up is good at creating materials but not for creating devices. Biological nano-objects have this ability, but it is extraordinarily difficult to reverse-engineer them and use the knowledge to create synthetic analogs.

These three methods can sometimes be advantageously combined; for example, self-assembly could be used to make a mask for photolithography more cheaply and quickly than electron beam writing, which is then used in top–down fabrication (e.g., to create superhydrophobic surfaces).

8.5 **FURTHER READING**

A.W. Castleman Jr., S.N. Khanna, Clusters, superatoms and building blocks of new materials, J. Phys. Chem. C 113 (2009) 2664–2675.

H. Frauenfelder, From atoms to biomolecules, Helv. Phys. Acta 57 (1984) 165–187.

E. Kellenberger, Assembly in biological systems, in: Polymerization in Biological Systems, CIBA Foundation Symposium 7 (new series), pp. 189–206, Elsevier, Amsterdam, 1972.

A.G. Mamalis, A. Markopoulos, D.E. Manolakos, Micro and nanoprocessing techniques and applications, Nanotechnol. Percept. 1 (2005) 63–73.

R.C. Merkle, Molecular building blocks and development strategies for molecular nanotechnology, Nanotechnology 11 (2000) 89–99.

P. Schaaf, J.-C. Voegel, B. Senger, Irreversible deposition/adsorption processes on solid surfaces, Ann. Phys. 23 (1998) 3–89.

J. Storrs Hall, Architectural considerations for self-replicating manufacturing systems, Nanotechnology 10 (1999) 323–330.

Carbon-Based Nanomaterials and Devices

9

Carbon-based materials and devices are dealt with in this separate chapter because of their unique importance and versatility. They represent the epitome of nanotechnology.

Carbon has long been an intriguing material because of its two very well known allotropes, diamond and graphite, which are so different from one another that if one did not happen to know they have the same elemental composition one might imagine that they are different elements. Although researchers working on carbon have long been aware of other forms, typically dubbed "carbon filaments" and "atypical char" (along with "amorphous carbon"), usually they were regarded as a nuisance and discarded if their formation could not be avoided. The importance of the recent "discoveries" of fullerenes (soluble carbon) and carbon nanotubes (carbon filaments) resides in the fact that their structures were elucidated for the first time. Even nowadays, with ever improving methods of industrial synthesis, yields of carbon nanotubes (CNT) and fullerenes are significantly less than 100%; the desired product is embedded in a gangue of uncharacterized or defective material that still merits the name "amorphous carbon". The most recent "discovery" is that of graphene, which is simply one of the many parallel sheets constituting graphite; here the ingenuity resided in the preparation of a single isolated sheet, which opened the possibility of examining experimentally what was already a well-studied material theoretically.

The electron configuration of carbon is $1s^1 2s^2 2p^2$. Its four outer shell electron orbitals are 2s, $2p_x$, $2p_y$ and $2p_z$; the four valence electrons may hybridize them into sp^1, sp^2 and sp^3, corresponding to a carbon atom bound to 2, 3 and 4 neighboring atoms, respectively. Diamond is composed entirely of sp^3 orbitals: it is ultrahard, a wide band-gap dielectric, has good thermal conductivity and is rather

189
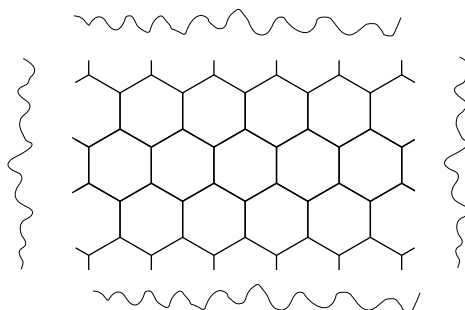
FIGURE 9.1

Part of a graphene sheet, showing the sp^2 chemical bonds. There is a carbon atom at each intersection. The length of each bond is about 0.3 nm.

inert chemically. Graphite is composed entirely of sp^2 orbitals: it is soft, and an electrical conductor, has highly anisotropic thermal conductivity and is very reactive chemically.

The three new materials, graphene, carbon nanotubes and fullerenes, can be called "nanocarbon". Like graphite, they are composed entirely of sp^2 orbitals, but fullerenes contain 12 pentagons and have some sp^3 character. They correspond to nanoplates, nanofibers and nanoparticles, respectively (see Figure 6.2). Rolling up graphene (Figure 9.1) into the smallest possible tube makes a single-walled carbon nanotube, and curling it up into the smallest possible sphere makes a fullerene—conceptually, that is; the materials are not actually fabricated that way. These nanomaterials have no bulk equivalents, discounting the fact that graphite is made up of endlessly repeating stacks of graphene.

Carbon nanotubes have hitherto attracted most commercial interest. Table 9.1 summarizes some of the remarkable properties of these materials. At present, they are most commonly commercially exploited by embedding them in a matrix to form some kind of composite.

Table 9.1 Some properties of bulk and nanoscale carbon materials[a]

Property	Unit	Diamond	Graphite	CNT
Young's modulus	$N\,m^{-2}$	10^9	10^{10}	10^{12}
Thermal conductivity	$W\,m^{-1}\,K^{-1}$	2000	20	3000
Electrical resistivity	$\Omega\,m$	10^{12}	10^{-5}	$<10^{-6}$

[a]*The given values are only approximate, in order to enable a rough comparative idea of the material properties to be formed. Actual measured values still depend on many experimental details.*

9.1 GRAPHENE

Inspired by learning about naphthalene and anthracene, countless school children have doubtless doodled endless fused polyaromatic rings. It has long been known that graphite is composed of stacks of such polyaromatic sheets, which are called graphene. Due to convincing theoretical work, it was however long believed that two-dimensional crystals cannot exist. The ostensive demonstration of their existence (graphene sheets) have, *post hoc*, led to the explanation that their stability is due to undulations of the sheet.

The graphene lamellae stacked to make bulk graphite were known from the ease of their detachment (as shown, e.g., by writing with graphite on paper) to be only weakly bound to each other. Individual sheets of graphene can actually be peeled off graphite using adhesive tape. Alternatively, a crystal of silicon carbide can be heated under vacuum to $1300\,°C$; the silicon evaporates and the remaining carbon slowly reorganizes to form some graphene.

9.2 CARBON NANOTUBES

The synthesis and characterization of carbon nanotubes were first reported in the scientific literature by Iijima in 1991. Since then they have been intensively studied both theoretically and experimentally. Great advances in fabrication techniques have been made, and nowadays it is possible to produce high-quality carbon nanotubes in reasonable quantities at least for research purposes.

It was long known that carbon filaments are formed by passing hydrocarbons over hot metal surfaces, especially iron and nickel. The actual nature of carbon nanotubes was however only established relatively recently (by Iijima in 1991): they are seamlessly rolled up tubes of graphene (Figure 9.2). The ends of the tubes may be open or "capped" with what is essentially a hemisphere of fullerene. Multiwalled carbon nanotubes (MWCNT) most typically consist of several concentric tubes of graphene nested inside each other (Figure 9.3). A form in which graphene is rolled up to give a spiral cross-section is also known.

An important feature of single walled carbon nanotubes (SWCNT) is the chiral (or wrapping or rollup) vector $\mathbf{C}_\chi$, defined as

$$\mathbf{C}_\chi = n\mathbf{a}_1 + m\mathbf{a}_2 \equiv (n, m) \qquad (9.1)$$

where the unit vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ have, respectively, the directions from the ith to the $i+2$th carbon on any hexagon, and from the ith to the $i+4$th carbon; $|\mathbf{a}_1| = |\mathbf{a}_2| = \sqrt{3}a_{C-C} = 0.246$ nm, where a_{C-C} is the nearest-neighbor distance between two carbon atoms, and thus the magnitude of C_χ is $0.246\sqrt{n^2 + nm + m^2}$; n and m are integers with $m \leq n$. The chiral vector connects crystallographically equivalent sites on the graphene lattice; $n\,(\geq 3)$ ("twist") and $0 \leq m \leq n$ ("countertwist") are the integers required to ensure translational symmetry of the lattice. $\mathbf{C}_\chi$ defines the

FIGURE 9.2

A single walled carbon nanotube (SWCNT): a single graphene layer rolled into a seamless tube. Reproduced with permission from [24].
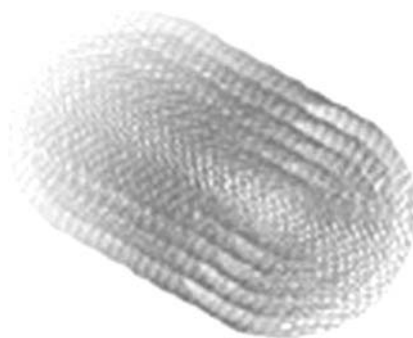
FIGURE 9.3

A multiwalled carbon nanotube (MWCNT): concentric single wall nanotubes of different diameters nested within each other. Reproduced with permission from [24].

disposition of the hexagonal rings of the carbon nanotube or in other words describes how the graphene sheet is wrapped. If $n = m$ the SWCNT is called armchair; if $m = 0$ then the SWCNT is called zig-zag; otherwise it is simply called chiral. The diameter in nanometers is given by

$$d = C_\chi / \pi \tag{9.2}$$

the chirality by

$$\chi = \sin \pi (1 - a_2/a_1); \tag{9.3}$$

and the chiral angle is

$$\theta_\chi = \arcsin \frac{\sqrt{3m}}{2\sqrt{n^2 + nm + m^2}}. \qquad (9.4)$$

The ratio a_2/a_1 provides a measure of helicity; it can range from zero (zig-zag) to 1 (armchair). The chirality ranges from zero (for zigzag and armchair nanotubes) to 1 if $a_2/a_1 = 0.5$. The corresponding chiral angles are 0 for armchair and 30° for zigzag.

The distance between the walls in multiwalled carbon nanotubes is about 0.34 nm, similar to the distance between graphene layers in graphite.

Single walled carbon nanotubes change their electrical properties depending on $\mathbf{C}_k$. If $n = m$, then the nanotube is metallic and can bear an electrical current density more than 1000 times greater than metals such as silver and copper. Conductance is ballistic; i.e., electrons can pass through the nanotube without heating it. If $n-m=3$, then the nanotube is semiconducting. Otherwise the nanotube is a moderate semiconductor.

The three methods for producing carbon nanotubes are the laser furnace, the carbon arc (i.e., vaporizing graphitic electrodes), and (plasma-enhanced) chemical vapor deposition (Figures 9.4 and 9.5). The latter has turned out to be the best for large-scale production of relatively pure material. Carbon nanotubes are often closed at one or both ends by a hemisphere of fullerene.

Major problems remain with the large-scale utilization of carbon nanotubes. The most severe are:

- making pure preparations;
- dispersing them in solvent (since they can scarcely be solvated (cf. Section 3.2) they tend to be strongly aggregated into bundles);
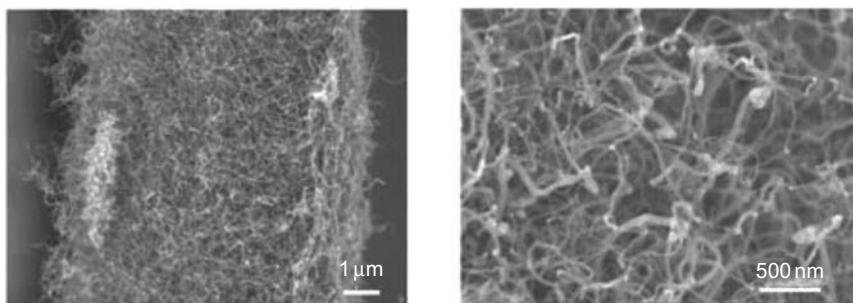
FIGURE 9.4

Scanning electron micrographs of carbon nanotubes grown on the surface of a carbon fiber using thermal chemical vapor deposition. The right-hand image is an enlargement of the surface of the fiber, showing the nanotubes in more detail. Reproduced with permission from [24].
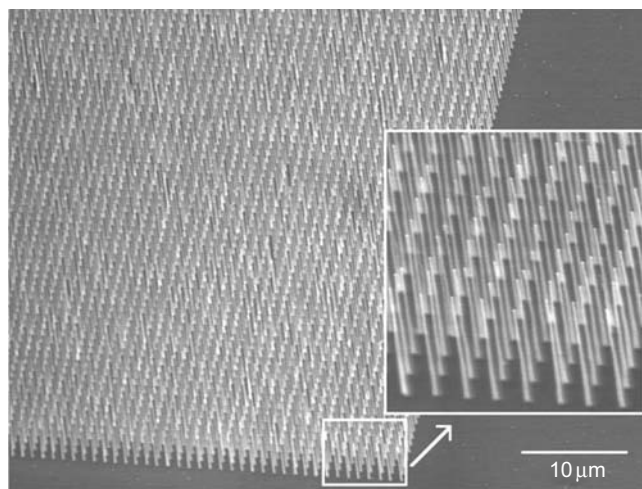
FIGURE 9.5

A forest of carbon nanotubes produced by plasma-enhanced chemical vapor deposition (PECVD). The substratum must first be covered with metal (e.g., Fe or Ni) catalyst islands. Hydrocarbon feedstock (acetylene) is then passed over the substratum heated to several hundred °C. The acetylene decomposes at the surface of the catalyst and the carbon nanotubes grow up from the catalyst particle, or grow up beneath it (pushing it up). Illustration courtesy of Dr Ken Teo, AIXTRON.

- reducing their length (a 20 nm diameter tube may be 20 μm long as fabricated, unnecessary for many applications);
- manipulating them into a desired position.

Postsynthesis processing of nanotube material therefore typically requires:

- Purification—methods include thermal annealing in air or oxygen; acid treatment, microfiltration; these processes typically reduce the mass by around 50%.
- De-agglomeration to separate the tubes. Methods include ultrasonication (but this can damage the tubes), electrostatic plasma treatment, electric field manipulation and polymer wrapping, ball milling (can damage the tubes); these methods can also reduce their length.
- Chemical functionalization (with electron-donating or electron-accepting groups) to improve interactions with a solid or liquid matrix.

As well as chemical functionalization, the lumen of carbon nanotubes may also be filled with different materials. Their use as templates for nanorods or nanotubes has already been mentioned.

Region of π electrons

C_{60} carbon atoms position
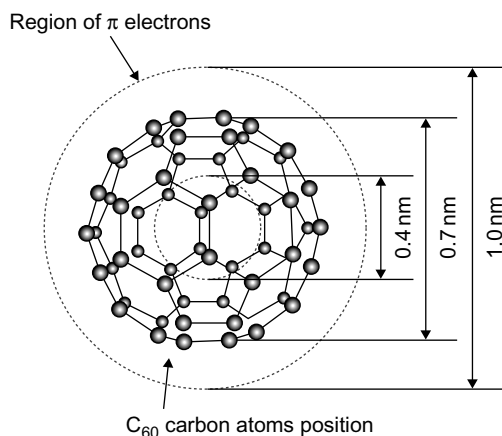
FIGURE 9.6

The structure of buckminsterfullerene, C_{60}. Note the presence of both hexagons and pentagons. Reproduced with permission from [9].

9.3 CARBON NANOPARTICLES (FULLERENES)

Fullerenes (also known as buckyballs) exist as C_{60}, C_{70}, etc. (Figure 9.6). The canonical structure, C_{60}, has iscosahedral symmetry and an electronic structure similar to that of graphene. Fullerenes can be made in a carbon arc, but burning a hydrocarbon feedstock with strict control of the oxygen supply is a more controllable method. The fullerenes can be separated from coproduced soot ("amorphous carbon") by dissolving them out.

The possibility of modifying fullerenes by substitutionally doping the carbon lattice (e.g., with boron), adding an endohedral atom, or functionalizing the surface (e.g., with an amine) greatly expands the ways in which fullerenes can be used [34].

9.4 MATERIALS APPLICATIONS

Carbon Black. This is a relatively old-established product that even today constitutes the overwhelming bulk of the nanotechnology industry (if it is included as part of it, as is done by many market analysts). It consists of carbon particles ranging in size from a few to several hundred nanometers; the bulk of the mass is therefore present in the form of particles bigger than the consensual upper limit of nanoparticles. It is added to the rubber tires for road vehicles as reinforcing filler; the volume of its manufacture is thus directly related to the popularity of such vehicles. Substitution by fullerene (its true nano equivalent), cf. Figure 1.3, would not be an economical

proposition (except perhaps in Formula 1), even if the properties of the tires were significantly enhanced, which seems rather unlikely.

Polymers with Dispersed Carbon Nanotubes. Dispersion of carbon nanotubes in polymer composites may improve their strength, stiffness and thermal and electrical conductivities.

Several processing methods are available. Examples include melt mixing, extrusion, internal mixing, injection molding techniques are all used free of solvents and contamination found in solution processing. There is less fiber breakage than in conventional carbon fibers. High shear mixing is required as increase in viscosity is greater for nanotubes than for nanoparticles. Controlled alignment is possible using spinning extrusion injection moulding. *In situ* polymerization and solution processing have also been used (cf. Section 6.5.1).

The strength improvement depends on the degree of load transfer and on the level of dispersion achieved in the matrix. Additions of nanotubes from 1–5% vol can produce increases in stiffness and strength in polypropylene (PP), polystyrene, polycarbonate and epoxies of up to 50% in stiffness and 20% in strength. If the nanotubes are aligned perpendicular to a crack they can slow down propagation by bridging crack faces. This occurs notably in carbon epoxy composites. In some polymers, nanotubes may decrease the coefficient of friction and decrease the wear rate.

Improvements in electrical properties are dramatic even at very low volume fractions since due to their strong affinity for each other the nanotubes tend to form networks rather than being randomly dispersed. Even without that effect, however, their extreme elongation would enable percolation to be achieved at low vol% additions (cf. Section 3.7). In PP or nylon 1–3 vol% dispersed produce resistivities of 1–100 Ω cm.

Typical current applications for these materials include electrically conducting paint, conducting polymer structures, lighter and stiffer structures (the ultimate application of which would be the space elevator), heat sinks for electronics, motor components, and smart polymer coatings.

9.5 DEVICE COMPONENTS AND DEVICES

Nano Field Emitters. Due to their extremely high curvature, carbon nanotubes can emit electrons at much lower voltages (a few volts) compared with conventional field emission devices. The principle envisaged application is in flat display screens, competing with liquid crystal technology. They are also attractive as electron guns for scanning electron microscopes, in high-power microwave amplifiers, and for miniature X-ray sources. Carbon nanotube-based electron guns for electron microscopes are undoubtedly the best available (but the global market is insignificant in terms of the volume of carbon nanotubes required). *In situ* growth on a flat substrate (Figure 9.5) might diminish the manipulation required to create the final device in this application.

Nanocomposite Transparent Electrodes. Display technology is inevitably going to have to change in the near future because of the global dearth of indium, presently

used for doping tin oxide (at a fairly high level, of the order of 10%) to create electrically conducting transparent "indium tin oxide" (ITO) thin films on glass, which are used as the counterelectrode in display devices. Annual consumption is around 800 tonnes, yet the total known reserves are less than 3000 tonnes. As much recycling as possible is carried out, but extraction of the indium is becoming increasingly difficult as devices become smaller and more integrated. Due to the very low percolation threshold of highly elongated objects dispersed in a matrix, carbon nanotube-doped polymers can be made adequately conductive at levels low enough for the material to remain transparent, which should therefore be able to replace current indium tin oxide-based conducting glass technology. Carbon nanotube-doped polymers are therefore attractive for organic photovoltaic solar cells and other optoelectronic devices.

Others. Potentially larger volume applications are constructing electrochemical capacitors, and in very large-scale integrated electronic circuits (e.g., as connectors between components, especially vertical ones ("vias") to connect stacked layers). In some of these other applications, however, notably supercapacitors, existing materials (e.g., the much cheaper carbon black) already offer performance close to the theoretical limit. More promising are logic components based entirely on carbon nano-objects, such as single electron tunneling transistors (Section 7.4.3) made entirely from graphene.

9.6 SUMMARY

Buckminsterfullerene, the single walled carbon nanotube and graphene epitomize, respectively, nanoparticles, nanofibers and nanoplates. The full potential of their remarkable properties, especially electronic ones, in devices will, however, have to await the development of viable nano-object manipulation technologies. Progress in the field of carbon nanomaterials is rapid. One of the latest developments is tough composites incorporating graphene fragments rather than carbon nanotubes—the former not only bond better to the matrix but also intercept cracks more effectively.

9.7 FURTHER READING

K.H. An, Y.H. Lee, Electronic-structure engineering of carbon nanotubes, NANO: Brief Rep. Rev. 1 (2006) 115–138.

M. Arikawa, Fullerenes—an attractive nano carbon material and its production technology, Nanotechnol. Percept. 2 (2006) 121–114.

B.O. Boscovic, Carbon nanotubes and nanofibers, Nanotechnol. Percept. 3 (2007) 141–158.

M. Kumar, Y. Ando, Carbon nanotube synthesis and growth mechanism, Nanotechnol. Percept. 6 (2010) 7–28.

M.A. Rafiee, et al. Fracture and fatigue in graphene nanocomposites, Small 6 (2010) 179–183.

Z.F. Wang, H. Zheng, Q.W. Shi, J. Chen, Emerging nanodevice paradigm: graphene-based electronics for nanoscale computing, ACM J. Emerging Technol. Comput. Syst. 5 (2009) no. 1, article 3.

This page intentionally left blank

Nanosystems and their Design

10

CHAPTER CONTENTS

This chapter comprises rather a miscellany of topics, because nanotechnology is not yet so far advanced as to comprise systems to any meaningful extent. Key elements in developing a viable industrial system are explored.

10.1 SYSTEMS

The essence of a system is that it cannot be usefully (for the purposes of analyzing its function or optimizing its design) decomposed into its constituent parts. Two or more entities (or activities) constitute a system if the following four conditions, enumerated by R.L. Ackoff, are satisfied:

1. We can talk meaningfully of the behavior of the whole of which they are the only parts.
2. The behavior of each part can affect the behavior of the whole.
3. The way each part behaves and the way its behavior affects the whole depends on the behavior of at least one other part.
4. No matter how we subgroup the parts, the behavior of each subgroup will affect the whole and depends on the behavior of at least one other subgroup.

The word "nanosystem" can be defined as a system whose components are in the nanoscale.

An example of a system that might justifiably be called "nano" is the foot of the reptile called the gecko, many species of which can run up vertical walls and upside down across ceilings. Their feet are hierarchically divided into tens of thousands of minute pads that allow a large area of conformal contact with irregular surfaces. The adhesive force is provided by the Lifshitz–van der Waals interaction (see Section 3.2), normally considered to be weak and short range, but additive and hence sufficiently strong in this embodiment if there are enough points of contact. Attempts to mimic the foot with a synthetic nanostructure have only had very limited success, because the real foot is living and constantly adjusted to maintain the close range conformal contact needed for the interaction to be sufficiently strong to bear the weight of the creature, whereas the synthetic foot is static, and typically irreversibly damaged merely upon detachment. Each nanoscale footlet (the smallest subdivision, the part that is actually in contact with the surface) is only part of a system in the living creature, whose brain is involved in maintaining adhesion.

10.2 MATERIALS SELECTION

When confronted with a macroscale design problem, one may use charts of the type shown in Figure 10.1 to select a suitable material fulfilling the functional requirements. Complex problems may impose more than two constraints on properties, hence many such diagrams may be required, since it is visually problematical to construct them in more than two dimensions. However, this procedure imposes some decomposability on the problem, which is contrary to the spirit of it being a system (Section 10.1). Furthermore, the decomposition is arbitrary, there being numerous possible binary combinations of properties, which may influence the design choices in different ways. Finally, we notice that the entire property space is by no means comprehensively covered. Light and strong, or heavy and weak materials remain elusive, for example.

The engineer working in the nanoscale has, in principle, far vaster possibilities available. These will only be fully realized with bottom-to-bottom nanofacture (Section 8.3), but even nanocomposites (Section 6.5) greatly expand the typically small region occupied by classical composites in diagrams like Figure 10.1.

As components become very small, the statistical nature inherent in present-day "top–down" methods (including the fabrication of a composite by blending nanoparticles into a polymer matrix) becomes more apparent. Ultraprecision engineering mostly achieves its high machining accuracy by rigorous control of the distributions of control errors making additive contributions to the final result, but ultimate nanoprecision is only achievable within a eutactic environment, via bottom-to-bottom nanofacture.
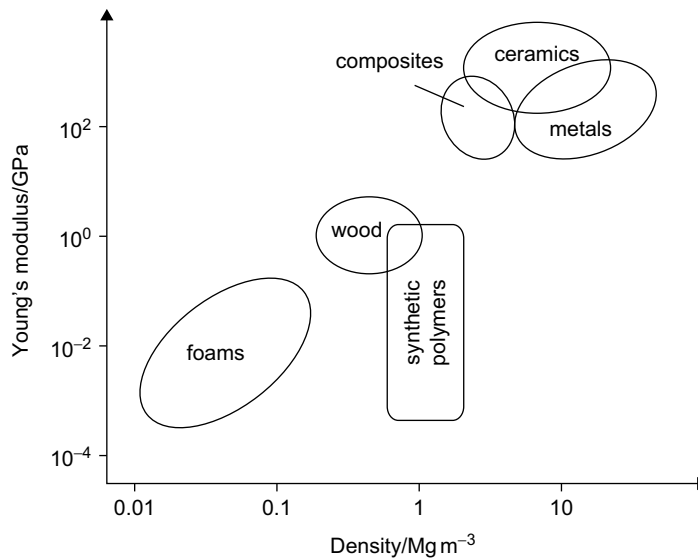
FIGURE 10.1

An example of a materials selection chart used in classical engineering design at the macroscale.

Similar arguments apply to semiconductor processing. Standard semiconductor devices depend on doping (i.e., the introduction of impurities or defects into the pure, perfect semiconductor) for setting the controlling framework for the mobile electron population. The doping level is typically quite low, hence as working volumes shrink a statistical distribution of dopants is no longer adequate (Section 10.3). Again, the solution is assembly within a eutactic environment, in which each dopant can be placed in a position precisely defined with respect to atomic coordinates.

Charts like Figure 10.1 really only deal with quantities that are scalars; that is, they can be quantified on a straightforward linear or logarithmic scale, such as density. Anisotropic materials with direction-dependent moduli, electrical conductivities, etc. require separate charts according to the orientation of the material. Esthetic qualities are not even considered because of their complexity. Even color requires three dimensions; other attributes such as "feel" are more difficult still; indeed there is still no standard way of representing them. Yet the history of technology starting from the earliest known man-made artifacts shows that purely decorative objects designed to be appreciated esthetically have always preceded objects of utility, including weapons and tools, and even entities of a strongly utilitarian nature (such as a motor car) are often chosen on the basis of esthetic attributes, provided the minimum requirements of the utility function are met. This must be borne in mind by the designer working in the nanoworld.

10.3 **DEFECTS IN NANOGRAINS**

The simplest kinds of grains (nanoparticles) are made from a single material, e.g. germanium or silicon. Generally the only issues are whether the surface has a different chemical composition, either due to the adsorption of impurities from the environment or due to the generation of lattice defects at the surface. In the bulk, the two kinds of defects are vacancies and interstitial ions.

In compound materials, formed from a cation C and an anion A, the occurrence of defects is governed by the constraint of the need to maintain electrical neutrality; i.e., vacancies must be compensated by interstitial ions of the same sign (Frenkel type) or by vacancies of the opposite sign (Schottky type). More specifically, they may be cations on interstitial sites and vacancies in the cation half-lattice (Frenkel) or anions on interstitial sites and vacancies in the anion half-lattice (anti-Frenkel); cations and anions on interstitial sites (anti-Schottky) or vacancies in the cation and anion half-lattices (Schottky). Compounds can be created deliberately non-stoichiometric; e.g., by growth of a granular metal oxide film via physical vapor deposition (PVD) in an oxygen-poor atmosphere, thereby assuring an abundance of defects.

Notation (Schottky). The superscript represents the electrical charge of the defect, relative to the undisturbed, overall electrically neutral lattice: $\prime, \bullet, \times$ represent negative, positive and neutral (zero) excess electrical charge. Subscript $\square$ is a vacancy, subscript $\circ$ is an interstitial. Hence $C_\circ^\bullet$ is an interstitial cation, and $C_\square^\prime$ is a cation vacancy, 'null' signifies the undisturbed lattice, and CA represents the addition (at the surface of the nanoparticle) of a new lattice molecule. A filled circle is used to represent the substitution of a native ion by a foreign one, e.g., $X_\bullet^\bullet(C)$ would be a divalent cation X on a cation lattice site, $Cu_\bullet^\times(Ag)$ would be (mono)valent copper on a silver lattice site in (say) AgCl, etc.

Relations between Defects. The following relations are possible [71]

$$C_\circ^\bullet + C_\square^\prime \rightleftharpoons \text{null} \tag{10.1a}$$

$$A_\circ^\prime + A_\square^\bullet \rightleftharpoons \text{null} \tag{10.1b}$$

$$C_\circ^\bullet + A_\circ^\prime \rightleftharpoons CA \tag{10.1c}$$

The above three equations can be used to derive some more:

$$\text{null} \rightleftharpoons C_\square^\prime + A_\square^\bullet + CA \tag{10.2a}$$

$$C_\circ^\bullet \rightleftharpoons CA + A_\square^\bullet \tag{10.2b}$$

$$A_\circ^\prime \rightleftharpoons CA + C_\square^\prime \tag{10.2c}$$

To each of these there corresponds a mass action law, i.e., for the last three, writing x for the mole fraction, identified by its subscript:

$$x_{C_\square^\prime} x_{A_\square^\bullet} = K \tag{10.3a}$$

$$x_{A_\square^\bullet}/x_{C_\circ^\bullet} = K_1 \tag{10.3b}$$

$$x_{C_\square^\prime}/x_{A_\circ^\prime} = K_2 \tag{10.3c}$$

where the K are the equilibrium constants. The defect concentrations are supposed to be sufficiently small so that the chemical potentials μ depend on the logarithm of x; e.g., for vacancies:

$$\mu_\Box - E_\Box = -RT \ln x_\Box \tag{10.4}$$

where $E_\Box$ is the energy to annihilate one mole of vacancies. Of course the usual electroneutrality condition holds:

$$x_{C_\circ^\bullet} - x_{C'_\Box} = x_{A'_\circ} - x_{A_\Box^\bullet} \tag{10.5}$$

Impurities. This formalism can be powerfully applied to the effect of impurities on defects. Examples:

$$NaCl \rightleftharpoons Na_\bullet^\times(Ag) + AgCl \tag{10.6}$$
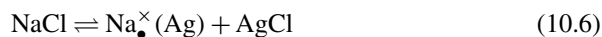
(silver chloride doped with sodium chloride): here there would be practically no change in defect concentration as a result of the doping; but if it is with divalent cadmium, there would evidently be a significant increase in silver vacancies and a decrease in interstitial silver ions:

$$CdCl_2 \rightleftharpoons Cd_\bullet^\bullet(Ag) + Ag'_\Box + 2AgCl \tag{10.7}$$

and

$$Ag_\circ^\bullet + CdCl_2 \rightleftharpoons Cd_\bullet^\bullet(Ag) + 2AgCl \tag{10.8}$$
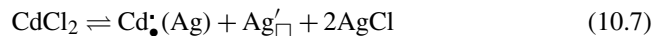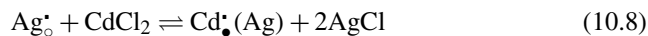
In these mixed phases, the electroneutrality condition becomes

$$x_{C_\circ^\bullet} - x_{C'_\Box} = x_{A'_\circ} - x_{A_\Box^\bullet} + y \tag{10.9}$$

where y is the lattice concentration of the dopant. This relation, together with the three equations (10.3), can be used to derive explicit relations for the concentration dependence of each of the four defect types in equation (10.8) on y.

Surfaces. In very small particles, a significant fraction of the atoms of the particle may be actually surface atoms. Because of their different bonding, the surface atoms may be sources or sinks of defects, thereby perturbing the equilibria (10.1a). The adsorption of impurities at the surface of a nanoparticle will generate defects. For example, adsorption of two monovalent Ag^+ onto the surface of a ZnS nanoparticle will require the formation of a zinc vacancy in the cation sublattice. The environment of the nanoparticle (the matrix) thus plays a crucial role in determining its properties.

10.4 SPACIAL DISTRIBUTION OF DEFECTS

If p is the probability that an atom is substituted by an impurity, or the probability of a defect, then the probability of exactly k impurities or defects among n atoms is

$$b(k;n,p) = \binom{n}{k} p^k q^{n-k} \tag{10.10}$$
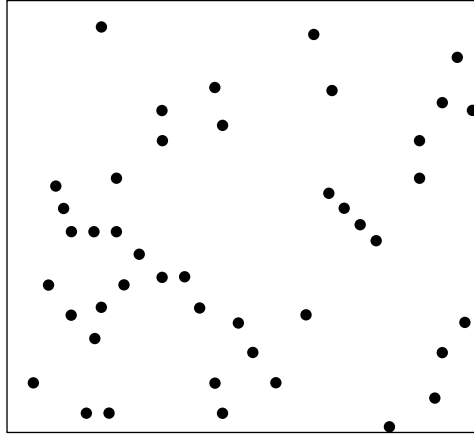
FIGURE 10.2

A two-dimensional piece of matter (white) in which dopants (black) have been distributed at random.

where $q = 1 - p$. If the product $np = \lambda$ is of moderate size (~ 1), the distribution can be simplified to:

$$b(k;n,p) \approx \frac{\lambda^k}{k!} e^{-\lambda} = p(k;\lambda) \tag{10.11}$$

the Poisson approximation to the binomial distribution. Figure 10.2 shows an example. Hence, the smaller the device, the higher the probability that it will be defect-free. The relative advantage of replacing one large device by m devices each $1/m$th of the original size is $m^{1-k}e^{np(1-1/m)}$, assuming that the nanification does not itself introduce new impurities.

10.5 STRATEGIES TO OVERCOME COMPONENT FAILURE

Similar reasoning to that of Section 10.4 can be applied to the failure of individual components (transistors, etc.) on a processor chip, which become "defects". Equation (10.10) can be used to estimate likely numbers of failures, at least as a first approximation, considering them to all occur independently of each other. As the number of components on a "chip" is increased, instead of applying more rigorous manufacturing standards to reduce the probability of occurrence of a defective component, it may become more cost-effective to build in functional redundancy, such that failures of some of the components will not affect the performance of the whole. One strategy for achieving that is to incorporate ways for their failure to be detected by their congeners, who would switch in substitutes. One of the reasons for the remarkable robustness of living systems is the exploitation of functional redundancy, in neural

circuits and elsewhere, although the mechanisms for achieving it are still obscure. The oak tree producing millions of acorns each year, of which a mere handful may survive to maturity, provides another example of functional redundancy.

Standard techniques for achieving reliable overall performance when it is not possible to decrease the number of defects include the NAND multiplexing architecture proposed by John von Neuman [124] and simple approaches reminiscent of overcoming errors in message transmission by repeating the message and instructing the receiver to apply a majority rule. Thus instead of one logic gate, R gates work in parallel and feed their input into a device that selects a majority output. This is called R-fold modular redundancy (RMR), R being an odd number. For even higher reliability the redundancy can be cascaded; thus one might have three groups of three modules feeding into a majority selector, feeding into a further majority selector. A more sophisticated approach is to arrange for the circuit to be reconfigurable [124]. In all these approaches, the higher the probability of a component being defective, the greater the redundancy required to achieve reliable computation.

10.6 **COMPUTATIONAL MODELING**

One of the attractions of nanotechnology is that a nanodevice is small enough for it to be possible to explicitly simulate its operation with atomic resolution (molecular dynamics), using present-day computing resources. This has been done extensively for the Feynman–Drexler diamondoid devices being considered as a way to realize productive nanosystems. Whereas biological nanodevices (e.g., a ribosome) of a similar size are extremely difficult to simulate operationally because of relaxation modes at many different timescales, extending to tens of seconds or longer because the structures are "soft", diamondoid devices such as a gear train typically operate in the GHz frequency domain.

The main atomistic simulation technique is molecular dynamics: the physical system of N atoms is represented by their atomic coordinates, whose trajectories $x_i(t)$ are computed by numerically integrating Newton's second law of motion

$$m_i \mathrm{d}^2 x_i / \mathrm{d}t^2 = \partial V / \partial x_i \tag{10.12}$$

where m_i is the mass of the ith atom and V is the interatomic potential. A general weakness of such atomic simulations is that they use predefined empirical potentials, with parameters adjusted by comparing predictions of the model with available experimental data. It is not possible to give a general guarantee of the validity of this approach. Since no complete diamondoid nanodevices have as yet been constructed, the output of the simulations cannot be verified by comparison with experiment. One way round this difficulty is to develop first principles molecular dynamics simulation code. Density functional theory (DFT), which defines a molecule's energy in terms of its electron density, is a current favorite for calculating the electronic structure of a system.

In many cases a mesoscale simulation may be adequate. For example, if multi-atom nanoblocks are used as units of assembly, it may not be necessary to explicitly simulate each atom within the block.

10.7 "EVOLUTIONARY" DESIGN

Although the most obvious consequence of nanotechnology is the creation of very small objects, an immediate corollary is that in most cases of useful devices, there must be a great many of these objects (vastification). If r is the relative device size, and R the number of devices, then usefulness may require that $rR \sim 1$, implying the need for 10^9 devices. This corresponds to the number of components (with a minimum feature length of about 100 nm) on a very large-scale integrated electronic chip, for example. At present, all these components are explicitly designed and fabricated. But will this still be practicable if the number of components increases by a further two or more orders of magnitude?

Because it is not possible to give a clear affirmative answer to this question, alternative routes to the design and fabrication of such vast numbers are being explored. One of the consequences of vastification is that explicit design of each component, as envisaged in the Feynman–Drexler vision of productive nanosystems, is not practicable. The human brain serves as an inspiration here. Its scale is far vaster than even the largest scale integrated chips: it has $\sim 10^{11}$ neurons, and each neuron has hundreds or thousands of connexions to other neurons. There is insufficient information contained in our genes (considering that each base, chosen from four possibilities, contributes $-0.25 \log_2 0.25$ bits) to specify all these neurons (suppose that each neuron is chosen from two possibilities) and their interconnexions (each one of which needs at least to have its source and destination neurons specified). Rather, our genes cannot do more than specify an algorithm for generating the neurons and their interconnexions.

In this spirit, evolutionary design principles may become essential for designing nanodevices. An example of an evolutionary design algorithm, or simply evolutionary algorithm (EA) for short, is shown in Figure 10.3. It might be initialized by a collection of existing designs, or guesses at possible new designs. Since new variety within the design population is generated randomly, the algorithm effectively expands the imagination of the human designer.

Essential requirements of evolutionary design are, then:

1. Encoding the features of the entity to be designed. Each feature is a "gene", and the ensemble of genes is the "genome".
2. Generating a population of variants of the entity; in the simplest genomes, each possible feature may be present or absent, in which case the genome is simply a binary string; more elaborate ones may allow features to take a range of possible values.
3. Selection. If the selection criterion or criteria can be applied directly to the genome, the process is very rapid, especially if the criterion is a simple threshold
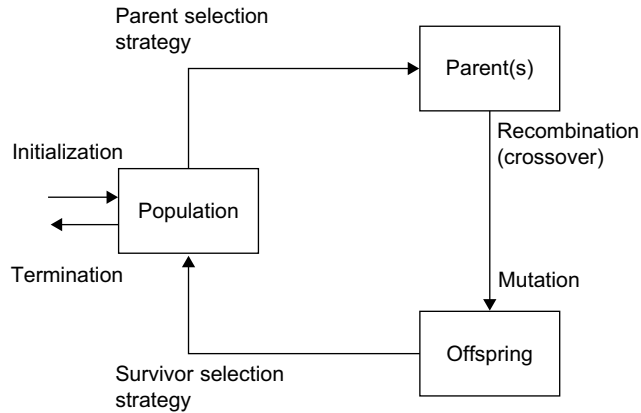
FIGURE 10.3

An evolutionary design algorithm. All relevant design features are encoded in the genome (a very simple genome is for each gene to be a single binary digit indicating absence (0) or presence (1) of a feature). The genomes are evaluated ("survivor selection strategy") —this stage could include human (interactive) as well as automated evaluation—and only genomes fulfilling the evaluation criteria are retained. The diminished population is then expanded in numbers and in variety—typically the successful genomes are used as the basis for generating new ones via biologically inspired processes such as recombination and mutation.

    (e.g., is the ultimate functional requirement met?). It may be, however, that the genome needs to be translated into the entity it encodes, and the selection criterion or criteria applied to the entity itself; an intermediate situation is if the genome can be translated into a model of the entity, which is then subjected to the selection criterion or criteria. Only those genomes fulfilling the criteria are retained; the rest are discarded.

4. Regeneration of variety. Since the selection (3.) implies more or less drastic elimination, the survivors (there must be at least one) beome the parents of the next generation, which may be asexual ("mutations", i.e., randomly changing the values of genes) or sexual (in which two or more parents are selected, according to some strategy, which may be random, and fragments of the parents' genomes, again selected according to some strategy, are combined into the offspring's genome) or both.

5. Return to 3. and continue iterating. If the selection criterion is graded, one can continue until no further increase in quality takes place. It is almost inevitable that human intervention must then make the final choice (unless there is a single clear "winner". Indeed, human intervention may be incorporated at every selection stage, although it makes the process very slow. Typically human intervention takes place after, say, one hundred iterations of automatic selection and regeneration; the human selector is presented with a small number (~ 10) of the best

according to the automatic selection process, and selects a few from those (e.g., [25]). This "interactive evolutionary computation" (IEC) is an excellent way to incorporate expert human knowledge into the design process without having to define and quantify it.

Although the EA design strategy was inspired by biological ("Darwinian") evolution, there is a crucial qualitative difference in that the latter is open ended, whereas the former is generally applied to achieve a preset goal, even though that goal may be specified at a high level. In fact, the main current use of evolutionary algorithms is multi-objective optimization of designs whose main features are already fixed.

The evolutionary paradigm can operate at various levels. "Genetic algorithms" merely allow the parameters of the system's description to evolve (e.g., the coefficients of equations). "Genetic programming" allows more freedom—the algorithm itself may evolve. One may also allow the genome's structure to evolve, and so forth. EAs are generally implemented synchronously (the entire population is simultaneously subjected to the same process), which is of course different from evolution in nature. In a real evolving system, "bad" species may persist, and "good" species may be eliminated for no good reason [5]. These and other biological features [14] will hopefully be implemented in the future as EAs themselves continue to evolve. Ultimately one might reach the ability to create self-organizing functionality.

Although this strategy of evolutionary design enables the design size (i.e., the number of individual features that must be explicitly specified) to be expanded practically without limit, one sacrifices knowledge of the exact internal workings of the system, introducing a level of unpredictability into device performance that may require a new engineering paradigm to be made acceptable. One should, however, bear in mind that even deterministically designed complex systems (e.g., a motor car, ship or airplane), at current levels of technological sophistication, have a vast behavior space and every possible combination of control parameters cannot be explicitly tested. Given that therefore we already in practice sacrifice complete knowledge of the system (even though in principle it is attainable), and yet still have a high degree of confidence in its ability to function safely, it may be unreasonable to object to using an evolutionarily designed artifact.

10.8 PERFORMANCE CRITERIA

The evolutionary algorithm approach demands "fitness" criteria against which the performance of offspring can be judged. Although a system's performance can only be meaningfully measured by considering its ultimate functional output, it may still be helpful to consider component performance. Thus, a logic gate may be assessed according to its rectification ratio and switching speed. A memory cell can be assessed according to its flip energy (between states) and the stability of states. Sensors may be assessed by (in order of increasing preference) gain, signal to noise ratio [73] and detection efficiency [114]. Note that the performance of microelectromechanical systems (MEMS) such as accelerometers is degraded if they are further miniaturized down to the nanoscale [73].

Criteria applied in interactive evolutionary computation are not usually specifiable—if they were, there would be no need to use IEC. It is especially important when considering the esthetic features of a design.

An important criterion applicable to nanobiotechnological devices in medicine (Section 4.2) is resistance to opsonization. Opsonization is that process whereby a foreign object inside a living organism becomes coated with proteins that signal to the immune system that the object must be eliminated. Thus, a key parameter for drug delivery nanoparticles is how long they can circulate in the body. Nanoparticles circulating in a complex fluid such as blood typically attract a protein corona around them, a complex multicomponent composite whose structure is essentially dynamic, continuously varying. Particularly if protein denaturation takes place (Section 4.1.4 and Figure 4.5) the particle is likely to be recognized as foreign.

10.9 SCALEOUT

Traditional process design begins with small-scale laboratory experiments, continues with a pilot scale fabrication, in some factories fortunate to have the capacity it may then be possible to run an experiment on the existing manufacturing plant, before finally a dedicated production unit is designed and put into operation. At each transition to the next scale up, it is scarcely to be expected that control parameters can simply be multiplied by the length or volume (etc.) scaling factor. In complicated processes, the behavior may even change qualitatively. Scaleup is therefore very problematical.

On the other hand, a process implemented at the nanoscale requires only to be multiplied (this is called scaleout). Scaling performance up to the level of human utility is simply a matter of massive parallelization. For example, nanoreactors synthesizing a medicinal drug simply need to work in parallel to generate enough of the compound for a therapeutically useful dose. This introduces new problems of connexions and synchronization, but whereas practically each scaleup problem needs to be solved *de novo* in an *ad hoc* fashion, once the general principles of scaleout are established they are universally valid.

With information processors, the problem is the user interface: a visual display screen must be large enough to show a useful amount of legible information, a keyboard for entering instructions and data must be large enough for human fingers, and so forth—these kinds of problems have already been addressed with microsystems technology (e.g., for reading information stored on microfilm).

10.10 STANDARDIZATION

Standardization is the key to any viable system. The richness and robustness of bacterial life is due to a high degree of standardization of their genomes, such that genetic material can be readily exchanged between them. Engineers require a standard

vocabulary in order to work together and create standard specifications for inter-changeable components of industrial systems.

Nanotechnology is no exception. Even though the nanotechnology industry is still in its infancy, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) are jointly preparing a multipart Technical Specification (as a precursor to an International Standard) for the vocabulary of nanotechnology (ISO/TS 80004), some parts of which have already been published. These specify terms and definitions relating to nano-objects and nano-structured materials, carbon nano-objects, the nano/bio interface, nanometrology, nanomanufacturing processes, and so forth.

10.11 CREATIVE DESIGN

New ways of connecting known components together could yield new functions. The Nanotechnology Age may usher in a new era of creative leaps, reminiscent of the great Victorian era of engineering. When Richard Trevithick designed and built, in 1803, the first railway locomotive in the world (a replica of which is displayed in the hall of Telford Central Station in the UK) he had to solve design problems that no one had ever encountered before—and many of his solutions have persisted in essence to this day. The same applies to the bridges of totally new structure conceived by Brunel, the Stephenson brothers, and others. Will we now embark on a new panopticon of creative design comparable to the immense creative energy of that earlier era?

One needs to examine how the novel possibilities of nanotechnology can be fully exploited from the design viewpoint. While structure is familiarly a fixed given that determines function, living cells present hints of function retroacting on structure. Can this principle be extended to the inanimate world? Can the "structure" intermediate be short-circuited in nanotechnology?; that is, can we determine the components needed to assemble a device with specified functional properties directly, without considering what structure it should have? This may be especially relevant when considering structures that are not static, but act dynamically (e.g., an enzyme, cf. Section 11.3). The methods described in Section 10.7 are especially appropriate to such an approach.

10.12 PRODUCEABILITY

Any viable design must be associated with a practicable route to fabrication. The evolutionary design process described in Section 10.7 could readily be extended to encompass manufacturability. A database of existing manufacturing systems could be a starting point for evaluating the fitness of a design, but the design process could be extended to encompass the production system itself.

It is a corollary of the principles enunciated in Section 10.4 that very small features (with characteristic dimensions of a few nm) will have fluctuations of the order

of tens of percent in atom numbers (and, hence, in any properties dependant on those numbers) if mass-produced by a top-down approach. In the clearly foreseeable future the only way to counter such manufacturing variability is to harness self-limiting phenomena such as described in Section 8.2.9; in the longer term eutectic environments mimicking what nature achieves in, for example, protein synthesis should have the capability of providing reliable production.

Similar considerations apply to marketability. Existing domain knowledge might best be incorporated into the evaluation step interactively, using commercial marketing experts, but the design process could itself devise appropriate and novel approaches to creating a market for the product.

10.13 SUMMARY

Systems and nanosystems are defined. Nanotechnology implies many departures from traditional engineering practice. Regarding choice of materials, one should ultimately no longer be constrained to merely select suitable materials; rather, materials with exactly the required properties can be specified and produced. Computation is much more intimately connected with nanotechnology than with larger scale technologies, because explicit and reliable simulation of complete systems is often possible. To cope with vastification, evolutionary computation offers a solution, vastly amplifying human capabilities (it can be considered as an "intelligence amplifier", in the same way that muscular power is amplified by a traction engine, etc.).

With any nanosystem, it needs to be considered how its function can be made useful for human beings. Scaleout—massive parallelization—eliminates the barriers associated with traditional scaleup.

Any viable, integrated technological system requires standardization, first of vocabulary, then of the components used in constructing the artifacts of the technology.

Nanotechnology offers magnificent opportunities for creative design, possibly allowing a new era of innovation to nucleate and grow. Such growth will be driven not only by the possibilities inherent in the technology but also by the fact that a far greater proportion of the population than hitherto should be able to contribute to conventional design and realization.

Manufacturing variability represents a considerable challenge at the nanoscale and feasible, high-volume production routes have yet to be achieved.

10.14 FURTHER READING

D.C. Sayle, et al., Mapping nanostructure: a systematic enumeration of nanomaterials by assembling nano building blocks at crystallographic positions, ACS Nano 2 (2008) 1237–1251.

This page intentionally left blank

Bionanotechnology

11

CHAPTER CONTENTS

Bionanotechnology is defined as the application of biology to nanotechnology (note that biotechnology is the directed use of organisms to make useful products, typically achieved by genetically modifying organisms); that is, the use of biological molecules in nanomaterials, nanoscale devices or nanscale systems. It should be contrasted with nanobiotechnology (Chapter 4); if the bionanotechnology is then applied to human health (nanomedicine or nanobiotechnology), consistency in terminology would demand that we call it bionanobiotechnology.

The discovery of some of the mechanistic details of complicated biological machinery such as the ribosome, which encodes the sequence of nucleic acids as a sequence of amino acids (called "translation" in molecular biology), was happening around the time that Eric Drexler was promoting his assembler-based view of nanotechnology, and these biological machines provided a kind of living proof of principle that elaborate and functionally sophisticated mechanisms could operate at the nanoscale. Some of these biological machines are listed in Table 11.1. There are many others, such as the mechanism that packs viral DNA ultracompactly in the head of bacteriophage viruses.

The machines listed in Table 11.1 are proteins (polypeptides); some are considered to be enzymes (e.g., ATPase). Enzymes (and possibly other machines as well) can also be constructed from RNA, and some known machines, such as the ribosome, are constructed from both polypeptides and RNA. RNA and polypeptides

213

Table 11.1 Examples of biological nanosized machines

Name	Natural function	State of knowledge[a]
Muscle (myosin)	Pulling	C, S, T
Kinesin	Linear motion	C, S, T
Nerve	Information transmission	T
ATPase	Synthesis of ATP from proton e.p.g.[b]	C, S, T
Bacteriorhodopsin	Generation of proton e.p.g. from light	C,T
Transmembrane ion pump	Moving selected ions against an adverse e.p.g.	C,T
Hemoglobin	Oxygen uptake and release	C,T

[a]C, crystal structure determined; S, single-molecule observation of operation; T, theoretical mechanism(s) available.
[b]Electrochemical potential gradient.

are synthesized (naturally or artificially) as linear polymers, most of which can adopt their functional three-dimensional structure via a self-assembly process (Section 8.2.11) that occurs spontaneously (and often reversibly).

Some of these machines show consummate scaling out to the macroscopic realm. Muscle is a good example: although the actin–myosin pair that is the molecular heart of muscular action develops a force of a few piconewtons, by arranging many "molecular muscles" in parallel, large animals such as elephants can sustainably develop kilowatts of power, as humans have known and made use of for millennia.

These natural nanomachines are inspiring in their own right; their existence and the detailed study of their mode of operation have driven efforts to mimic them using artifically designed and constructed systems—this is called bio-inspired nanotechnology or biomimetic nanotechnology. Many structures and especially devices produced in living systems are constituted from biopolymers designed to fit to congeners with exquisite specificity and precise stoicheiometry. One of the challenges of biomimetic nanotechnology is to recreate these attributes with simpler artificial systems—without much success until now. Could one, for example, create a synthetic oxygen carrier working like hemoglobin but with a tenth or fewer the number of atoms? Possibly, although one wonders whether such a "lean" carrier would be as resilient to fluctuations in its working environment.

Returning to our definition of bionanotechnology (the incorporation of biological molecules into nanoartifacts), after recalling the basics of biological structure and biomolecular mechanism, we shall survey three example areas in which biological molecules have been used structurally or incorporated in nanoscale devices: DNA as a self-assembling construction material; biosensors; and biophotonic memory and

logic gates. Although a rather exotic system of a motile bacterium harnessed to push a tiny rotor has been reported [75], the main current areas of nanotechnological significance are biosensors and biophotonics.

11.1 THE STRUCTURAL NATURE OF BIOMOLECULES

Polypeptides (PP) (Proteins). These are linear polymers of amino acids (H_2N–CHR–COOH, where R (bonded to the central C) is a variable side chain ("residue")—there are 20 different natural ones. To polymerize them, water is eliminated between –COOH and H_2N– to form the peptide bond, hence there is a common backbone (linked via the "peptide" bond) with variable side chains—short aliphatic groups, small aromatic groups, carboxylate, amine, hydroxyl, etc. functionalities. Template-directed synthesis with a very high yield is used in nature, with the templates being closely related to genes via the genetic code (triplets of nucleotide bases encode each amino acid). After synthesis (polymerization), they fold, often spontaneously, to a compact structure according to a least-action principle (see Section 8.2.11). Typical natural proteins have $50 \sim 500$ amino acids. Depending on their sequence, they adopt a definite remembered conformation (proteins acting as devices, rather than having a passive structural role, have two or more stable conformations) and can carry out varied functions, ranging from essentially structural or scavenging to enzymes and motors. Some proteins (called glycoproteins) are branched with oligosaccharides attached to certain residues.

Nucleic Acids (NA). These are polymerized from nucleotides (the monomers) each constituted from a sugar, a phosphate group, and a "base" derived from a purine or pyrimidine (aromatic heterocycle). The sugar and phosphate are polymerized by eliminating water to form a linear backbone, with the bases playing the role of the residues in PP. There are 4 natural bases, abbreviated A, C, G, T (in deoxyribonucleic acid, DNA) and A, C, G, U (in ribonucleic acid, RNA). The bases pair preferentially: A with T (or U), via 2 hydrogen bonds, and C with G via 3 hydrogen bonds (complementary base pairing, CBP). Template-directed synthesis with a very high yield is used in nature to create the polymers. The templates are the genes (DNA), and operate according to the principle of CBP. During polymerization RNA spontaneously folds to a definite compact structure according to a least-action principle (see Section 8.2.11), in which base-pairing via hydrogen bonding is equivalent to the potential energy, and loop and hairpin formation is equivalent to the kinetic energy. DNA forms the famous double helix in which genetic information is stably stored in living cells and many viruses.

Polysaccharides (PS) and Oligosaccharides (OS). These are linear or branched polymers of diverse sugar (cyclic oligoalcohol) monomers, linked via water elimination ("condensation") at any of the several hydroxyl groups. The problem of predicting their structure is not yet solved. Polymerization is not templated (i.e., not

under direct genetic control) and there is variability (to a degree that is only poorly characterized) in sequence and length of the polysaccharides found fulfilling the same function in comparable organisms.

11.2 SOME GENERAL CHARACTERISTICS OF BIOLOGICAL MOLECULES

The energy contained in a given system can be divided into two categories: (i) the multitude of microscopic or thermal motions sufficiently characterized by the temperature; and (ii) the (usually small number of) macroscopic, highly correlated motions, whose existence turns the construction into a machine (a device). The total energy contained in the microscopic degrees of freedom may be far larger than those in the macroscopic ones, but nevertheless the microscopic energy can usually be successfully neglected in the analysis of a construction (in informational terms, the macrostates are remembered but the microstates are not).

Biological molecules are constructions. Hence a statistical approach, in which the motions of an immense number of individual particles are subsumed into a few macroscopic parameters such as temperature and pressure, is inadequate. A construction uses only an insignificant fraction of the Gibbs canonical ensemble and hence is essentially out of equilibrium. This is different from thermodynamic nonequilibrium—it arises because the system is being investigated at timescales much shorter than those required for true statistical equilbrium. Such systems exhibit "broken ergodicity", as epitomized by a cup of coffee in a closed room to which cream is added and then stirred. The cream and coffee equilibrate within a few seconds (during which vast amounts of microinformation are generated within the whorled patterns); the cup attains room temperature within tens of minutes; and days may be required for the water in the cup to saturate the air in the room.

Broken ergodicity may be regarded as a generalization of *broken symmetry*, which leads to a new thermodynamic quantity, the *order parameter* ξ whose value is zero in the symmetrical phase. ξ may be thought of as conferring a kind of generalized rigidity on a system, allowing an external force applied at one point to be transferred to another. Some protein molecules demonstrate this very clearly: flash photolysis of oxygenated hemoglobin (causing the oxygen molecule to dissociate from the iron core of the porphyrin (the heme) to which it is bound) causes motion of the iron core of the heme, which results in (much larger) movement at the distant intersubunit contacts, leading ultimately to an overall change in the protein conformation involving hundreds of atoms.

11.3 THE MECHANISM OF BIOLOGICAL MACHINES

Active proteins (i.e., all proteins apart from those fulfilling a purely passive structural or space-filling role) have two or more stable conformations. Unfortunately there is

very little information about these multiple conformations available. Useful as X-ray diffraction is for determining the structure of proteins, it has the disadvantage of usually fixing the protein in just one of these conformations in the crystal that is used to diffract the X-rays. If a fraction of the proteins does happen to be present in one of the other stable conformations, this is usually regarded as "disorder" and any information about it is lost during structural refinement of the diffraction data. However, these multiple conformations are the key to the general mechanism of biological machines, originally formulated by L.A. Blumenfeld to explain enzymatic catalysis [23].

The prototypical example is an enzyme E catalyzing (say) the decomposition of a molecule A−B (called the substrate of the enzyme in the biochemical literature) into products A + B. In this case, work has to be done to break the chemical bond A−B; the principle can equally well be applied to any action where work is done, such as pulling on a spring (as in muscle). The binding and release of oxygen to and from hemoglobin also works on this principle.

The enzyme consists of an active site and the rest of the protein, which may be considered to be much bigger than the active site (Figure 11.1). The "rest" has two stable conformations, E and $\tilde{E}$. The mechanism proceeds in four stages:

1. A complex is formed between the substrate and the enzyme, A−B + E → (A−B)E*. A−B binds to the active site, releasing free energy and resulting in a local conformational change, which creates a strain between the active site and the rest of the protein. Local fast vibrational relaxation takes place on the picosecond timescale, but the active site is no longer in equilibrium with the rest of the molecule and the resulting strain modifies the energy surface on which the enzymatic reaction takes place. The asterisk denotes that the protein is overall in a strained, nonequilibrium state. Strain creation requires energy, but its magnitude must of course be less than the energy of binding.

2. The complex slowly relaxes to a new conformation $\tilde{E}$, releasing the energy to drive the energy-requiring breaking of the A−B bond: (A−B)E* → A$\tilde{E}$B. This is the elementary act of the enzymatic reaction. This conformational relaxation involves making and breaking a multiplicity of weak bonds, but at a slower rate than the reaction being catalyzed.
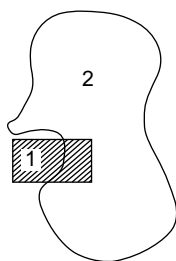
FIGURE 11.1

Model of an enzyme (or motor), consisting of an active site (1) and the rest (2).

3. The product–enzyme complex is decomposed, i.e., the products are released: $A\tilde{E}B \rightarrow A + B + \tilde{E}^*$. Release of the products from the active site again creates strain between it and the rest of the protein molecule.

4. Finally, the strained enzyme slowly relaxes back to its initial conformation: $\tilde{E}^* \rightarrow E$.

An interesting prediction of this mechanism is that the rate of the overall reaction $A-B \rightarrow A + B$ should exhibit an inverse Arrhenius temperature dependence, because increasing the temperature accelerates conformational relaxation (step 2), and hence shortens the time during which the strained molecule is able to accelerate the reaction enzymatically.

11.3.1 **Biological Motors**

Some of the most remarkable devices in nature are not enzymes *per se*—after all, many efficient synthetic catalysts have been made—but miniature motors, which perform prodigious feats. Both linear and rotary motors are known. Figure 11.2 shows the working cycle of the myosin linear motor that powers muscle. Key details have been observed via an experimental approach based on single molecule manipulation. For example, myosin is immobilized on the substratum and allowed to interact with actin tethered to beads held in optical traps (e.g., [171, 58, 125]). It was long

FIGURE 11.2

Simplistic model of the working cycle of muscle. M denotes myosin and A denotes actin; (A) is actin weakly bound to myosin; the asterisk denotes myosin in a strained, nonequilibrium state. Binding of adenosine triphosphate (ATP) to the myosin results in weakened binding to actin, and hydrolysis of the ATP to adenosine diphosphate (ADP) and the subsequent release of phosphate generate strain between the ATP-binding site and the rest of the myosin molecule; the relaxation of this strain drives the movement of the motor (ΔX, 8–10 nm) during which mechanical work is done. The hydrolysis of one ATP molecule yields a chemical energy E of 5–20 k_BT; the force $E/\Delta X$ exerted by one myosin motor can therefore be estimated as 2–10 pN.

assumed that there was a direct correlation between the hydrolysis of ATP and the mechanical work performed by the motor; that is, each hydrolyzed molecule resulted in one unit of displacement ΔX ("tight coupling model"). This is, however, rather naive and essentially treats the biomolecule as a miniature newtonian construction. Single molecule experimental observations do not support this assumption. Simultaneous monitoring of individual hydrolysis events (by microscopic observation of fluorescently labeled adenosine) and bead displacements due to the mechanical force exerted on the actin have clearly shown that mechanical force might be generated several hundred milliseconds after release of ADP [86]. Apparently, myosin can store the chemical energy from several ATP hydrolysis events, and may then subsequently perform several mechanical work steps ("loose coupling model"). The diagram in Figure 11.2 is therefore somewhat simplistic. The energy storage demonstrates that the muscle motor follows the Blumenfeld mechanism. Other mechanical motors such as kinesin moving on microtubules also operate on this principle. The general principle of conformational strain between a binding site and the rest of the protein (generated by binding or release of a small molecule) driving (electro)chemical or mechanical work can be expected to be universal in living systems.

11.3.2 Microtubule Assembly and Disassembly

Another example of a process dependent on two conformational substates is the assembly and disassembly of microtubule filaments from the globular protein tubulin. During the normal state of the eukaryotic cell, these filaments pervade the cytoplasm, acting as tracks for kinesin motors transporting molecules and supramolecular complexes. Prior to eukaryotic cell division (mitosis) the duplicated genome is compactified into chromosomes; the nuclear membrane and the microtubule filament network are degraded. The duplicated genome must be separated and its two halves relocated in the two halves of the cell that will become separate cells upon division. How can this be accomplished? Two centrosomes (protein complexes) form asteriated poles at opposite ends of the cell, and microtubules repeatedly elongate at a speed v_g out from them in random directions, followed by catastrophic disassembly leading to abrupt shrinkage with speed $v_s, v_s \gg v_g$. The process continues until a microtubule filament reaches a chromosome, upon which it attaches itself and drags half of it towards the centrosome. The result is each duplicated genome located in separate halves of the cell, after which the rest of the division process takes place.

   The dynamic instability (assembly–disassembly) is characterized by length fluctuations of the order of the mean microtubule length, hinting at a phase transition [78]. Let f_{gs} denote the frequency of switching from growth to shrinkage and f_{sg} the frequency of switching from shrinkage to growth. When $v_g f_{sg} = v_s f_{gs}$ growth switches from unbounded (corresponding to the assembly of the microtubule filament network) to bounded. At this point the average microtubule length $\bar{\ell} = v_g v_s/(v_s f_{gs} - v_g f_{sg})$ diverges. The molecular origin of growth and shrinkage lies in the fact that tubulin monomers can bind to guanosine triphosphate (GTP) and the complex can spontaneously assemble to form filaments. But the GTP slowly hydrolyzes

spontaneously to guanosine diphosphate (GDP), thereby somewhat changing the tubulin conformation such that it prefers to be monomeric. However, the monomers can only be released from the end; disassembly can be initiated if the rate of GTP hydrolysis exceeds that of tubulin addition for a while. The overall process is a remarkably effective way of searching a restricted volume for an object when no prior information about the location of the object exists.

11.3.3 **The Cost of Control**

The force F that has to be applied to a molecular lever requires accurate knowledge of its position x if reversible work is to be performed [64]. Specifying the positional accuracy as Δx, the uncertainty principle gives the energy requirement as

$$\Delta E \geq hc/(4\Delta x) \tag{11.1}$$

where h is Planck's constant ($= 6.63 \times 10^{-34}$ J s) and c the speed of light in vacuum ($= 3.00 \times 10^8$ m/s). ΔE is obviously negligible for macroscopic systems millimeters in size. The uncertainty in the force $F(x)$ generated at x is

$$\Delta F = F(x) \pm \Delta x(dF/dx). \tag{11.2}$$

To compute the work W done by the system, equation (11.2) is integrated over the appropriate interval. The first term on the right-hand side yields the reversible work W_{rev} and the second term yields $-\Delta x \sum_j |F_j - F_{j+1}|$ for any cycle involving j steps.

The energy conversion factor ϵ is

$$\epsilon = W/(Q + \Delta E) \tag{11.3}$$

where Q is the net energy input during the cycle. With the help of inequality (11.1), the ratio of this to the classical conversion factor $\epsilon_{\text{rev}} = W_{\text{rev}}/Q$ is

$$\epsilon/\epsilon_{\text{rev}} \leq (1 - \alpha/z)/(1 + z) \tag{11.4}$$

where

$$\alpha = hc \sum_j |F_j - F_{j+1}| /(4QW_{\text{rev}}) \tag{11.5}$$

and the relative energy cost of control is

$$z = \Delta E/Q. \tag{11.6}$$

The maximum possible value of the ratio $\epsilon/\epsilon_{\text{rev}}$ is obtained by replacing z by its optimal value z_{opt}, obtained from the turning point of equation (11.4):

$$z_{\text{opt}} = \alpha(1 + \sqrt{1 + 1/\alpha}); \tag{11.7}$$

it is

$$\left(\frac{\epsilon}{\epsilon_{\text{rev}}}\right)_{\text{max}} = \frac{1 - 1/(1 + \sqrt{1 + 1/\alpha})}{1 + \alpha(1 + \sqrt{1 + 1/\alpha})}. \tag{11.8}$$

If more energy than z_{opt} is used, then α decreases because of the energy cost of information; if less, then ϵ decreases because of the irreversibility (dissipation, etc.).

For a macroscopic system these quantities are insignificant. But consider the myosin motor (Figure 11.2): taking $F_j \approx 2$ pN, the displacement $x \approx 10$ nm and $Q \approx 0.067$ aJ (the energy released by hydrolyzing a single ATP molecule), then the energy cost of optimum control, Qz_{opt}, is equivalent to hydrolyzing almost 150 ATP molecules and $(\epsilon/\epsilon_{rev})_{opt} = 0.0033$. Reversible operation is evidently far from optimal; chemical to mechanical conversion occurs at a finite rate that may essentially be uncontrolled, i.e., determined intrinsically. This analysis and conclusion allows the loose coupling model for muscle (Section 11.3.1) to be rationalized.

11.4 DNA AS CONSTRUCTION MATERIAL

The specific base-pairing of DNA, together with the ease of nucleotide polymerization (it can be accomplished artificially using automated equipment) and the relative robustness of the molecule, has engendered interest in the design and construction of artificial nanoscale artifacts of arbitrary shape made from DNA, as Nadrian Seeman was the first to point out. A drawback is that the design of the required DNA strands is a laborious, empirical process (at least at present); but in principle both DNA and RNA could become universal construction materials (provided they are not required to be stable under extreme conditions). The fact that enzymes constructed from RNA are known to exist in nature suggests that ultimately devices could also be made. Once synthesized (according to what are now straightforward, routine procedures—albeit not completely free from errors), it suffices to randomly stir a solution of the components together at an appropriate temperature; their assembly then proceeds in a unique (if the sequences have been correctly designed) fashion (Figure 11.3). This field has recently grown enormously to encompass very elaborate constructions. The process is connected with tile assembly and computation [76].

The specific base-pairing has also been exploited by fastening fragments of DNA to nanoparticles to confer selective affinity upon them, resulting in the formation of
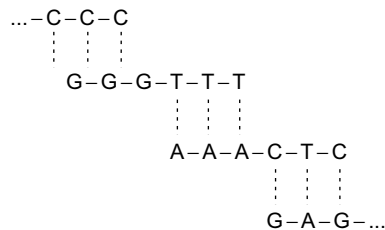
FIGURE 11.3

Four oligonucleotides, which can only assemble in the manner shown. The dashes represent strong covalent bonds, and the dashed lines represent weak hydrogen bonds.

specific clusters of particles, although unless the fastening is geometrically precise, rather than merely statistical, the clusters are also statistical in their composition, except for the simplest ones.

11.5 BIOSENSORS

The dictionary definition of a biosensor is "a device which uses a living organism or biological molecules, especially enzymes or antibodies, to detect the presence of chemicals" [35] (cf. Section 7.9.2). The classic biosensor is the amperometric glucose sensor, comprising glucose oxidase coating an electrode; the enzyme oxidizes glucose. The main reason for wishing to use an enzyme is to exploit the exquisite specificity of biomolecular binding interactions ("molecular recognition").

The Holy Grail of research in the field is to couple the enzyme directly to the electrode such that it can be regenerated by passing electrons to it; in current practice the enzyme concomitantly reduces water to hydrogen peroxide, which is in turn reduced at the electrode, engendering the measured amperometric signal. This is not nanoscale technology, but if the enzyme could indeed be coupled directly to the electrode, this would typically require the active site of the enzyme to be within $\sim$1 nm of the electrode, hence it enters the realm of nanoengineering, in which a carbon nanotube might be used as the electrode, and which opens the way to reducing the size of the device, such that ultimately it might incorporate a single enzyme, able to detect single glucose molecules.

Another kind of biosensor exploits the combinatorial uniqueness of base strings of even fairly modest length to fabricate "gene chips" [33] used to identify genes and genomes. In these devices, the sample to be identified (e.g., the nucleic acids extracted from bacteria found in the bloodstream of a patient) is dispersed over the surface of the chip, which comprises an array of contiguous microzones containing known oligomers of nucleic acids complementary to the sought-for sequences (e.g., a fragment GATTACA is complementary to CTAATGA). Binding can be detected by double helix-specific dyes.

11.6 BIOPHOTONIC DEVICES

Apart from the marvellous intricacy of the biological machinery that converts light into chemical energy, which at present only serves to inspire nanotechnological mimics, there are other, simpler, photoactive proteins, robust enough to be incorporated into artificial devices. Molecules based on the chromophore rhodopsin (such as the primary optical receptor in the eye) seem to have a special place here.

One of the most remarkable of these photoactive proteins is bacteriorhodopsin, which constitutes about a third of the outer membranes of the archaeon (extremophilic prokaryote) *Halobium salinarum*, living in salt lakes. The optically

active site of the protein is the conjugated polyene rhodopsin, and when it absorbs a photon of red light, there is a conformational change generating strain between it and the rest of the protein, which translocates a proton across the membrane (according to the mechanism outlined in Section 11.3). The process is called the bacteriorhodopsin photocycle, and a key intermediate state is called M; the altered interaction between the chromophore and its protein environment gives it an absorption maximum of 410 nm (Figure 11.4).

H. salinarum can be easily grown and the bacteriorhodopsin harvested in the form of "purple membrane fragments"—pieces of outer membrane consisting of an array of bacteriorhodopsin with the membrane lipid filling the interstitial volume. These fragments can be oriented and dried, in which state they can be kept under ambient conditions for 10 years or more without any loss of activity; they have already generated considerable interest as a possible optical storage medium, using a bacteriorhodopsin mutant, the M state of which is almost indefinitely thermally stable. In such an optical memory, the ground state could represent "0" and the M state could represent "1".

Indeed, the overwhelming amount of work in biophotonics has been carried out using the photoactive archaeal protein bacteriorhodopsin (bR). The two main applications are holographic optical memories with ultrahigh data storage density and optical switches. In the former, the biological part is a block of bR and the nonliving part interacting with it is light [68]. Native bacteriorhodopsin can be used to construct an optically switched optical switch (Figure 11.5). Not only can the switch operate extremely rapidly (at megahertz frequencies and above), but only weak light is needed. The remarkable optical nonlinearity of the protein is manifested by exposing it to a single photon! These switches can be used to construct all-optical logic gates [168] and, hence, optical computers.
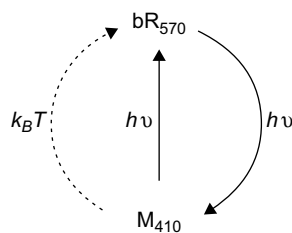
FIGURE 11.4

Simplified view of the bacteriorhodopsin photocycle. A 570 nm photon absorbed by the ground state bR_{570} (the subscript indicates the wavelength of maximum adsorption of the molecule) rapidly (within a few microseconds) transforms (through a series of intermediate stages) the molecule to the relatively stable intermediate M_{410}. This state slowly relaxes thermally back to the ground state, but it can also be rapidly converted by a 410 nm photon. The thermal stability of the M state can be extended almost indefinitely by genetically modifying the protein.
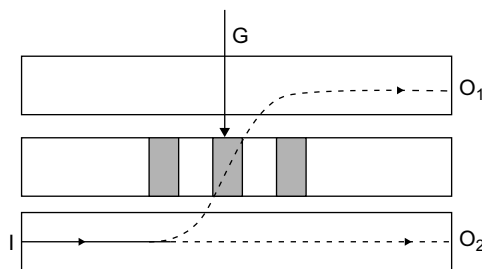
FIGURE 11.5

An optically switched optical switch. The basic construction is a slab of bacteriorhodopsin-containing purple membrane fragments sandwiched between two optical waveguides. Initially, let us suppose that an optical wave introduced at the input I is guided through the structure to emerge at output O_2. If a grating (indicated by the stripes) is holographically generated in the bacteriorhodopsin slab by illuminating with light of 570 nm (from G, the "gate"), light will be coupled out from the lower waveguide and coupled into the upper waveguide, emerging at output O_1. Destroying the grating by illuminating with light of 410 nm will cause the output to revert to O_2.

11.7 SUMMARY

Bionanotechnology is defined as the incorporation of biological molecules into nanoartifacts. The highly refined molecular binding specificity is particularly valued, and used to facilitate the assembly of unique structures from a solution of precursors and for capturing chemicals from the environment prior to registering their presence via a transducer (biosensors). A further application involves using the widely encountered ability of biomolecules to easily accomplish actions associated with difficult and extreme conditions in the artificial realm, such as the catalysis of many chemical reactions, and optical nonlinearity with single photons, a feature which can be exploited to construct optical computers. One might also mention the kidneys as a marvellous example of biological nanoengineering that functions to extract certain substances from highly dilute solutions, an operation which may become of increasing importance as conventionally processable ores become depleted [161].

11.8 FURTHER READING

F.A. Armstrong, H.A.O. Hill, N.J. Walton, Reactions of electron-transfer proteins at electrodes, Q. Rev. Biophys. 18 (1986) 261–322.

L.A. Blumenfeld, Problems of Biological Physics, Springer, Berlin, 1981.

J. Chen, N. Jonoska, G. Rozenberg (Eds.), Nanotechnology: Science and Computation, Springer, Berlin, 2006.

A. Dér, L. Keszthelyi (Eds.), Bioelectronic Applications of Photochromic Pigments, IOS Press, Amsterdam, 2001.

F. Scheller, F. Schubert, Biosensoren, Akademie-Verlag, Berlin, 1989.

J.-M. Valleton, Information processing in biomolecule-based biomimetic systems. From macroscopic to nanoscopic scale, React. Polym. 12 (1990) 109–131.

M. Yoshida, E. Muneyuki, T. Hisabori, ATP synthase—a marvellous rotary engine of the cell, Nature Rev. Mol. Cell Biol. 2 (2001) 669–677.

This page intentionally left blank

The Impact of Nanotechnology

12

CHAPTER CONTENTS

One might well wonder, after having been confronted by all this marvellous technology and potential technology, whether it has the potential for positively benefiting civilization in contributing to the elevation of society. This final chapter examines the actual and potential impacts of nanotechnology. These introductory paragraphs provide an overview; scientific and technical revolutions are considered, in order to decide whether nanotechnology merits being considered as one; more detailed coverage of individual areas is then given: scientific, technical, economic, environmental, social and finally ethical aspects. There is of course overlap between these areas and the division between them is to be somewhat arbitrary. Technical impacts, in other words applications, are covered in the greatest detail, focusing attention on the "big three" areas—computing, energy and health. Here too there is overlap between them.

It is necessary to consider both what might be called "soft" and "hard" aspects of nanotechnology, in the sense of unexceptionable and controversial, respectively. The former corresponds to the near-term—in some cases what is already realized; the

latter corresponds to the long-term; that is, productive nanosystems, as embodied by the personal nanofactory. As with any long-term extrapolation of technology, there is a considerable element of speculation concerning the latter, especially regarding timescales.

Applications can be considered as both direct and indirect. An example of the former is a nanoparticle that functions as a medicinal drug and can be injected as such directly into the bloodstream of a patient. An example of the latter is an information processor (computer) based on very large-scale integrated chips with individual circuit components in the nanoscale (but the overall size of the device and that of many essential peripheral components is bigger than the nanoscale); the direct application of the nanotechnology is to the realization of the integrated circuit; the many applications of the circuitry count as indirect applications of nanotechnology.

Can nanotechnology help to solve the great and pressing problems of contemporary humanity? Although, if ranked, there might be some debate about the order, most people would include rapid climate change, environmental degradation, energy and other resource depletion, unfavorable demographic trends, insufficiency of food, and nuclear proliferation among the biggest challenges. Can these problems be solved if nanotechnology is the continuation of technological progress, which might ultimately be revolutionary if the quantitative change becomes big enough to rank as qualitative? For example, atom-by-atom assembly of artifacts implies, conversely, that discarded ones can be disassembled according to a similar principle, hence the problem of waste (and concomitant environmental pollution) vanishes. More advanced understanding at the nanoscale should finally allow us to create artificial energy-harvesting systems, mimicking photosynthesis, hence the potential penury of energy disappears. If the manufacture of almost everything becomes localized, the transport of goods (a major contributor to energy consumption and environmental degradation) should dwindle to practically nothing. Localized energy production would have a similar effect, eliminating the need for a vast distribution infrastructure. However, the achievement of this ultimate state of affairs depends on the advent of the personal nanofactory, or something resembling it, which is by no means inevitable. Perhaps the miniature medical robot (often called the nanobot) is somewhat closer to realization—in essence it is simply a more advanced version of the responsive drug delivery nanoparticles that are already being deployed. Would indefatigably circulating nanobots inside our bodies enable our lives to be extended almost indefinitely? And if so, what would be the consequences?

Reports published during the last few years are typically euphoric about nanotechnology and all the benefits it will bring. Many of the examples are, however, of a relatively trivial nature and do not seem to represent sufficient breakthrough novelty to constitute a revolution. Thus, we already have nanostructured textiles that resist staining, self-cleaning glass incorporating nanoparticulate photocatalysts capable of decomposing dirt (Figure 7.24); nanoparticle-based sun creams that effectively filter out ultraviolet light without scattering it and are therefore transparent; even lighter and stronger tennis rackets made with carbon fiber or even carbon nanotube composites; and so forth. None of these developments can be said to be truly revolutionary

in terms of impact on civilization. Indeed, they are rather low-profile; one might not even notice them if they were not pointed out. In contrast, the Industrial Revolution was very visible because of the colossal size of its products: gigantic bridges (e.g., the Forth bridge), gigantic steamships (e.g., the *Great Eastern*) and, most gigantic of all if the entire network is considered as a single machine, the railway. And the steel for these constructions was produced in gigantic works; a modern chemical plant or motor-car factory may cover the area of a medium-sized town. In sharp contrast, the products of nanotechnology are, by definition, very small. Individual assemblers would be invisible to the naked eye. But of course the *products* of the assemblers would be highly visible and pervasive—such as the ultralight strong materials from which our built environment might be constructed.

12.1 TECHNICAL REVOLUTIONS

The anatomically modern human being, *Homo sapiens*, emerged some 200,000 years ago. Over 100,000 years apparently elapsed before his superior mental capacities began to be exploited, as evidenced by sophisticated hunting tools, figurative arts, bodily embellishment (jewelery, etc.) and musical instruments, which have been found in Africa dating from 90,000 years before the Christian era (BCE), although the initial bursts seem to have later died out; about 45,000 years ago one finds similar evidence in Europe and western Asia (the so-called upper Palaeolithic explosion). The first writing (cuneiform script) dates from as recently as around 3000 BCE, and comprehensive written records of the past (i.e., the beginning of history) only emerged around 600 (China)–500 BCE (Greece).

The development of man is marked by technological breakthroughs, especially those concerning materials processing. So important are they that the technologies give their names to the successive epochs of prehistory: the Stone Age (predating the emergence of *Homo sapiens*), so-called because of simple stone implements, for example axes, made by knapping flint; pottery first seems to have been made around 20,000 BCE (but it took another 15,000 years before the potter's wheel was invented), the Bronze Age (starting about 3000 BCE), the Iron Age (starting about 1000 BCE, coincidentally with the emergence of glassmaking), rather than modes of life such as hunting, agriculture (starting about 12,000 BCE), pastoralism (around 5000 BCE), urbanization, etc.

The most significant change in our way of life during the last two or three millennia was probably that brought about by the Industrial Revolution, which began in Britain around the middle of the 18th century and marked the beginning of the Industrial Age; by the middle of the 19th century it was in full swing in Britain and, at first patchily, but later rapidly, elsewhere in Europe and North America. Note that this revolution, unlike its predecessors, was very much production-oriented (in other words, manufacturability was as much a consideration for what was produced as usefulness). This in turn was replaced in the latter half of the 20th century by the still ongoing Information Revolution, which ushered in the Information Age, marked by

the development of unprecedented capabilities in the gathering, storage, retrieval and analysis of information, and heavily dependent upon the high-speed electronic digital computer. The next revolution already appears to be on the horizon, and it is thought that it will be the Nano Revolution.

Within a revolution capabilities grow exponentially—one could even say that the duration of the epoch of such exponential growth temporally defines the revolution. This is sometimes quite difficult to perceive, because an exponential function is linear if examined over a sufficiently small interval, and if the technology (or technological revolution) unfolds over several generations, individual perceptions tend to be strongly biased towards linearity. Nevertheless, empirical examination of available data shows that exponential development is the rule (Ray Kurzweil has collected many examples, and in our present epoch the best demonstration is probably Moore's law), although it does not continue indefinitely, but eventually levels off.

Very often a preceding technological breakthrough provides the key to a successive one. For example, increasing skill and knowledge in working iron was crucial to the success of the steam power and steel that were the hallmarks of the Industrial Revolution, which ultimately developed the capability for mass production of the very large-scale integrated electronic circuits needed for realizing the Information Revolution.

Why do people think that the next technological revolution will be that of nanotechnology? Because once the technology has been mastered, the advantages of making things "at the bottom", as Feynman proposed [56], will be so overwhelming it will rapidly dominate all existing ways of doing things. Once iron-making and iron-working had been mastered, no one would have considered making large, strong objects out of bronze; mechanical excavators now reign supreme on building sites; no one uses a slide rule now that electronic calculators are available, and even domestic appliances such as washing machines are controlled by a microprocessor.

Is it to be expected that information technology will be crucial for the realization of nanotechnology? Very probably yes. As explained in Chapter 10, the design of nanomaterials and systems will be heavily dependent upon computation. Furthermore, nanofacture is scarcely conceivable without computer-enabled automation of (bottom-to-bottom) assembly.

Another consideration is that, the Nano Revolution will consummate the trend of science infiltrating industry that began with the Industrial Revolution. As J.D. Bernal has pointed out, this infiltration can be roughly described in four stages of increasing complexity (Table 12.1). Clearly nanotechnology belongs to Stage 4, at least in its aspirations. Traditional or conventional technologies (as we can label everything that is not nanotechnology) also have Stage 4 as their goal but in most cases are still quite far from realizing it.

Note that Stage 4 also encompasses the cases of purely scientific discoveries (e.g., electricity) being turned to industrial use. Nanotechnology is the consummation of Stage 4; a corollary is that nanotechnology should enable science to be applied at the level of Stage 4 to even those very complicated industries that are associated with the most basic needs of mankind, namely food and health.

Stage	Description	Characteristic feature(s)
Table 12.1 The infiltration of science into industry (after Bernal)		
1	Increasing the scale of traditional industries	Measurement and standardization
2	Some scientific understanding of the processes (mainly acquired through systematic experimentation in accord with the scientific method)	Enables improvements to be made
3	Formulation of an adequate theory (implying full understanding of the processes)	Possibility of completely controlling the processes
4	Complete integration of science and industry, extensive knowledge of the fundamental nature of the processes	Entirely new processes can be devised to achieve desired ends

Consideration of the anticipated impacts of nanotechnology on society needs to be set in the general context of technology impacts. Bernal has pointed out the difficulties that arise from the discrepancy between the primitive needs of man, which are actually extraordinarily complex from the scientific viewpoint (e.g., the biochemistry of food preparation, and the animal psychology involved in hunting and domestication), and the need for understanding to proceed from the simple to the complex: what can be understood rationally must necessarily be simple, at least to begin with. Unfortunately the simplest sciences, astronomy and mechanics, appeared (around 3000 and 400 BCE, respectively) only after the main techniques of human life had already been fixed. As a result, these techniques—encompassing such things as agriculture, cookery, husbandry, metalwork, pottery and textiles—remained almost unchanged for many centuries at least until the early 18th century, largely untouched by the scientific method. Subsequent attempts to improve them "scientifically" have actually led to a mixture of benefits and disbenefits, and rational expectations of the impacts of nanotechnology must be tempered by this past history. The Industrial Revolution led to a focus on machine production rather than direct human needs such as food and health. In this regard a fundamental difference between the Nanotechnology Revolution and the Industrial Revolution is that the former is consumer-oriented, unlike the production orientation of the latter. The ultimate stage of nanotechnology, productive nanosystems, in essence abolishes the difference between consumer- and production-orientation. Nanotechnology has the potential of refocusing the way society satisfies its needs on the more human aspects, which is itself a revolutionary enough departure from what has been going on during the last 300 years to warrant the label Revolution. Furthermore, the Nanotechnology Revolution is supposed to usher in what I.J. Good referred to as the "intelligence explosion", when human intelligence is first surpassed by machine intelligence, which then rapidly spreads throughout the entire universe. This is what Kurzweil calls the singularity—the ultimate revolution.

12.2 **SCIENTIFIC IMPACTS**

Nanotechnology implies scrutinizing the world from the viewpoint of the atom or molecule, while remaining cognizant of structure and process at higher levels. Practically speaking, this should have a huge impact on applied science, in which domain the most pressing problems facing humanity, such as food, energy and other resource security, fall (and which are covered in Section 12.3). Nanotechnology will hopefully give humanity new impetus to the problems in a more rational way, according to which it is first ascertained whether the problem requires new fundamental knowledge for its solution, or whether it "merely" requires the application of existing fundamental knowledge.

Nanotechnology is often considered not in isolation, but as one of a quartet: nanotechnology, biotechnology, information technology and cognitive science (nano-bio-info-cogno, NBIC). Nanotechnology is itself inter-, multi- and trans-disciplinary and associating it with this quartet of emerging technologies further emphasizes that catholicism.

There is already impact on science from nanometrology, the instrumentation of which is useful in fields other than that of nanotechnology itself.

12.3 **TECHNICAL IMPACTS**

This section covers the main anticipated fields of applications. Apart from the "big three" (information technology (IT), health and energy), there are other areas that are also due to benefit from nanotechnology, notably chemical processing (e.g., through better, rationally designed and fabricated catalysts). This will partly be given consideration under the heading Energy. General purpose material manipulation capability will doubtless enable the anticipated future shortages of rare elements required for current technologies to be overcome.

12.3.1 **Information Technologies**

IT applications are often called indirect, since the main impact does not arise directly from the nanoscale features of a very large-scale integrated circuit, but from the way that circuit is used. Nanotechnology is well expressed by the continuing validity of Moore's law, which asserts that the number of components on a computer chip doubles every 18 months. Feature sizes of individual circuit components are already below 100 nm; even if the basic physics of operation of such a nanotransistor is the same as that of its macroscale counterpart, the ability, through miniaturization, of packing a very large number of components on a single chip enables functional novelty.

There is little difference between hard and soft in this case, because there is a continuous drive for miniaturization and whether a nanoscale transistor is made by the current top–down methodology of the semiconductor industry or atom-by-atom

assembly should not affect the function. Other developments, notably quantum computing, that are not specifically associated with nanotechnology but have the same effect of increasing the processing power of a given volume of hardware, will also compete.

This indirect nanotechnology is responsible for the ubiquity of internet servers (and, hence, the World Wide Web) and cellular telephones. The impact of these information processors is above all due to their very high-speed operation, rather than any particular sophistication of the algorithms governing them. Most tasks, ranging from the diagnosis of disease to ubiquitous surveillance, involve pattern recognition, something that our brains can accomplish swiftly and seemingly effortlessly for a while, until fatigue sets in, but which requires huge numbers of logical steps when reduced to a form suitable for a digital processor. Sanguine observers predict that despite the clumsiness of this "automated reasoning", ultimately artificial thinking will surpass that of humans—this is Kurzweil's "singularity". Others predict that it will never happen. To be sure, the singularity is truly revolutionary, but is as much a product of the Information Revolution as of the Nano Revolution, even though the latter provides the essential enabling technology.

Information processing and storage constitutes the most "classical" part of nanotechnology applications, in the sense that it was the most readily imaginable at the time of the Feynman lecture [56]. The physical embodiment of one bit of information could be in principle the presence or absence of a single atom (the main challenge is reading the information thus embodied). The genes of the living world, based on four varieties of deoxyribonucleic acid (DNA), come quite close to this ultimate limit and the reading machinery is also nanosized.

Ever since the invention of writing, man has been storing information but the traditional technologies, whether clay tablets or books, are voluminous, whereas the miniaturization of storage that has already been envisaged creates what is essentially unlimited capacity (including, for example, the "life log"—a quasicontinuous record of events and physiological variables of every human being).

12.3.2 Energy

Nanotechnology has the opportunity to contribute in several ways to the problem of energy, which can be succinctly expressed as the current undersupply of usable energy and the trend for the gap to get worse. The principle near-term technical impacts of nanotechnology will be:

Renewable Energies. There is expected to be direct impact on photovoltaic cells. The main primary obstacle to their widespread deployment is the high cost of conventional photovoltaic cells. Devices incorporating particles (e.g., Grätzel cells) offer potentially much lower fabrication costs. The potential of incorporating further complexity through mimicry of natural photosynthesis, the original inspiration for the Grätzel cell, is not yet exhausted. The main secondary obstacle is that except for a few specialized applications (such as powering air-conditioners in Arabia) the electricity thus generated needs to be stored, hence the interest in simultaneous conversion

and storage (e.g., Figure 7.24). Undoubtedly natural photosynthesis is only possible through an extremely exact arrangement of atoms within the photosystems working within plant cells, and the more precisely artificial light harvesters can be assembled, the more successful they are likely to be.

Fuel Cells. Although the scientific basis of this technology, whereby fuel is converted to electricity directly, was established over 150 years ago by Christian Schönbein, it has been very slow to become established. As with photovoltaic cells, the main primary obstacle is the high cost of fabrication. Nanotechnology is expected to contribute through miniaturization of all components (especially reducing the thickness of the various laminar parts), simultaneously reducing inefficiencies and costs, and through realizing better catalysts for oxygen reduction and fuel oxidation. A particular priority is developing fuel cells able to use feedstocks other than hydrogen.

Energy Storage. The primary means of storing energy is as fuel, but unless photoelectrochemical cells generating fuel from sunlight receive renewed impetus, renewable sources will mainly produce electricity and except for some special cases at least some of the electricity will have to be stored to enable the supply to match demand. Supercapacitors based on carbon nanotubes have attracted interest, but the impact of nanotechnology is likely to be small since using ordinary carbon black has already enabled over 90% of the theoretical maximum charge storage capacity to be achieved, at much lower cost. In any case, the classic photoelectrochemical process generates hydrogen (from water), the storage of which is problematical. Through the design and fabrication of rational storage matrix materials, nanotechnology should be able to contribute to effective storage, although whether this will tip the balance in favor of the hydrogen economy is still questionable.

Energy Efficiency. This heading comprises a very heterogeneous collection of technical impacts. Nanostructured coatings with very low coefficients of friction and extremely good wear resistance will find application in all moving machinery, hence improving its efficiency and reducing the energy required to achieve a given result. Examples include electricity-generating wind turbines and electric motors. For all applications where collateral heat production is not required, nanotechnology-enabled light-emitting diodes can replace incandescent filaments. They can achieve a similar luminous output for much less power than incandescent lamps. The heat produced by the latter may be of value in a domestic context (e.g., contributing to space heating in winter) that is simply wasted in the case of outdoor street lighting (although the esthetic effect is very agreeable). Note that, however, the actual operational efficiency of a device in a given functional context typically represents only a fraction of the overall effectiveness in achieving the intended function. For example, if the intended function of street lighting is to reduce road accidents, there is probably an ergonomic limit to the number of lamps per unit length of street above which energy saving becomes insignificant. Although data are hard to come by, it seems that few amenities have been properly analyzed in this manner. It may well be that a 50% reduction in the number of lamps could be effected without diminution of the functional effect. Such a reduction would be equivalent to a really significant technological advance in the device technology. Miniaturizing computer chips diminishes

the heat dissipated per floating point operation (but, at present, not by as much as the increased number of floating point operations per unit area enabled by the miniaturization, cf. Section 7.5). Devices in which bits are represented as electron spins rather than as electron charges will dissipate practically no heat.

Resource Extraction. Current technologies used to extract metal from ores use vastly more energy than is theoretically required. Nanotechnology can be brought to bear on this problem in many different ways. Biomimicry seems a very attractive route to explore especially since living organisms are extremely good at extracting very dilute raw materials from their environment, operating at room temperature and, seemingly, close to the thermodynamic limit.

Localized Manufacture. Possibly the greatest ultimate contribution of nanotechnology, once the stage of the personal nanofactory has been reached, to energy conservation will be through the great diminution of the need to transport raw materials and finished products around the world. The amount of energy currently consumed by transport in one form or another is something between 30 and 70% of total energy consumption. A reduction by an order of magnitude is perhaps achievable.

The above are realizable to a degree using already-available nanotechnology—the main issue is whether costs are low enough to make them economically viable.

12.3.3 **Health**

Impacts can be summarized under the following headings:

Diagnosis. Many diagnostic technologies should benefit from the contributions of nanotechnology. Superior contrast agents based on nanoparticles have already demonstrated enhancement of tissue imaging. Nanoscale biochemical sensors offer potentially superior performance and less invasiveness, to the extent that implanted devices may be able to continuously sense physiological parameters. More powerful computers able to apply pattern recognition techniques to identify pathological conditions from a multiplicity of indicators provide an indirect diagnostic benefit.

Therapy. The most prominent development has been the creation of functionally rich nanostructured drug packaging, enabling the more effective delivery of awkward molecules to their targets. Implants with controlled drug-eluting capability enabled by nanostructured coatings have also been demonstrated. Indirect therapeutic impacts include more powerful computers accelerating the numerical simulation stage of drug discovery, and nano-enabled microreactors facilitating the affordable production of customized medicines, which depends at least in part on the availability of individual patient genomes.

Surgery. Miniaturized devices are making surgery less and less invasive. The basic surgical tools are unlikely to be miniaturized below the microscale, however, but their deployment may be enhanced by nanoscale features such as ultralow friction coatings and built-in sensors for *in situ* monitoring of physiological parameters. The autonomous robot ("nanobot") is still some way in the future. Superior nanostructured tissue scaffolds will enhance the ability to regenerate tissue for repair purposes.

In medicine, scrutinizing the world from the viewpoint of the atom or molecule amounts to finding the molecular basis of disease, which has been underway ever since biochemistry became established, and which now encompasses all aspects of disease connected with the DNA molecule and its relatives. There can be little doubt about the tremendous advance of knowledge that it represents. It, however, is part of the more general scientific revolution that began in the European universities founded from the 11th century onwards—and which is so gradual and ongoing that it never really constitutes a perceptible revolution. Furthermore, it is always necessary to counterbalance the reductionism implicit in the essentially analytical atomic (or nano) viewpoint by insisting on a synthetic systems approach at the same time. Nanotechnology carried through to productive nanosystems could achieve this, because the tiny artifacts produced by an individual assembler have somehow to be transformed into something macroscopic enough to be serviceable for mankind.

The application of nanotechnology to human health is usually called nano-medicine, and is thus a subset of nanobiotechnology. To recall the dictionary definition, medicine is "the science and art concerned with the cure, alleviation and prevention of disease, and with the restoration and preservation of health, by means of remedial substances and the regulation of diet, habits, etc." Therefore, in order to arrive at a more detailed definition of nanomedicine, one needs to simply ask which parts of nanotechnology (e.g., as represented by Figure 1.1) have a possible medical application.

Much of nanomedicine is concerned with materials. Quantum dots and other nano-objects, surface-functionalized in order to confer the capability of specific binding to selected biological targets, are used for diagnostic purposes in a variety of ways, including imaging and detection of pathogens. Certain nano-objects such as nanosized hollow spheres are used as drug delivery vehicles. Based on the premise that many biological surfaces are structured in the nanoscale, and hence it might be effective to intervene at that scale, nanomaterials are being used in regenerative medicine both as scaffolds for constructing replacement tissues and directly as artificial prostheses. Nanomaterials are also under consideration as high-capacity absorbents for kidney replacement dialysis therapy. Other areas in which materials are important and which can be considered to be sufficiently health-related to be part of nanomedicine include the design and fabrication of bactericidal surfaces (e.g., hospital paints incorporating silver nanoparticles).

The advantages of miniaturized analytical devices for medicine ("labs-on-chips"), especially for low-cost point-of-care devices, are already apparent at the microscale. Further miniaturization would continue some of these advantageous trends, resulting *inter alia* in the need for even smaller sample volumes and even lower power consumption, although some of the problems associated with miniaturization such as unwanted nonspecific binding to the internal surfaces of such devices would be exacerbated. However, whereas microscale point-of-care devices rely on the patient (or his or her physician) to take a sample and introduce it into the device,

miniaturization down to the nanoscale would enable such devices to be implanted in the body and, hence, able to monitor a biomarker continuously. The apotheosis of this trend is represented by the nanobot, a more or less autonomous robotic device operating in large swarms of perhaps as many as 10^9 in the bloodstream.

As well as analysis, synthesis (especially of high-value pharmaceuticals) is also an area where microtechnology is contributing through microscale mixers. This technology is very attractive to the pharmaceutical manufacturing industry. Enough evidence has accumulated for it to be generally recognized that many drugs are typically efficacious against only part of the population. This may partly be due to genetic diversity, and partly to other factors, which have not yet been characterized in molecular detail. In the case of genetic diversity, it is possible that almost all the pharmaceutically relevant DNA sequence variants occur in haplotype blocks, regions of 10,000 to 100,000 nucleotides in which a few sequence variants account for nearly all the variation in the world human population (typically, five or six sequence variants account for nearly all the variation). If a drug that has been found to be efficacious against the majority haplotype variant can be made to be efficacious against the others by small chemical modifications, then the task of the drug developer would not be insuperable.

At present, clinical trials do not generally take account of haplotype variation. Advances in sequencing, which nanotechnology is helping both through the development of nanobiotechnological analytical devices (although it seems that microtechnological "labs-on-chips" may be adequate to fulfil needs) and through more powerful information processing, should make it possible in the fairly near future for haplotype determination to become routine. It is, however, not known whether (and perhaps rather improbable that) small modifications to a drug would adapt its efficacity to other haplotype variants. At any rate, different drugs will certainly be needed to treat different groups of patients suffering from what is clinically considered to be the same disease. Micromixers represent a key step in making custom synthesis of drugs for groups of patients, or even for individual patients, economically viable.

The precise control of the hydrodynamic regime that is attainable (see Section 2.4) typically enables reactions that normally run to what are considered to be quite good yields of 90%, say, to proceed at the microscale without by-products; that is, a yield of 100%. The advantages for the complicated multistep syntheses typically required for pharmaceuticals are inestimable. Added to that advantage is the ease of scaling up the volume of synthesis from a laboratory experiment to that of full-scale industrial production simply by scaleout; that is, multiplication of output by the addition of identical parallel reactors.

It is not, however, obvious that further miniaturization down to the nanoscale will further enhance performance. Similar problems to those afflicting nanoscale analytical labs-on-chips would become significant, especially the unwanted adsorption of reagents to the walls of the mixers and their connecting channels.

Two further aspects of nanomedicine deserve a mention. One is automated diagnosis. It is an inevitable corollary of the proliferation of diagnostic devices enabled

by their miniaturization that the volume of data that needs to be integrated in order to produce a diagnosis becomes overwhelming for a human being. Interestingly, according to a University of Illinois study by Miller and McGuire (quoted by Fabb [49]), about 85% of medical examination questions require only recall of isolated bits of factual information. This suggests that automated diagnosis to at least the level currently attainable by a human physician would be rather readily realizable. The program would presumably be able to determine when the situation was too complicated for its capabilities and would still be able to refer the case to one or more human agents in that case. Indeed, medicine has already become accustomed to depending on heavy computations in the various tomographies that are now routine in large hospitals. Diagnosis is essentially a problem of pattern recognition: an object (in this case, the disease) must be inferred from a collection of features. Although there have already been attempts to ease the work of the physician by encapsulating his or her knowledge in an expert system that makes use of the physician's regular observations, significant progress is anticipated when measurements from numerous implanted biosensors are input to the inference engine. This is an example of indirect nanotechnology: the practical feasibility depends on the availability of extremely powerful processors, based on chips having a very high degree of integration enabled by nanoscale components on the chips.

The second aspect is drug design. In many cases useful targets (enzymes, ion channels, promoter sites on DNA) for therapeutic action are already known through molecular biology work. The key tasks for the designer in current practice are: (i) to generate a collection of molecules that bind to the target; and (ii) to select those that bind negligibly to structurally related but functionally unrelated or antagonistic targets. Task (i) may nowadays be largely accomplished automatically through molecular modeling (trial and error docking), and thus requires high-performance (possibly nano-enabled) computing. Task (ii) is a germane problem, the solution of which is mainly hampered through ignorance of the rival targets. However, the growing importance of scrutinizing biological processes not only from the viewpoint of what is now traditional molecular biology but also from that of the nano-engineer is yielding new insight (a particularly valuable example of such new insight is the characterization of the propensity of biomolecules to bind according to their dehydron density [54]).

Since diet is an important contribution to health, nanotechnology applied to foodstuffs could also be considered to be part of nanomedicine. This is of course a vast area, ranging from nanoscale sensors for discreetly probing food quality to nanoscale food additives (to create "nutriceuticals") to nanoscale field additives such as fertilizers, pesticides, and enhancers of agriculturally advantageous natural symbioses. Since water is ingested by humans in greater quantity than any other substance, it might also be considered as food, making nanotechnologies applied to water purification also part of nanomedicine. Furthermore, creating micronutrients in nanoparticulate form permits substances that are normally considered too insoluble to be of nutritional value to be administered (cf. equation 2.7).

12.4 **COMMERCIAL AND ECONOMIC IMPACTS**

Although if the many nanotechnology market reports are to be believed, the technology sector is growing robustly, this growth will be unsustainable unless issues of standardization and risk are addressed. At present, the nanoscale products (e.g., nanoparticles and nanofibers) supplied by many companies are essentially research grades sold on the basis of *caveat emptor*. Standardization implies robust nanometrology (cf. Chapter 5) in order to be able to verify compliance with specifications; the existence of the specifications in turn implies the possibility of freely tradeable commodities via an exchange. Once this begins to happen on a significant scale, the nanotechnology industry will have reached a certain stage of maturity and will be able to continue to develop more and more sophisticated commodities, such as sensorial materials.

The economic impacts of the technological innovations considered in the previous section are likely to be relatively minor as far as the "soft" implementations of nanotechnology are concerned. There will be many innovations but these will in turn breed new demands and one cannot therefore expect a dramatic change in work–life balance. At least, this has been the experience of the developed world for the last century or so and there is no reason to expect radical change from what are essentially incremental improvements. For example, many of the new nanostructured drug delivery vehicles will lead to therapeutic enhancement, but once the old ailments are cured, new ones are likely to be found requiring further new drugs. In contrast, the "hard" implementation—the personal nanofactory—will lead to very dramatic changes in terms of personal empowerment. The entire system of the joint stock company concentrating resources into large, central, capital-intensive factories will become obsolete.

It is often stated that the driver behind Moore's law is purely economic. This is not, however, self-evident. It presumably means that growth is dominated by "market pull" rather than "technology push", but this raises a number of questions. It is doubtful whether there are intrinsic markets, and even if there were why should the technology advance exponentially? How could the actual doubling time be predicted? Does it depend on the number of people working on the technology? If they are mostly employed by large corporations (such as semiconductor chip manufacturers) the ability of those corporations to pay the salaries of the technologists indeed depends on economic factors, but this would not hold in the case of open source development, whether of software or hardware. Similar questions apply to all technical innovations, which generally grow exponentially. Technology push seems to require fewer assumptions as an explanation than market pull. However, it may be erroneous to say that the dynamics must reflect either technology push or market pull. Both may be in operation; in analogy to the venerable principle of supply and demand, push and pull may also "equilibrate", as illustrated in Figure 12.1.

Further aspects connected with "push" and "pull" are the degree to which technology changes society, and the degree to which society is ready for change. The internet—first e-mail and now social networking sites—has enormously changed the
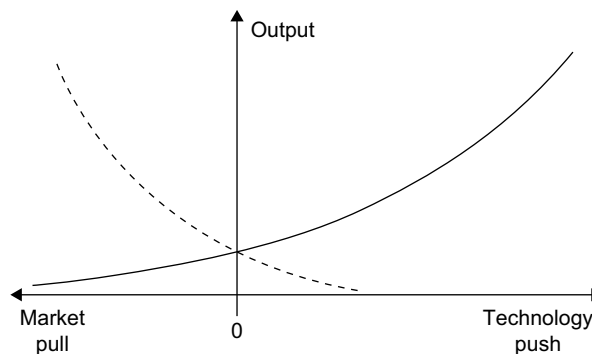
FIGURE 12.1

Proposed quasi-equilibrium between technology push (solid line) and market pull (dashed line). The ideal level of output occurs where they exactly match each other. This diagram neglects consideration of possible temporal mismatch between push and pull.

way people communicate with one another as individuals. It cannot be said that there was a market for these innovations, or for mobile telephony; the market was somehow created. The latter in particular epitomizes many aspects of a sustainable commercial activity; most users are eager to keep up with the latest technological innovations that imply a high degree of personal enthusiasm among the technology developers. Although these technologies are already nano-enabled (through their ultraminiature information processors) the economic model behind them is basically that of the Industrial Revolution. The new communications media are heavily infiltrated by commercial advertising. Internet search engines and social networking sites are not actually floating somewhere in cyberspace but are firmly anchored to strong financial interests and user preferences can easily be linked to advertising via the cellular phone. All this activity has taken on a strong cultural dimension, since for it to be widely acceptable culture has somehow had to become dominated by its commercial and popular side, virtually squeezing out anything else. Although this general trend appears to be fully in the spirit of Adam Smith's "invisible hand"—every member of society acting in his or her best interest and thereby contributing to the general good—there seems to be a growing asymmetry between "producers" (production itself is anyway becoming more and more automated) and "consumers". Whereas previous visions extrapolating the trend of ever more efficient production, which implies ever increasing leisure time, have thought of this leisure time being used for personal cultural enrichment and "working out one's own salvation", presumably because of the preponderance of information technology today personal leisure is also nowadays characterized by "producers" and "consumers"—even though the technology actually enables anyone to be an author, film-maker, journalist or singer, we have not achieved increasing microproduction of culture, with works produced by small audiences, maybe of just a few dozen, with everyone more or less on an

equal footing; these potentially very enriching aspects of nanotechnology have been, at least until now, largely thrown away.

Information technologies have already enabled the "personal nanofactory" of culture to be realized (but little use has been made of this possibility by the vast majority). If and when the real personal nanofactory capable of fabricating virtually any artifact becomes established, will things be different? Open source software demonstrably works and its products seem on the whole to be superior to those produced in closed, "professional" companies. Attempts are already being made to mimic this model through open source hardware. Nanotechnology implies that everyone can be a designer, or at least contribute to design in the case of complex technology. Among other things, this implies the complete obsolescence of patents and intellectual property laws. At present, nanotechnology is at best seen as an enabler of "mass customization", taking a little further existing technologies that allow the purchaser of house paint, and possibly of a T-shirt or a motor car, to specify a unique, individual color. Ultimately, though, everyone *could* be as much a producer as a consumer, of hardware as well as software. This would render the present commercial model obsolete.

12.5 ENVIRONMENTAL IMPACTS

One often hears it stated that nanotechnology will enable the environment to be returned to a pristine state, without explaining the process by which this rather vague assertion might be realized. It seems that there are going to be two principal impacts of nanotechnology on the environment. The first is immediate and direct, the second long-term and indirect. The first is concerned with the use of nanoparticles for environmental remediation. In particular, iron-containing nanoparticles are being promulgated as superior alternatives to existing remediation procedures for soil contaminated with chlorinated hydrocarbons using a more or less comminuted scrap iron. This proposed technology raises a number of questions—to start with there does not seem to be any conclusive evidence that it is actually efficacious, and furthermore there is the still open question of the effect of dispersing a significant concentration of nanoparticles (and it has to be significant, otherwise there would be no significant remediation) on the ecosystem, especially microbial life.

The long-term and indirect effects follow from the obvious corollary of atomically precise technologies—they essentially eliminate waste. This applies not only to the actual fabrication of artifacts for human use, but also to the extraction of elements from the geosphere (should those elements still be necessary, cf. Section 12.3.2). Furthermore, the localized fabrication implied by the widespread deployment of productive nanosystems and personal nanofactories should eliminate almost all of the currently vast land, sea and air traffic involved in wholesale and retail distribution; this elimination (and commensurate downscaling of transport infrastructure) will bring about by far the greatest benefit to the environment.

Nano-engineered "artificial kidneys" can be used not only to extract desirable elements from very dilute sources, such as seawater, but also to extract harmful elements or extractable compounds from natural water polluted with them.

12.6 SOCIAL IMPLICATIONS

The focus in this section is on some of the collective activities of humankind, those that have not been covered in the preceding sections. Clearly the technical advances sketched out in Section 12.3 also have social implications; for example, localized medicine—diagnosis and treatment—will make large central hospitals obsolete, which will be quite a big social change.

If the present commercial model becomes obsolete, what are the political implications? The political arrangements in most countries have co-evolved with commercial developments. If centralized production becomes obsolete, does that imply the same for centralized government?

12.6.1 Regulation

There are already widespread calls for stricter regulation of the deployment of nano-technology, particularly nano-objects in consumer products. These calls are driven by a growing awareness of the potential dangers of nano-objects penetrating into the human body, and by the realization that understanding of this process (cf. Section 4.3) is still rather imperfect, and prediction of the likely effects of any new kind of nano-object is still rather unreliable. Furthermore, there have been a sufficient number of cases, albeit individually on a relatively small scale, of apparently unscrupulous entrepreneurs promoting nano-object-containing products that have turned out to be quite harmful.

These calls, while seemingly reasonable, raise a number of difficulties. One is purely practical: once nano-objects are incorporated into a product they are extraordinarily difficult to trace. Traceability is only feasible up to the point of manufacture, and even then only if the manufacturer has sourced materials through a regulated or self-regulated commercial channel such as a commodity exchange. Establishing the provenance of nanoparticles that might turn up in a waste dump, for example, poses a very difficult forensic challenge.

Furthermore, regulation will become essentially meaningless if productive nano-systems become established: every individual would be producing his or her own artifacts according to his or her own designs and it is hard to see how this could be regulated.

12.6.2 Military Implications

It is a striking fact that the mechanization of war has resulted in a reversion of society's involvement in fighting to an early epoch when every member of a tribe was essentially a warrior. During the last 3000 years, the trend in civilized countries was

for fighting to become a highly specialized activity practiced by a small minority. This changed dramatically in the First World War. Not only was the actual number of fighters a significant proportion of the active working population, but virtually the entire rest of the nation was actively working for the war effort, supporting the fighters. The nature of warfare also changed. Military activity was beginning to become highly mechanized in the First World War, but it still saw battlefields of the kind epitomized by the Battle of Waterloo; that is, places where pitched battles between the professionals took place, in a manner analogous to a game played on a sports field, but in the Second World War the tendency was rather to speak of battle theaters, and that war was marked by the wholesale bombing of great civilian centers, made relatively easy by advances in the design and production of aircraft. More recently, this trend to involve the entire civilian population has continued through urban resistance movements turning civilian cities into battlegrounds.

Against this background, it is highly pertinent to inquire how nanotechnology will affect the social organization implicated in warfare. Clearly many of the technical advances brought about by nanotechnology are useful for military purposes. At present, however, even though the entire population might be involved in a war, it is still only the professionals who are armed; the civilian population might at best have handguns, but they are essentially defenseless. Advances in nanotechnology that are put to belligerent use would appear to mainly serve the professional military services; the performance of munitions can be enhanced or made cheaper. But in the era of the personal nanofactory (productive nanosystems), any individual could design and fabricate offensive weapons, such as small pilotless flying objects. Thus the ultimate impact of military nanotechnology is much greater for the civilian population, who will in principle be able to defend themselves rather effectively against aggression. Nanotechnology represents the apotheosis of mechanization. On the other hand, the universal plenty that the Nano Revolution is supposed to usher in (Section 12.4) should largely remove the raison d'être for warfare in the first place.

12.6.3 Technical Literacy

The real benefits of nanotechnology, which by changing the way we understand technology and science will also impact on the way we understand our relationship with the rest of the universe, can only be realized if there is a significant—that is, at least tenfold—increase in the technical literacy of the general population. Technical literacy means the ability to understand an intelligible account of the technology (i.e. an account in plain language not requiring any prior specialized knowledge). It would be tragic if the technology remains—like so many technologies today—a mysterious, impenetrable black box (cf. Section 12.2).

12.6.4 Education

There are doubtless implications at all levels, starting with the newborn infant and continuing throughout life. The commercial and political reorganization implied by the growth of nanotechnology provides an opportunity to thoroughly rethink the

whole structure of education, especially the formal years of schooling from the ages of six to eighteen. The content of what is currently taught seems very far from what would be required for future designers of nanoartifacts.

During the last century or so, we have become accustomed to highly formalized education in which a growing state bureaucracy has played a large part, not only in building and running schools, but also in devising and managing curricula. Yet, if the technical literacy associated with nanotechnology (Section 12.6.3) develops to a significant degree, the need for such tutelage will wither away, and parents will be capable, and hopefully willing, to once again take on the principal responsibility for their children's education. This is just another example of how the Nanotechnology Revolution will lead to far-reaching social changes that will go way beyond merely devising new educational materials and objects with the help of nanotechnology, while leaving the fundamental aims and processes unchanged. The paradigm exemplified by every citizen designing their own material artifacts for fabrication on their personal nanofactory applies equally well to the possibility of their designing every aspect of their lives.

12.7 IMPACTS ON INDIVIDUAL PSYCHOLOGY

Much of what has already been written in this chapter impacts psychology, both positively and negatively. At present, the negative aspects seem to be slightly dominant. People worry about information overload, which might be quite as harmful as overloading the environment with nano-objects. "Theranostics", therapy combined with diagnostics, as exemplified by ingested nanoparticles that can both diagnose an internal ailment and release a drug to treat it, is seen as disempowerment of the patient (whereas if he or she is prescribed a drug to take as a tablet, it is after all possible to refuse to take it). Nanotechnology may offer a clear solution to solving, say, a nutritional problem, but implementation stalls. Yet the potential of nanotechnology is surely positive, because it offers the opportunity for all to fully participate in society. The answer to the question how one can move more resolutely in that direction would surely be that under the impetus of gradually increasing technical literacy in an era of leisure, in which people are as much producers as consumers, there will be a gradually increasing level of civilization, including a more profound understanding of nature. The latter in particular must inevitably lead to revulsion against actions that destroy nature, and that surely is how the environment will come to be preserved. Elevation of society implies other concomitant advances, such as in the early (preschool) education of infants, which has nothing to do *per se* with nanotechnology, but which will doubtless be of crucial importance in determining whether humanity survives.

12.8 SOME ETHICAL ISSUES

There are quite a few books on the subject of "nanoethics". Is there actually such a subject? The contents of these books mostly deal with matters such as whether

it is right to release nanoparticles into the environment when the effects of inhaling those nanoparticles are imperfectly understood. Clearly that is wrong if (for example) one's choice of ethics is based on Hippocrates' dictum, "Primum, nil nocere." Most of these books seem, however, to overlook the fact that the choice of ethics is ultimately arbitrary. As Herbert Dingle has sagely pointed out, there exists no final sanction for any particular answer to the question, "How shall I choose what to do?"

It follows that books dealing with ethics can be of two types: one type can present a system (possibly with the intention of convincing others that it is worthwhile adopting it). The second type starts with a particular choice of ethical system and examines the consequences for action in a particular area (e.g., nanotechnology). It seems that most of the books dealing with nanoethics are of the second type, but the defect of omitting to clearly state what are the ethics on which they are based seems to be widespread.

Perhaps there is an underlying current of thought that the new era of rationality which might be ushered in by nanotechnology and its way of looking at the universe would lead to an indefeasible and unique set of "natural" ethical principles. Certainly the technology encourages a more consequential attitude: if one knows a structure with atomic precision then its properties should be more predictable than those of conventional or statistically assembled materials; likewise the "atoms" of ethical convictions have certain consequences for human survival, and if they lead to the destruction of humanity those convictions should be discarded. At the same time the use of evolutionary design principles (Section 10.7) means that we shall become surrounded by devices, the mechanisms of whose inner workings may not be known. This will perhaps focus attention on the real features of reality, notably its microdiversity, open-endedness and continuous evolution, which the preponderance of engineering based on fixed, uniform systems and science based on linearity and isolation has tended to eclipse.

It should not be forgotten that morals are essentially stationary, and dependent on the state of intellectual knowledge for their interpretation. Therefore, if there is some agreement on a common set of ethical principles, nanotechnology as part of intellectual knowledge could at least contribute to, and hopefully enhance, their interpretation and effectiveness.

12.9 SUMMARY

It is fairly easy to list likely technical impacts of nanotechnology. The focus on information processing, energy and health implies that these are the areas in which the greatest impact is foreseen. The most important impacts will, however, be on a broader plane, because nanotechnology epitomizes a deeper, more rational view of the universe, including all the practical aspects of our environment. Nevertheless, it is far from certain that humanity will take up the challenge of the new opportunities, especially for everyone to participate in shaping his or her own environment, as much as a producer as a consumer. For this reason, those that hold this vision and see

such a development as the only one offering a way forward for humanity (by which I mean its very survival) have a special responsibility to promote the vision and its realization.

12.10 FURTHER READING

J. Altmann, Military Nanotechnology, Routledge, London, 2006.

J. Baumberg, et al., Where is nano taking us? Nanotechnol. Percept. 3 (2007) 3–14.

J.D. Bernal, The Social Function of Science, Routledge & Kegan Paul, London, 1939.

R.A. Freitas Jr., Economic impact of the personal nanofactory, in: N. Bostrom, et al., Nanotechnology Implications: More Essays, Collegium Basilea, Basel, 2006, pp. 111–126.

S.L. Gillett, Nanotechnology, resources, and pollution control, Nanotechnology 7 (1996) 177–182.

OECD, Nanotechnology: an Overview Based on Indicators and Statistics, STI Working Paper 2009/7.

OECD, The Impacts of Nanotechnology on Companies, 2010.

J.J. Ramsden, Applied Nanotechnology, Elsevier, Amsterdam, 2009.

T.T. Toth-Fejel, A few lesser implications of nanofactories, Nanotechnol. Percept. 5 (2009) 37–59.

Appendix: Nano Neologisms

Any new technology almost inevitably introduces new words. A great many of the terms that have been invented for nanotechnology are simply existing words with "nano" prefixing them. Their meaning is therefore self-evident; a formal definition is only needed in some cases to remove ambiguity. This Appendix describes these and some other new words, given in alphabetical order.

Errification. A change in the parameters of a system that promotes a higher incidence of defective components, or errors in operation.

Eutactic. (adj.) Describes the ideal nano-environment, in which every atom is positioned and oriented as desired. The eutactic environment is an essential precursor for carrying out mechanosynthetic chemistry (cf. Section 8.3).

Micro-Enabled Nanotechnology. A general challenge of nanotechnology is how to scale it up or out in order to provide artifacts usable by human beings. Microtechnology has the potential to act as a key intermediary in this process. For example, nanocatalysts may line the inner tubes of a microchemical reactor [84].

Nanification. To approach the meaning of the word "nanification", think of miniaturization but in a more all-encompassing fashion. To nanify electronics, for example, is not only to make individual components smaller (right down to the nanoscale) but also to adapt all parts of the industry to that situation, including design aspects. In short, nanification means introducing nanotechnology in an integrated rather than a piecemeal fashion. Hence, to nanify manufacture is, ultimately, to introduce molecular manufacturing, which involves not only the actual assembly devices themselves, but also logistics, indeed the entire supply chain, and the (re)organization of the economic system. Nanification of physical systems (designed for human use) implies vastification (q.v.)—many systems are required to work in parallel in order to achieve something of practical use. But in each individual system fluctuations will play an enhanced role (see, e.g., Section 2.4).

Nanobiology. The obvious meaning is the investigation of biological objects at the nanoscale. It is difficult to perceive how this differs from molecular biology. The application of nanotechnology to biological research is called nanobiotechnology (see Chapter 4). We deprecate use of the word "nanobiology" as a synonym for molecular biology. On the other hand, the mindset of molecular biology, which is rooted in biology and chemistry, is very different from that of nanotechnology, which is rooted in engineering and physics. Nanobiology has a valuable meaning describing the scrutiny of biological processes at the molecular level using the formal scientific method associated with physics, namely the mapping of the observed phenomena onto numbers, and the manipulation of those numbers using the rules of mathematics. The relationship of nanobiology to molecular biology is the same as that of Sommerhoff's analytical biology [154] to biology.

Nanoblock. Nanostructured object, typically made by a bottom-to-bottom assembly process, and designed to be capable of self-assembling into a finished object, or being conveniently processed using conventional microscopic or macroscopic techniques.

Nanochemistry. If one allows nanobiology, it seems highly appropriate to also allow this term, defined as chemistry with explicit consideration of processes involving only a few molecules. It could therefore be said to have been founded by Alfred Rényi with his landmark paper on reactions involving very small numbers [147]. Micromixing as studied by chemical engineers, but extended down to the nanoscale, would also belong to nanochemistry.

Nanocriterion. A very useful word, meaning a criterion that can be applied to determine whether something is nanotechnology or not.

Nanofacture. Contraction of nanomanufacture. Linguistic purists prefer this word because the etymology of manufacture implies the involvement of human hands in the making process (however, manufacture is currently used to describe operations taking place in a wholly automated factory).

Nanology. Surprisingly, given its obvious parentage, this word has been rather rarely used so far. It is useful as an umbrella term covering both nanotechnology and nanoscience, and could be used wherever it is awkward to distinguish between them.

Nanonutrition. The exact quantification of all elements and compounds (in principle, by enumerating them) present in food. The word is contrasted with micronutrition, which is defined as the micronutrients (vitamins, trace elements, etc.) which are required for assimilation and transforming the macronutrients (the energy-rich proteins, carbohydrates and lipids), from which the micronutrients cannot be synthesized by our bodies. This term does not yet appear to be in use, but is included in this list to illustrate the essential difference between "micro" and "nano" approaches.

Nanoscience. Is there a need for this term? Sometimes it is defined as "the science underlying nanotechnology". If so, is it not those parts of biology, chemistry and physics that can be grouped under the term of "molecular sciences"? It is the *technology* of designing and making functional objects at the nanoscale that is new; *science* has long been working at this scale and below; for some time in universities and research institutes a movement has existed to unite departments of chemistry and molecular biology, and some parts of physics, into a "department of molecular sciences". No one is arguing that *fundamentally* new physics emerges at the nanoscale; rather, it is the new combinations of phenomena manifesting themselves at that scale that constitute the new technology. The term "nanoscience" therefore appears to be superfluous if used in this sense. A possible meaning could, however, be to denote those parts of molecular sciences that are useful (at any particular epoch) for nanotechnology. Another possible meaning is *the science of mesoscale approximation*, which amounts to roughly the same as virtual nanotechnology (see the legend to Figure 1.1). The description of a protein as a string of amino acids provides a good example: at the mesoscale, one does not need to inquire into details of the internal structure (at the atomic and subatomic levels) of the amino acids. This is perhaps the most useful meaning that can be assigned to the term.

Nanoscope. Device for observing in the nanoscale. The term "microscope" is clearly misleading: an optical microscope cannot resolve features in the nanoscale. Unfortunately the word was (rather unthinkingly but perhaps understandably) applied to new instruments such as the electron microscope and the scanning tunneling microscope

well able to resolve objects in the nanoscale. It may be too late to change this usage, in which case the term "ultramicroscope" might be used instead, a word which is already established in German.

Scaleout. Massive parallelization of a nanoscale production process.

Sensorization. The embedding or incorporation of large numbers of sensors into a structure (which might be something fixed like a bridge or a building, or a structure such as an aircraft, or even a living organism). It is only feasible if the sensors are in, or close to, the nanoscale, from the viewpoints of both cost and space requirements; that is, nanotechnology is an enabling concept. Sensorization is likely to lead to a qualitatively different way of handling situations in at least four areas:

- Structural (civil) engineering: bridges, walls, buildings, etc. Sensors—typically optical fiber Bragg gratings, the technology of which already exists—will be incorporated throughout the structure (e.g., embedded in the concrete, or in the wings of an aircraft). The output of these sensors is indicative of strain, the penetration of moisture, and so forth.
- Process (including chemical) engineering: sensors embedded throughout machinery and reaction vessels will monitor physical (e.g., temperature) and chemical (e.g., the concentration of a selected substance) variables.
- Sensors will be incorporated into the human body, continuously monitoring physiological variables (cf. Section 4.2).
- Sensors will be dispersed throughout the environment (e.g., along rivers and in lakes), reporting the purity of water. The concept is somewhat analogous to what is already taking place in agriculture ("microfarming"; that is, intervention guided by high-resolution satellite images of fields, indicating local moisture, etc.).

In most, or perhaps all, cases where sensorization is envisaged, at present we simply do not have data of the spacial and temporal intensity that will be obtainable. Its availability will almost certainly qualitatively change our views. It will perhaps come to seem primitive to base an assessment of health on a single analysis of key physiological biomarkers. Vehicle health—as appraised by analyzing sensor readouts—will become the criterion for the airworthiness of aircraft (etc.).

Apart from making the miniature sensors themselves, the two main challenges of sensorization are (i) how to deal with the vast proliferation of data (an example of vastification) and (ii) what about the reliability of the sensors? The first challenge can presumably be dealt with by automated processing of the data, and human intervention will only be alerted in the event of some unusual pattern occurring. This implies vast data processing capacity, which is, however, anyway an envisaged development of nanotechnology. The second challenge may be dealt with in the same way: the system will come to be able to determine from the pattern of its readouts when sensors are malfunctioning (cf. Chapter 10).

Ultramicroscope. Synonym for nanoscope; that is, an instrument able to reveal features in the nanoscale.

Vastification. An enormous increase in the number of elementary units or components in a system. Typically, vast numbers of objects are a corollary of their being very small (cf. Moore's law, which is as much about the vast increase in the

number of components fabricated in parallel on a chip as the diminishing size of the components). The main qualitatively distinctive consequence of vastification is that explicit specification and control become impracticable. To tackle this problem, an evolutionary approach will be required, notably in design (see Section 10.3), but very possibly also in operation (e.g., a stigmergic approach). Biomimicry of the human brain may become the favored approach. Vastification usually implies complexification (although typically not found in dictionaries, this term is already used by mathematicians and does not therefore rank as a neologism) and, sometimes, errification (q.v.).

Abbreviations

AES	Auger emission spectroscopy
AFM	atomic force microscope or microscopy
AIC	algorithmic information content
APED	alternating polyelectrolyte deposition
APFIM	atom probe field ion microscopy
APT	atomically precise technologies
BCE	before the Christian era
CMOS	complementary metal–oxide–semiconductor
CNT	carbon nanotubes
CVD	chemical vapor deposition
DFT	density functional theory
DLA	diffusion-limited aggregation
DNA	deoxyribonucleic acid
DPI	dual polarization interferometer
DPN	dip pen nanolithography
EA	evolutionary algorithm
ECM	extracellular matrix
EDS, EDX	energy dispersive spectroscopy
ESEM	environmental scanning electron microscope or microscopy
FET	field effect transistor
FIB	fast atom bombardment
GBD	generalized ballistic deposition
GMR	giant magnetoresistance
IEC	interactive evolutionary computation
IT	information technology
LB	Langmuir–Blodgett
LS	Langmuir–Schaefer
MBE	molecular beam epitaxy
MEMS	micro-electromechanical systems
MMC	metal–matrix composites
MOSFET	metal–oxide–semiconductor field effect transistor
MRAM	magnetic random access memory
MST	microsystems technology
MTJ	magnetic tunnel junction
MWCNT	multiwalled carbon nanotubes
NEMS	nano-electromechanical systems
NSOM	near-field scanning optical microscope or microscopy
OWLI	optical waveguide lightmode interferometry
OWLS	optical waveguide lightmode spectroscopy
PDB	protein data bank
PECVD	plasmon-enhanced chemical vapor deposition

PN	productive nanosystems, personal nanofactory
PSA	programmable self-assembly
PVD	physical vapor deposition
QCA	quantum cellular automaton
QD	quantum dot
RFN	random fiber network
RLA	reaction-limited aggregation
RNA	ribonucleic acid
RRAM	resistive random access memory
RSA	random sequential addition or adsorption
RWG	resonant waveguide grating
SAM	self-assembled monolayer
SAR	scanning angle reflectometry
SEM	scanning electron microscope or microscopy
SET	single electron transistor
SFG	sum frequency generation
SICM	scanning ion conductance microscope or microscopy
SIMS	secondary ion mass spectroscopy
SMEL	sequential minimization of entropy loss
SNOM	scanning near-field optical microscope or microscopy
SPM	scanning probe microscope or microscopy
STEM	scanning transmission electron microscope or microscopy
STM	scanning tunneling microscope or microscopy
SWCNT	single wall carbon nanotubes
TEM	transmission electron microscope or microscopy
TIM	thermal interface materials
UPMT	ultraprecision machine tool
VLSI	very large scale integration

Bibliography

[1] E. Abad, et al., NanoDictionary, Collegium Basilea, Basel, 2005.

[2] N. Aggarwal, et al., Protein adsorption on heterogeneous surfaces, Appl. Phys. Lett. 94 (2009) 083110.

[3] S. Alexandre, C. Lafontaine, J.-M. Valleton, Local surface pressure gradients observed during the transfer of mixed behenic acid/pentadecanoic acid Langmuir films, J. Biol. Phys. Chem. 1 (2001) 21–23.

[4] C. Allain, M. Cloitre, Characterizing the lacunarity of random and deterministic fractal sets, Phys. Rev. A 44 (1991) 3552–3558.

[5] P.M. Allen, W. Ebeling, Evolution and the stochastic description of simple ecosystems, Biosystems 16 (1983) 113–126.

[6] L. Amico, et al., Entanglement in many-body systems, Rev. Mod. Phys. 80 (2008) 517–576.

[7] P.W. Anderson, More is different, Science 177 (1972) 393–396.

[8] A. Aref, et al., Optical monitoring of stem cell-substratum interactions, J. Biomed. Opt. 14 (2009) 010501.

[9] M. Arikawa, Fullerenes—an attractive nano carbon material and its production technology, Nanotechnol. Percept. 2 (2006) 121–114.

[10] R.W. Ashby, Principles of the self-organizing system, in: H. von Foerster, G.W. Zopf (Eds.), Principles of Self-Organization, Pergamon Press, Oxford, 1962, pp. 255–278.

[11] J. Bafaluy, et al., Effect of hydrodynamic interactions on the distribution of adhering Brownian particles, Phys. Rev. Lett. 70 (1993) 623–626.

[12] S. Bandyopadhyay, Single spin devices—perpetuating Moore's law, Nanotechnol. Percept. 3 (2007) 159–163.

[13] L. Banyai, S.W. Koch, Absorption blue shift in laser-excited semiconductor microspheres, Phys. Rev. Lett. 57 (1986) 2722–2724.

[14] W. Banzhaf, et al., From artificial evolution to computational evolution, Nat Rev Genet 7 (2006) 729–735.

[15] A.A. Berezin, Stable isotopes in nanotechnology, Nanotechnol. Percept. 5 (2009) 27–36.

[16] C.R. Berry, Effects of crystal surface on the optical absorption edge of AgBr, Phys. Rev. 153 (1967) 989–992.

[17] C.R. Berry, Structure and optical absorption of AgI microcrystals, Phys. Rev. 161 (1967) 848–851.

[18] W.C. Bigelow, et al., Oleophobic monolayers, J. Colloid Sci. 1 (1946) 513–538.

[19] J.J. Bikerman, The criterion of fracture, SPE Trans. 4 (1964) 290–294.

[20] J.J. Bikerman, Surface energy of solids, Phys. Stat. Sol. 10 (1965) 3–26.

[21] G. Binnig, H. Rohrer, Scanning tunneling microscopy, Helv. Phys. Acta 55 (1982) 726–735.

[22] C. Binns, Prodding the cosmic fabric with nanotechnology, Nanotechnol. Percept. 3 (2007) 97–105.

[23] L.A. Blumenfeld, D.S. Burbajev, R.M. Davydov, Processes of conformational relaxation in enzyme catalysis, in: E.R. Welch, (Ed.), The Fluctuating Enzyme, Wiley, New York, 1986, pp. 369–402.

[24] B.O. Boscovic, Carbon nanotubes and nanofibres, Nanotechnol. Percept. 3 (2007) 141–158.

[25] A.M. Brintrup, et al., Evaluation of sequential, multi-objective, and parallel interactive genetic algorithms for multi-objective optimization problems, J. Biol. Phys. Chem. 6 (2006) 137–146.

[26] R. Bruinsma, Physical aspects of adhesion of leukocytes, in: T. Riste, D. Sherrington (Eds.), Physics of Biomaterials: Fluctuations, Self-Assembly and Evolution, Kluwer, Dordrecht, 1996, pp. 61–101.

[27] L.E. Brus, Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest excited electronic state, J. Chem. Phys. 18 (1984) 4403–4409.

[28] M. Calis, M.P.Y. Desmulliez, Haptic sensing technologies for a novel design methodology in micro/nanotechnology, Nanotechnol. Percept. 1 (2007) 141–158.

[29] C. Calonder, J. Talbot, J.J. Ramsden, Mapping the electron donor/acceptor potentials on protein surfaces, J. Phys. Chem. B 105 (2001) 725–729.

[30] G. Carturan, et al., Inorganic gels for immobilization of biocatalysts: inclusion of invertase-active whole cells of yeast (*Saccharomyces cerevisiae*) into thin layers of SiO_2 gel deposited on glass sheets, J. Molec. Catal. 57 (1989) L13–L16.

[31] G. Catalan, A. Schilling, J.F. Scott, J.M. Gregg, Domains in three-dimensional ferroelectric nanostructures: theory and experiment, J. Phys.: Cond. Matter 19 (2007) 132–201.

[32] R.K. Cavin, V.V. Zhirnov, Generic device abstractions for information processing technologies, Solid State Electron. 50 (2006) 520–526.

[33] S. Chumakov, et al., The theoretical basis of universal identification systems for bacteria and viruses, J. Biol. Phys. Chem. 5 (2005) 121–128.

[34] S.A. Claridge, et al., Cluster-assembled materials, ACS Nano 3 (2009) 244–255.

[35] *Concise Oxford Dictionary*, tenth ed., University Press, Oxford, 1999.

[36] J. Corbett, et al., Nanotechnology: international developments and emerging products, Ann. CIRP 49 (2) (2000) 523–545.

[37] K. Cottier, R. Horvath, Imageless microscopy of surface patterns using optical waveguides, Appl. Phys. B 91 (2008) 319–327.

[38] G. Csúcs, J.J. Ramsden, Interaction of phospholipid vesicles with smooth metal oxide surfaces, Biophys. Biochim. Acta 1369 (1998) 61–70.

[39] A. Dér, et al., Interfacial water structure controls protein conformation, J. Phys. Chem. B 111 (2007) 5344–5350.

[40] R.H. Dennard, et al., Design of ion-implanted MOSFETs with very small physical dimensions, IEEE J. Solid State Circuits SC-9 (1974) 256–268.

[41] R. Dingle, C.H. Henry, Quantum effects in heterostructure lasers, US Patent 3,982,207 (1976).

[42] P.A.M. Dirac, The Principles of Quantum Mechanics, fourth ed., Clarendon Press, Oxford, 1967, (Ch. 1.).

[43] K.E. Drexler, Molecular engineering: an approach to the development of general capabilities for molecular manipulation, Proc. Natl Acad. Sci. USA 78 (1981) 5275–5278.

[44] X. Duan, C.M. Lieber, General synthesis of compound semiconductor nanowires, Adv. Mater. 12 (2000) 298–302.

[45] X. Duan, et al., Single nanowire electrically driven lasers, Nature 421 (2003) 241–245.

[46] S.F. Edwards, R.D.S. Oakeshott, Theory of powders, Physica A 157 (1989) 1080–1090.

[47] A.L. Efros, A.L. Efros, Interband absorption of light in a semiconductor sphere, Soviet Physics of Semiconductors 16 (1982) 772–775.

[48] A.L. Efros, B.I. Shklovskii, Coulomb gap and low-temperature conductivity of disordered systems, J. Phys. C 8 (1975) L49–L51.

[49] W.E. Fabb, Conceptual leaps in family medicine: are there more to come? Asia Pac. Fam. Med. 1 (2002) 67–73.

[50] M. Faraday, Experimental relations of gold (and other metals) to light (Bakerian Lecture), Phil. Trans. R. Soc. 147 (1857) 145–181.

[51] Y. Fang, et al., Resonant waveguide grating biosensor for living cell sensing, Biophys. J. 91 (2006) 1925–1940.

[52] A. Fernández, H. Cendra, In vitro RNA folding: the principle of sequential minimization of entropy loss at work, Biophys. Chem. 58 (1996) 335–339.

[53] A. Fernández, A. Colubri, Microscopic dynamics from a coarsely defined solution to the protein folding problem, J. Math. Phys. 39 (1998) 3167–3187.

[54] A. Fernández, R. Scott, Dehydron: a structurally encoded signal for protein interaction, Biophys. J. 85 (2003) 1914–1928.

[55] A. Fernández, J.J. Ramsden, On adsorption-induced denaturation of folded proteins, J. Biol. Phys. Chem. 1 (2001) 81–84.

[56] R. Feynman, There's plenty of room at the bottom, J. Microelectromech. Syst. 1 (1992) 60–66 (transcript of a talk given by the author on 26 December 1959 at the annual meeting of the American Physical Society at the California Institute of Technology).

[57] H. von Foerster, On self-organizing systems and their environments, in: M.C. Yorvitz, S. Cameron, (Eds.), Self-Organizing Systems, Pergamon Press, Oxford, 1960, pp. 31–50.

[58] T. Funatsu, et al., Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution, Nature 374 (1995) 555–559.

[59] S. Galam, A. Mauger, Universal formulas for percolation thresholds, Phys. Rev. E 53 (1996) 2177–2181.

[60] Y. Gefen, Y. Meir, A. Aharony. Geometric implementation of hypercubic lattices with noninteger dimensionality by use of low lacunarity fractal lattices, Phys. Rev. Lett. 50 (1983) 145–148.

[61] M. Gell-Mann, S. Lloyd, Information measures, effective complexity, and total information, Complexity 2 (1996) 44–52.

[62] J. Gorelik, et al., Non-invasive imaging of stem cells by scanning ion conductance microscopy: future perspective, Tissue Eng. C 14 (2008) 311–318.

[63] T. Graham, On liquid diffusion applied to analysis, J. Chem. Soc. 15 (1862) 216–255.

[64] B.F. Gray, Reversibility and biological machines, Nature (Lond.) 253 (1975) 436–437.

[65] A.Yu. Grosberg, T.T. Nguyen, B.I. Shklovskii, The physics of charge inversion in chemical and biological systems, Rev. Mod. Phys. 74 (2002) 329–345.

[66] M.S. Gudiksen, C.M. Lieber, Diameter-selective synthesis of semiconductor nanowires, J. Am. Chem. Soc. 122 (2000) 8801–8802.

[67] K.C. Hall, et al., Nonmagnetic semiconductor spin transistor, Appl. Phys. Lett. 83 (2004) 2937–2939.

[68] N. Hampp, Bacteriorhodopsin as a photochromic retinal protein for optical memories, Chem. Rev. 100 (2000) 1755–1776.

[69] T. Hashimoto, H. Tanaka, H. Hasegawa. Ordered Structure in Mixtures of a Block Copolymer and Homopolymers. 2. Effects of Molecular Weights of Homopolymers, Macromolecules 23 (1990) 4378–4386.

[70] D. Hanein, et al., Selective interactions of cells with crystal surfaces, J. Cell Sci. 104 (1993) 27–288.

[71] K. Hauffe, Fehlordnungsgleichgewichte in halbleitenden Kristallen vom Standpunkt des Massenwirkungsgesetzes, in: W. Schottky, (Ed.), Halbleiterprobleme, Vieweg, Brunswick, 1954, pp. 107–127.

[72] T.W. Healy, L.R. White, Ionizable surface group models of aqueous interfaces, Adv. Colloid Interface Sci. 9 (1978) 303–345.

[73] C. Hierold, From micro- to nanosystems: mechanical sensors go nano, J. Micromech. Microeng. 14 (2004) S1–S11.

[74] H. Hillman, The Case for New Paradigms in Cell Biology and in Neurobiology, Edwin Mellen Press, Lewiston, 1991.

[75] Y. Hiratsuka, et al., A microrotatory motor powered by bacteria, Proc. Natl Acad. Sci. USA 103 (2006) 13618–13623.

[76] B. Högberg, H. Olin, Programmable self-assembly—unique structures and bond uniqueness, J. Computat. Theor. Nanosci. 3 (2006) 391–397.

[77] T. Hogg, Evaluating microscopic robots for medical diagnosis and treatment, Nanotechnol. Percept. 3 (2007) 63–73.

[78] T.E. Holy, S. Leibler, Dynamic instability of microtubules as an efficient way to search in space, Proc. Natl Acad. Sci. USA 91 (1994) 5682–5685.

[79] R. Horvath, J.J. Ramsden, Quasi-isotropic analysis of anisotropic thin films on optical waveguides, Langmuir 23 (2007) 9330–9334.

[80] R. Horvath, et al., Multidepth screening of living cells using optical waveguides, Biosens. Bioelectron. 24 (2008) 805–810.

[81] R. Horvath, L.R. Lindvold, N.B. Larsen, Reverse-symmetry waveguides: theory and fabrication, Appl. Phys. B 74 (2002) 383–393.

[82] M.A. Hubbe, Adhesion and detachment of biological cells *in vitro*, Prog. Surf. Sci. 11 (1981) 65–138.

[83] H.S. Hwang, Resistance RAM having oxide layer and solid electrolyte layer, and method for operating the same. U.S. Patent Application Pub. No 2010/0002491.

[84] R.K. Iler, Multilayers of colloidal particles, J. Colloid Interface Sci. 21 (1966) 569–594.

[85] A. Iles, Microsystems for the enablement of nanotechnologies, Nanotechnol. Percept. 5 (2009) 121–133.

[86] A. Ishijima, et al., Simultaneous observation of individual ATPase and mechanical events by a single myosin molecule during interaction with actin, Cell 92 (1998) 161–171.

[87] I. Jäger, P. Fratzl, Mineralized collagen fibrils, Biophys. J. 79 (2000) 1737–1746.

[88] M. Jaschke, H.-J. Butt, Deposition of organic material by the tip of a scanning force microscope, Langmuir 11 (1995) 1061–1064.

[89] F.O Jones, K.O. Wood, The melting point of thin aluminium films, Br. J. Appl. Phys. 15 (1964) 185–188.

[90] K. Kendall, The impossibility of comminuting small particles by compression, Nature (Lond.) 272 (1978) 710–711.

[91] H.L. Khor, Y. Kuan, H. Kukula, K. Tamada, W. Knoll, M. Moeller and D.W. Hutmacher, Response of cells on surface-induced nanopatterns: fibroblasts and mesenchymal progenitor cells, Biomacromolecules 8 (2007) 1530–1540.

[92] A.I. Kitaigorodskii, Organic Chemical Crystallography, Consultants Bureau, New York, 1961 (especially Chs 3 and 4).

[93] E. Klavins, Directed self-assembly using graph grammars, in: Foundations of Nanoscience: Self-Assembled Architectures and Devices, Snowbird, Utah, 2004.

[94] E. Klavins, Universal self-replication using graph grammars, in: Int. Conf. on MEMs, NANO and Smart Systems, Banff, Canada, 2004.

[95] E. Klavins, R. Ghrist, D. Lipsky, Graph grammars for self-assembling robotic systems, in: Proc. Int. Conf. Robotics and Automation, New Orleans, 2004, pp. 5293–5300.

[96] W.D. Knight, et al., Electronic shell structure and abundances of sodium clusters, Phys. Rev. Lett. 52 (1984) 2141–2143.

[97] A.N. Kolmogorov, A new invariant for transitive dynamical systems, Dokl. Akad. Nauk SSSR 119 (1958) 861–864.

[98] V. Kondratyev, The Structure of Atoms and Molecules, Dover, New York, 1965.

[99] I. Ichinose, T. Kawakami, T. Kunitake, Alternate molecular layers of metal oxides and hydroxyl polymers prepared by the surface sol-gel process, Adv. Mater. 10 (1998) 535–539.

[100] R. Kurzweil, The Singularity is Near, Viking Press, New York, 2005.

[101] B. Laforge, Emergent properties in biological systems as a result of competition between internal and external dynamics, J. Biol. Phys. Chem. 9 (2009) 5–9.

[102] Ph. Lavalle, et al., Direct observation of postadsorption aggregration of antifreeze glycoproteins on silicates, Langmuir 16 (2000) 5785–5789.

[103] M. Lawo, et al., Simulation techniques for the description of smart structures and sensorial materials, J. Biol. Phys. Chem. 9 (2009) 143–148.

[104] S.C.G. Leeuwenburgh, et al., Morphology of calcium phosphate coatings for biomedical applications deposited using electrostatic spray deposition, Thin Solid Films 503 (2006) 69–78.

[105] P. Lehmann, G. Goch. Comparison of conventional light scattering and speckle techniques concerning an in-process characterization of engineered surfaces, Ann. CIRP 49 (1) (2000) 419–422.

[106] P. Lehmann, Surface roughness measurement based on the intensity correlation function of scattered light under speckle-pattern illumination, Appl. Optics 38 (1999) 1144–1152.

[107] P. Lehmann, S. Patzelt, A. Schöne, Surface roughness measurement by means of polychromatic speckle elongation, Appl. Optics 36 (1997) 2188–2197.

[108] K.K. Likharev, Single-electron devices and their applications, Proc. IEEE 87 (1999) 606–632.

[109] W.A. Lopes, H.M. Jaeger, Hierarchical self-assembly of metal nanostructures on diblock copolymer scaffolds, Nature 414, (2001) 735–738.

[110] P.O. Luthi, et al., A cellular automaton model for neurogenesis in *Drosophila*, Physica D 118 (1998) 151–160.

[111] P.O. Luthi, J.J. Ramsden, B. Chopard, The role of diffusion in irreversible deposition, Phys. Rev. E 55 (1997) 3111–3115.

[112] Yu. Lvov, et al., Molecular film assembly via layer-by-layer adsorption of oppositely charged macromolecules (linear polymer, protein and clay) and concanavalin A and glycogen, Thin Solid Films 284–285 (1996) 797–801.

[113] P.A. McKeown, et al., Ultraprecision machine tools—design principles and developments, Nanotechnol. Percept. 4 (2008) 5–14.

[114] S. Manghani, J.J. Ramsden, The efficiency of chemical detectors, J. Biol. Phys. Chem. 3 (2003) 11–17.

[115] E.K. Mann, et al., Optical characterization of thin films: beyond the uniform layer model, J. Chem. Phys. 105 (1996) 6082–6085.

[116] M. Máté, J.J. Ramsden, Addition of particles of alternating charge, J. Chem. Soc. Faraday Trans. 94 (1998) 2813–2816.

[117] M. Máté, J.J. Ramsden, Colloidal particles efficiently scavenge fatty acid Langmuir–Blodgett films, J. Dispers. Sci. Technol. 19 (1998) 875–883.

[118] M.L. Minsky, Where is our microtechnology? in: The Scientist Speculates, I.J. Good, (Ed.), Heinemann, London, 1962, pp. 139.

[119] A.M. Mood, The distribution theory of runs, Ann. Math. Statist. 11 (1940) 367–392.

[120] M.K. Monastyrov, et al., Electroerosion dispersion-prepared nano- and submicrometre-sized aluminium and alumina powders as power-accumulating substances, Nanotechnol. Percept. 4 (2008) 179–187.

[121] T.L. Morkved, et al., Mesoscopic self-assembly of gold islands on diblock-copolymer films, Appl. Phys. Lett. 64 (1994) 422–424.

[122] N.F. Mott, The electrical conductivity of transition metals, Proc. R. Soc. A 153 (1936) 699–717.

[123] National Nanotechnology Initiative: The Initiative and its Implementation Plan, pp. 19–20. Washington, D.C.: National Science and Technology Council, Committee on Technology, Subcommittee on Nanoscale Science, Engineering and Technology (2000).

[124] K. Nikolić, A. Sadek, M. Forshaw, Architectures for reliable computing with unreliable nanodevices, in: Proc. IEEE-NANO 2001, M2.1 Nano-Devices II, pp. 254–259.

[125] T. Nishizaka, et al., Unbinding force of a single motor molecule of muscle measured using optical tweezers, Nature 377 (1995) 251–254.

[126] S. Oh, et al., Stem cell fate dictated solely by altered nanotube dimension, Proc. Natl Acad. Sci. USA 106 (2009) 2130–2135.

[127] C.J. van Oss, Forces interfaciales en milieux aqueux, Masson, Paris, 1996.

[128] C.J. van Oss, R.F. Giese, Properties of two species of deadly nanoneedles, Nanotechnol. Percept. 5 (2009) 147–150.

[129] C.J. van Oss, et al., Impact of different asbestos species and other mineral particles on pulmonary pathogenesis, Clays Clay Minerals 47 (1999) 697–707.

[130] N. Oyabu, et al., Mechanical vertical manipulation of selected single atoms by soft nanoindentation using near contact atomic force microscopy, Phys. Rev. Lett. 90 (2003).

[131] N. Oyabu, et al., Single atomic contact adhesion and dissipation in dynamic force microscopy, Phys. Rev. Lett. 96 (2006).

[132] T.O. Paine, L.I. Mendelsohn, F.E. Luborsky, Effect of shape anisotropy on the coercive force of elongated single-magnetic-domain iron particles. Phys. Rev. 100 (1955) 1055–1059.

[133] M. Planck, The concept of causality, Proc. Phys. Soc. 44 (1932) 529–539.

[134] A. Politi, J.L. O'Brien, Quantum computation with photons. Nanotechnol. Percept. 4 (2008) 289–294.

[135] J.J. Ramsden, The nucleation and growth of small CdS aggregates by chemical reaction. Surf. Sci. 156 (1985) 1027–1039.

[136] J.J. Ramsden, The stability of superspheres, Proc. R. Soc. Lond. A 413 (1987) 407–414.

[137] J.J. Ramsden, Electron diffraction anomalies in small CdS clusters, J. Cryst. Growth 82 (1987) 569–572.

[138] J.J. Ramsden, Impedance of pore-containing membranes, Stud. Biophys. 130 (1989) 83–86.

[139] J.J. Ramsden, Molecular orientation in lipid bilayers, Phil. Mag. B 79 (1999) 381–386.

[140] J.J. Ramsden, M. Grätzel, Formation and decay of methyl viologen radical cation dimers on the surface of colloidal CdS, Chem. Phys. Lett. 132 (1986) 269–272.

[141] J.J. Ramsden, R. Horvath, Optical biosensors for cell adhesion, J. Recept. Signal Transduct. 29 (2009) 211–223.

[142] J.J. Ramsden, G.I. Bachmanova, A.I. Archakov, Kinetic evidence for protein clustering at a surface, Phys. Rev. E. 50 (1994) 5072–5076.

[143] J.J. Ramsden, Yu.A. Lvov, G. Decher, Optical and X-ray structural monitoring of molecular films assembled via alternate polyion adsorption, Thin Solid Films 254 (1995) 246–251; ibid., 261 (1995) 343–344.

[144] J.J. Ramsden, et al., An optical method for the measurement of number and shape of attached cells in real time, Cytometry 19 (1995) 97–102.

[145] J.J. Ramsden, et al., The design and manufacture of biomedical surfaces, Ann. CIRP 56 (2) (2007) 687–711.

[146] M. Reibold, et al., Carbon nanotubes in an ancient Damascus sabre, Nature 444 (2006) 286.

[147] A. Rényi, Kémiai reakciók tárgyalása a sztochasztikus folyamatok elmélete segítségével, Magy. Tud. Akad. Mat. Kut. Int. Közl. 2 (1953) 83–101.

[148] P.A. Revell, The biological effects of nanoparticles, Nanotechnol. Percept. 2 (2006) 283–298.

[149] G. Schwarz, Cooperative binding to linear biopolymers, Eur. J. Biochem. 12 (1970) 442–453.

[150] U.D. Schwarz, et al., Tip artefacts in scanning force microscopy, J. Microsc. 173 183–197 (1994).

[151] E.K. Schweizer, D.M. Eigler, Positioning single atoms with a scanning tunneling microscope, Nature (Lond.) 344 (1990) 524–526.

[152] O. Sinanoğlu, Microscopic surface tension down to molecular dimensions and micro thermodynamic surface areas of molecules or clusters, J. Chem. Phys. 75 (1981) 463–468.

[153] G.L. Snider, et al., Quantum dot cellular automata: review and recent experiments, J. Appl. Phys. 85 (1999) 4283–4285.

[154] G. Sommerhoff, Analytical Biology, Oxford University Press, London, 1950, pp. 124–126.

[155] J.P. Spatz, A.R. Khokhlov, R.G. Winkler, P. Reineker. Order-disorder transition in surface-induced nanopatterns of block copolymer films, Macromolecules 33 (2000) 150–157.

[156] K. Spratte, L.F. Chi, H. Riegler, Physisorption instabilities during dynamic Langmuir wetting, Europhys. Lett. 25 (1994) 211–217.

[157] Y. Sugimoto, et al., Complex patterning by vertical interchange atom manipulation using atomic force microscopy, Science 322 (2008) 413–417.

[158] S.M. Sze, Physics of Semiconductor Devices, second ed., Appendix G, Wiley, New York, 1981.

[159] N. Taniguchi, On the basic concept of nano-technology, in: Proc. Intl Conf. Prod. Engng, Part II (Jap. Soc. Precision Engng), pp. 18–23. Tokyo, 1974.

[160] A.I. Teixeira, P.F Nealey, C.J. Murphy, Responses of human keratinocytes to micro- and nanostructured substrates, J. Biomed. Mater. Res. 71A (2004) 369–376.

[161] S.R. Thomas, Modelling and simulation of the kidney, J. Biol. Phys. Chem. 5 (2005) 70–83.

[162] K. Tiefenthaler, W. Lukosz, Sensitivity of grating couplers as integrated-optical chemical sensors, J. Opt. Soc. Am. B 6 (1989) 209–220.

[163] K. Ichikawa, et al., Nonvolatile thin film transistor memory with ferritin, J. Korean Phys. Soc. 54 (2009) 554–557.

[164] Ph. Walter, et al., Early use of PbS nanotechnology for an ancient hair dying formula, Nano Lett. 6 (2006) 2215–2219.

[165] V.F. Weisskopf, Quality and quantity in quantum physics, Daedalus 88 (1959) 592–605.

[166] Y. de Wilde, et al., Thermal radiation scanning tunnelling microscopy, Nature (Lond.) 444 (2006) 740–743.

[167] S.A. Wilson, et al., Enhanced dc conductivity of low volume-fraction nano-particulate suspensions in silicone and perfluorinated oils, J. Phys. D 42 (2009) 062003.

[168] E.K. Wolff, A. Dér, All-optical logic, Nanotechnol. Percept. 6 (2010) 51–56.

[169] T. Xu, et al., The influence of molecular weight on nanoporous polymer films, Polymer 42 (2001) 9091–9095.

[170] G.E. Yakubov, et al., Viscous boundary lubrication of hydrophobic surfaces by mucin, Langmuir 25 (2009) 2313–2321.

[171] T. Yanagida, Y. Harada, A. Ishijima, Nanomanipulation of actomyosin molecular motors in vitro: a new working principle, Trends Biochem. Sci. (TIBS) 18 (1993) 319–323.

[172] A.D. Yoffe, Low-dimensional systems: quantum size effects and electronic properties of semiconductor microcrystallites (zero-dimensional systems) and some quasi-two-dimensional systems, Adv. Phys. 42 (1993) 173–266.

[173] R. Young, et al., The Topografiner: an instrument for measuring surface microtopography, Rev. Sci. Instrum. 43 (1972) 999–1011.

[174] V.V. Zhirnov, et al., Limits to binary logic switch scaling—a gedanken model, Proc. IEEE 91 (2003) 1934–1939.

[175] M. Zwolak, M. Di Ventra, Physical approaches to DNA sequencing and detection, Rev. Mod. Phys. 80 (2008) 141–165.

Index