CHAPTER

1

Molecular Biology for Computer Scientists

Lawrence Hunter

"Computers are to biology what mathematics is to physics."

— Harold Morowitz

One of the major challenges for computer scientists who wish to work in the domain of molecular biology is becoming conversant with the daunting intricacies of existing biological knowledge and its extensive technical vocabulary. Questions about the origin, function, and structure of living systems have been pursued by nearly all cultures throughout history, and the work of the last two generations has been particularly fruitful. The knowledge of living systems resulting from this research is far too detailed and complex for any one human to comprehend. An entire scientific career can be based in the study of a single biomolecule. Nevertheless, in the following pages, I attempt to provide enough background for a computer scientist to understand much of the biology discussed in this book. This chapter provides the briefest of overviews; I can only begin to convey the depth, variety, complexity and stunning beauty of the universe of living things.

Much of what follows is not about *molecular* biology per se. In order to

explain what the molecules are doing, it is often necessary to use concepts involving, for example, cells, embryological development, or evolution. Biology is frustratingly holistic. Events at one level can effect and be affected by events at very different levels of scale or time. Digesting a survey of the basic background material is a prerequisite for understanding the significance of the molecular biology that is described elsewhere in the book. In life, as in cognition, context is very important.

Do keep one rule in the back of your mind as you read this: for every generalization I make about biology, there may well be thousands of exceptions. There are a lot of living things in the world, and precious few generalizations hold true for all of them. I will try to cover the principles; try to keep the existence of exceptions in mind as you read. Another thing to remember is that an important part of understanding biology is learning its language. Biologists, like many scientists, use technical terms in order to be precise about reference. Getting a grasp on this terminology makes a great deal of the biological literature accessible to the non-specialist. The notes contain information about terminology and other basic matters. With that, let's begin at the beginning.

1. What Is Life?

No simple definition of what it is to be a living thing captures our intuitions about what is alive and what is not. The central feature of life is its ability to reproduce itself. Reproductive ability alone is not enough; computer programs can create endless copies of themselves—that does not make them alive. Crystals influence the matter around them to create structures similar to themselves but they're not alive, either. Most living things take in materials from their environment and capture forms of energy they can use to transform those materials into components of themselves or their offspring. Viruses, however, do not do that; they are nearly pure genetic material, wrapped in a protective coating. The cell that a virus infects does all the synthetic work involved in creating new viruses. Are viruses a form of life? Many people would say so.

Another approach to defining "life" is to recognize its fundamental interrelatedness. All living things are related to each other. Any pair of organisms, no matter how different, have a common ancestor sometime in the distant past. Organisms came to differ from each other, and to reach modern levels of complexity through *evolution*. Evolution has three components: inheritance, the passing of characteristics from parents to offspring; variation, the processes that make offspring other than exact copies of their parents; and selection, the process that differentially favors the reproduction of some organisms, and hence their characteristics, over others. These three factors define an evolutionary process. Perhaps the best definition of life is that it is

the result of the evolutionary process taking place on Earth. Evolution is the key not only to defining what counts as life but also to understanding how living systems function.

Evolution is a cumulative process. *Inheritance* is the determinant of almost all of the structure and function of organisms; the amount of variation from one generation to the next is quite small. Some aspects of organisms, such as the molecules that carry energy or genetic information, have changed very little since that original common ancestor several billion of years ago. Inheritance alone, however, is not sufficient for evolution to occur; perfect inheritance would lead to populations of entirely identical organisms, all exactly like the first one.

In order to evolve, there must be a source of *variation* in the inheritance. In biology, there are several sources of variation. Mutation, or random changes in inherited material, is only one source of change; sexual recombination and various other kinds of genetic rearrangements also lead to variations; even viruses can get into the act, leaving a permanent trace in the genes of their hosts. All of these sources of variation modify the message that is passed from parent to offspring; in effect, exploring a very large space of possible characteristics. It is an evolutionary truism that almost all variations are neutral or deleterious. As computer programmers well know, small changes in a complex system often lead to far-reaching and destructive consequences (And computer programmers make those small changes by design, and with the hope of improving the code!). However, given enough time, the search of that space has produced many viable organisms.

Living things have managed to adapt to a breathtaking array of challenges, and continue to thrive. *Selection* is the process by which it is determined which variants will persist, and therefore also which parts of the space of possible variations will be explored. Natural selection is based on the reproductive fitness of each individual. Reproductive fitness is a measure of how many surviving offspring an organism can produce; the better adapted an organism is to its environment, the more successful offspring it will create. Because of competition for limited resources, only organisms with high fitness will survive; organisms less well adapted to their environment than competing organisms will simply die out.

I have likened evolution to a search through a very large space of possible organism characteristics. That space can be defined quite precisely. All of an organism's inherited characteristics are contained in a single messenger molecule: deoxyribonucleic acid, or DNA. The characteristics are represented in a simple, linear, four-element code. The translation of this code into all the inherited characteristics of an organism (e.g. its body plan, or the wiring of its nervous system) is complex. The particular genetic encoding for an organism is called its *genotype*. The resulting physical characteristics of an organism is called its *phenotype*. In the search space metaphor, every point in the

space is a genotype. Evolutionary variation (such as mutation, sexual recombination and genetic rearrangements) identifies the legal moves in this space. Selection is an evaluation function that determines how many other points a point can generate, and how long each point persists. The difference between genotype and phenotype is important because allowable (i.e. small) steps in genotype space can have large consequences in phenotype space. It is also worth noting that search happens in genotype space, but selection occurs on phenotypes. Although it is hard to characterize the size of phenotype space, an organism with a large amount of genetic material (like, e.g., that of the flower Lily) has about 10^{11} elements taken from a four letter alphabet, meaning that there are roughly $10^{70,000,000,000}$ possible genotypes of that size or less. A vast space indeed! Moves (reproductive events) occur asynchronously, both with each other and with the selection process. There are many nondeterministic elements; for example, in which of many possible moves is taken, or in the application of the selection function. Imagine this search process running for billions of iterations, examining trillions of points in this space in parallel at each iteration. Perhaps it is not such a surprise that evolution is responsible for the wondrous abilities of living things, and for their tremendous diversity.*

1.1 The Unity and the Diversity of Living Things

Life is extraordinarily varied. The differences between a tiny archebacterium living in a superheated sulphur vent at the bottom of the ocean and a two-ton polar bear roaming the arctic circle span orders of magnitude in many dimensions. Many organisms consist of a single cell; a Sperm Whale has more than 10^{15} cells. Although very acidic, very alkaline or very salty environments are generally deadly, living things can be found in all of them. Hot or cold, wet or dry, oxygen-rich or anaerobic, nearly every niche on the planet has been invaded by life. The diversity of approaches to gathering nutrients, detecting danger, moving around, finding mates (or other forms of reproduction), raising offspring and dozens of other activities of living creatures is truly awesome. Although our understanding of the molecular level of life is less detailed, it appears that this diversity is echoed there. For example, proteins with very similar shapes and identical functions can have radically different chemical compositions. And organisms that look quite similar to each other may have very different genetic blueprints. All of the genetic material in an organism is called its *genome*. Genetic material is discrete and hence has a particular size, although the size of the genome is not directly related to the complexity of the organism. The size of genomes varies from about 5,000 elements in a very simple organism (e.g. the viruses SV40 or ϕx) to more than 10^{11} elements

*Evolution has also become an inspiration to a group of researchers interested in designing computer algorithms, e.g. Langton (1989).

in some higher plants; people have about $3x10^9$ elements in their genome.

Despite this incredible diversity, nearly all of the same basic mechanisms are present in all organisms. All living things are made of cells*: membrane-enclosed sacks of chemicals carrying out finely tuned sequences of reactions. The thousand or so substances that make up the basic reactions going on inside the cell (the core *metabolic pathways*) are remarkably similar across all living things. Every species has some variations, but the same basic materials are found from bacteria to human. The genetic material that codes for all of these substances is written in more or less the same molecular language in every organism. The developmental pathways for nearly all multicellular organisms unfold in very similar ways. It is this underlying unity that offers the hope of developing predictive models of biological activity. It is the process of evolution that is responsible both for the diversity of living things and for their underlying similarities. The unity arises through inheritance from common ancestors; the diversity from the power of variation and selection to search a vast space of possible living forms.

1.2 Prokaryotes & Eukaryotes, Yeasts & People

Non-biologists often fail to appreciate the tremendous number of different kinds of organisms in the world. Although no one really knows, estimates of the number of currently extant species range from 5 million to 50 million (May, 1988).† There are at least 300,000 different kinds of beetles alone, and probably 50,000 species of tropical trees. Familiar kinds of plants and animals make up a relatively small proportion of the kinds of living things, perhaps only 20%. Vertebrates (animals with backbones: fish, reptiles, amphibians, birds, mammals) make up only about 3% of the species in the world.

Since Aristotle, scholars have tried to group these myriad species into meaningful classes. This pursuit remains active, and the classifications are, to some degree, still controversial. Traditionally, these classifications have been based on the *morphology* of organisms. Literally, morphology means shape, but it is generally taken to include internal structure as well. Morhpology is only part of phenotype, however; other parts include physiology, or the functioning of living structures, and development. Structure, development and function all influence each other, so the dividing lines are not entirely clear.

In recent years, these traditional taxonomies have been shaken by information gained from analyzing genes directly, as well as by the discovery of an entirely new class of organisms that live in hot, sulphurous environments in the deep sea.

*A virus is arguably alive, and is not a cell, but it depends on infecting a cell in order to reproduce.

†May also notes that it is possible that half the extant species on the planet may become extinct in the next 50 to 100 years.

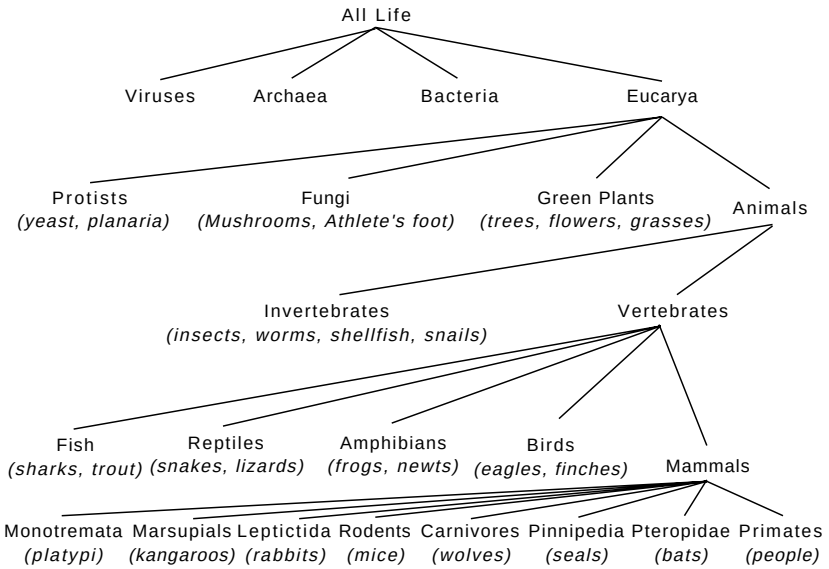

Figure 1. A very incomplete and informal taxonomic tree. Items in italics are common names of representative organisms or classes. Most of the elided taxa are Bacteria; Vertebrates make up only about 3% of known species.

Here I will follow Woese, Kandler & Wheelis (1990), although some aspects of their taxonomy are controversial. They developed their classification of organisms by using distances based on sequence divergence in a ubiquitous piece of genetic sequence As shown in Figure 1, there are three most basic divisions: the Archaea, the Bacteria and the Eucarya. Eucarya (also called eucaryotes) are the creatures we are most familiar with. They have cells that contain nuclei, a specialized area in the cell that holds the genetic material. Eucaryotic cells also have other specialized cellular areas, called organelles. An example of organelles are mitochondria and chloroplasts. Mitochondria are where respiration takes place, the process by which cells use oxygen to improve their efficiency at turning food into useful energy. Chloroplasts are organelles found in plants that capture energy from sunlight. All multicellular organisms, (e.g. people, mosquitos and maple trees) are Eucarya, as are many single celled organisms, such as yeasts and paramecia.

Even within Eucarya, there are more kinds of creatures than many non-biologists expect. Within the domain of the eucaryotes, there are generally held to be at least four kingdoms: animals, green plants, fungi and protists. From a genetic viewpoint, the protists, usually defined as single celled organisms other than fungi, appear to be a series of kingdoms, including at least the cili-

ates (cells with many external hairs, or cillia), the flagellates (cells with a single, long external fiber) and the microsporidia. The taxonomic tree continues down about a dozen levels, ending with particular species at the leaves. All of these many eucaryotic life forms have a great deal in common with human beings, which is the reason we can learn so much about ourselves by studying them.

Bacteria (sometimes also called eubacteria, or prokaryotes) are ubiquitous single-celled organisms. And ubiquitous is the word; there are millions of them everywhere — on this page, in the air you are breathing, and in your gut, for example. The membranes that enclose these cells are typically made of a different kind of material than the ones that surround eucarya, and they have no nuclei or other organelles (they do have ribosomes, which are sometimes considered organelles; see below). Almost all bacteria do is to make more bacteria; it appears that when food is abundant, the survival of the fittest in bacteria means the survival of those that can divide the fastest (Alberts, et al., 1989). Bacteria include not only the disease causing "germs," but many kinds of algae, and a wide variety of symbiotic organisms, including soil bacteria that fix nitrogen for plants and Escherichia coli, a bacterium that lives in human intestines and is required for normal digestion. E. coli is ubiquitous in laboratories because it is easy to grow and very well studied.

Archaea are a recently discovered class of organism so completely unlike both bacteria and eucarya, both genetically and morphologically, that they have upset a decades old dichotomy. Archaea live in superheated sulphur vents in the deep sea, or in hot acid springs, briney bogs and other seemingly inhospitable places. They are sometimes called *archebacteria* even though they bear little resemblence to bacteria. Their cell membranes are unlike either Bacteria or Eucarya. Although they have no nuclei or organelles, at a genetic level, they are a bit more like Eucarya than like Bacteria. These organisms are a relatively recent discovery, and any biological theories have yet to include Archaea, or consider them simply another kind of procaryote. Archaea will probably have a significant effect on theories about the early history of life, and their unusual biochemistry has already turned out to be scientifically and commercially important (e.g. see the discussion of PCR in the last section of this chapter).

Viruses form another important category of living forms. They are *obligatory parasites* meaning that they rely on the biochemical machinery of their host cell to survive and reproduce. Viruses consist of just a small amount of genetic material surrounded by a protein coat. A small virus, such as ϕX, which infects bacteria, can have as few as 5000 elements in its genetic material. (Viruses that infect bactieria are called *bacteriophages*, or just *phages*.) Their simplicity and their role in human disease make viruses an active area of study. They also play a crucial role in the technology of molecular biology, as is described in the last section in this chapter.

1.3 Evolutionary Time and Relatedness

There are so many different kinds of life, and they live in so many different ways. It is amazing that their underlying functioning is so similar. The reason that there is unity within all of that diversity is that all organisms appear to have evolved from a common ancestor. This fundamental claim underpins nearly all biological theorizing, and there is substantial evidence for it.

All evolutionary theories hold that the diversity of life arose by inherited variation through an unbroken line of descent. This common tree of descent is the basis for the taxonomy described above, and pervades the character of all biological explanation. There is a great deal of argument over the detailed functioning of evolution (e.g. whether it happens continuously or in bursts), but practically every biologist agrees with that basic idea.

There are a variety of ways to estimate how long ago two organisms diverged; that is, the last time they had a common ancestor. The more related two species are, the more recently they diverged. To the degree that phenotypic similarity indicates genotypic similarity, organisms can be classified on the basis of their structure, which is the traditional method. Growing knowledge of the DNA sequences of many genes in many organisms makes possible estimates of the time of genetic divergence directly, by comparing their genetic sequences. If the rate of change can be quantified, and standards set, these differences can be translated into a "molecular clock;" Li & Graur, (1991) is a good introduction to this method. The underlying and somewhat controversial assumption is that in some parts of the genome, the rate of mutation is fairly constant. There are various methods for trying to find these areas, estimate the rate of change, and hence calibrate the clock. The technique has mostly confirmed estimates made with other methods, and is widely considered to be potentially reliable, if not quite yet so. Most of the dates I will use below were derived from traditional (archaeological) dating.

In order to get a rough idea of the degrees of relatedness among creatures, it is helpful to know the basic timeline of life on Earth. The oldest known fossils, stromalites found in Australia, indicate that life began at least 3.8 billion years ago. Geological evidence indicates that a major meteor impact about 4 billion years ago vaporized all of the oceans, effectively destroying any life that may have existed before that. In effect, life on earth began almost as soon as it could have. Early life forms probably resembled modern bacteria in some important ways. They were simple, single celled organisms, without nuclei or other organelles. Life remained like that for nearly 2 billion years. Then, about halfway through the history of life, a radical change occurred: Eucarya came into being. There is evidence that eucarya began as symbiotic collections of simpler cells which were eventually assimilated and became organelles (see, e.g. Margolis (1981)). The advantages of these specialized cellular organelles made early eucarya very successful. Single-celled

Eucarya become very complex, for example, developing mechanisms for moving around, detecting prey, paralyzing it and engulfing it.

The next major change in the history of life was the invention of sex. Evolution, as you recall, is a mechanism based on the inheritance of variation. Where do these variations come from? Before the advent of sex, variations arose solely through individual, random changes in genetic material. A mutation might arise, changing one element in the genome, or a longer piece of a genome might be duplicated or moved. If the changed organism had an advantage, the change would propagate itself through the population. Most mutations are neutral or deleterious, and evolutionary change by mutation is a very slow, random search of a vast space. The ability of two successful organisms to combine bits of their genomes into an offspring produced variants with a much higher probability of success. Those moves in the search space are more likely to produce an advantageous variation than random ones. Although you wouldn't necessarily recognize it as sex when looking under a microscope, even some Bacteria exchange genetic material. How and when sexual recombination first evolved is not clear, but it is quite ancient. Some have argued that sexual reproduction was a necessary precursor to the development of multicellular organisms with specialized cells (Buss, 1987). The advent of sex dramatically changed the course of evolution. The new mechanism for the generation of variation focused nature's search through the space of possible genomes, leading to an increase in the proportion of advantageous variations, and an increase in the rate of evolutionary change.

This is probably a good place to correct a common misperception, namely that some organisms are more "primitive" than others. Every existing organism has, tautologically, made it into the modern era. *Simple modern organisms are not primitive.* The environment of the modern world is completely unlike that of earth when life began, and even the simplest existing creatures have evolved to survive in the present. It is possible to use groups of very distantly related creatures (e.g. people and bacteria) to make inferences about ancient organisms; whatever people and bacteria have in common are characteristics that were most likely shared by their last common ancestor, many eons ago. Aspects of bacteria which are not shared with people may have evolved as recently as any human characteristic not shared with bacteria. This applies to the relation between people and apes, too: apes are not any more like ancestral primates than we are. It is what we have *in common* with other organisms that tells us what our ancestors were like; the differences between us and other organisms are much less informative.

Whether or not it occurred as a result of the advent of sexual recombination, the origin of multicellular organisms led to a tremendous explosion in the kinds of organisms and in their complexity. This event occurred only about a billion years ago, about three quarters of the way through the history of life.

Of course, nearly all of the organisms people can see are multicellular (although the blue-green algae in ponds and swimming pools are a kind of bacteria). Multicellular organisms gain their main evolutionary advantage through cellular specialization. Creatures with specialized cells have the ability to occupy environmental niches that single-celled organisms cannot take advantage of. In multicellular organisms, cells quite distant from each other can exchange matter, energy or information for their mutual benefit. For example, cells in the roots of a higher plant exist in a quite different environment than the cells in the leaves, and each supplies the other with matter or energy not available in the local environment.

An important difference between multicellular organisms and a colony of unicellular organisms (e.g. coral) is that multicellular organisms have separated germ line (reproductive) cells from somatic (all the other) cells. Sperm and eggs are germ cells; all the other kinds of cells in the body are somatic. Both kinds of cells divide and make new cells, but only germ cells make new organisms. Somatic cells are usually specialized for a particular task; they are skin cells, or nerve cells, or blood cells. Although these cells divide, when they divide, they create more of the same kind of cell. The division of somatic cells and single celled organisms is a four stage process that ends with *mitosis,* resulting in the production of two identical *daughter cells.* The process as a whole is referred to as the *cell cycle.*

Only changes in germ cells are inherited from an organism to its offspring. A variation that arises in a somatic cell will affect all of the cell's descendents, but it will not affect any of the organism's descendents. Germ cells divide in a process called *meiosis;* part of this process is the production of sperm and egg cells, each of which have only half the usual genetic material. The advent of this distinction involved a complex and intricate balance between somatic cells becoming an evolutionary deadends and the improved competitive ability of a symbiotic collection of closely related cells.

Multicellular organisms all begin their lives from a single cell, a fertilized egg. From that single cell, all of the specialized cells arise through a process called cellular differentiation. The process of development from fertilized egg to full adult is extremely complex. It involves not only cellular differentiation, but the migration and arrangement of cells with respect to each other, orchestrated changes in which genes are used and which are not at any given moment, and even the programmed death of certain groups of cells that act as a kind of scaffolding during development. The transition from single-celled organism to multicellular creature required many dramatic innovations. It was a fundamental shift of the level of selection: away from the individual cell and to a collection of cells as a whole. The reproductive success of a single cell line within a multicellular individual may not correlate with the success of the individual.* Embryology and development are complex and important topics, but are touched on only briefly in this chapter.

Most of the discussion so far has focused on organisms that seem very simple and only distantly related to people. On a biochemical level, however, people are much like other eucaryotes, especially multicellular ones. Genetic and biochemical distance doesn't always correlate very well with morphological differences. For example, two rather similar looking species of frogs may be much more genetically distant from each other than are, say, people and cows (Cherty, Case & Wilson, 1978). A great deal of human biochemistry was already set by the time multicellular organisms appeared on the Earth. We can learn a lot about human biology by understanding how yeasts work.

We've now covered, very briefly, the diversity of living things, and some of the key events in the evolution of life up to the origin of multicellular organisms. In the next section, we'll take a closer look at how these complex organisms work, and cover the parts of eucaryotic cells in a bit more detail.

2. Living Parts: Tissues, Cells, Compartments and Organelles

The main advantage multicellular organisms possess over their single-celled competitors is cell specialization. Not every cell in a larger organism has to be able to extract nutrients, protect itself, sense the environment, move itself around, reproduce itself and so on. These complex tasks can be divided up, so that many different classes of cells can work together, accomplishing feats that single cells cannot. Groups of cells specialized for a particular function are *tissues*, and their cells are said to have *differentiated.* Differentiated cells (except reproductive cells) cannot reproduce an entire organism.

In people (and most other multicellular animals) there are fourteen major tissue types. There are many texts with illustrations and descriptions of the various cell types and tissue, e.g. Kessel and Kardon (1979) which is full of beautiful electron micrographs. Some of these tissue types are familiar: bones, muscles, cardiovascular tissue, nerves, and connective tissue (like tendons and ligaments). Other tissues are the constituents of the digestive, respiratory, urinary and reproductive systems. Skin and blood are both distinctive tissue types, made of highly specialized cells. Lymphatic tissue, such as the spleen and the lymph nodes make up the immune system. Endocrine tissue comprises a network of hormone-producing glands (for example, the adrenal gland, source of adrenaline) that exert global control over various aspects of the body as a whole. Finally, epithelium, the most basic tissue type, lines all of the body's cavities, secreting materials such as mucus, and, in the in-

*Cancer is an example where a single cell line within a multicellular organism reproduces to the detriment of the whole.

testines, absorbing water and nutrients.

There are more than 200 different specialized cell types in a typical vertebrate. Some are large, some small; for example, a single nerve cell connects your foot to your spinal cord, and a drop of blood has more than 10,000 cells in it. Some divide rapidly, others do not divide at all; bone marrow cells divide every few hours, and adult nerve cells can live 100 years without dividing. Once differentiated, a cell cannot change from one type to another. Yet despite all of this variation, all of the cells in a multicellular organism have exactly the same genetic code. The differences between them come from differences in *gene expression*, that is, whether or not a the product a gene codes for is produced, and how much is produced. Control of gene expression is an elaborate dance with many participants. Thousands of biological substances bind to DNA, or bind to other biomolecules that bind to DNA. Genes code for products that turn on and off other genes, which in turn regulate other genes, and so on. One of the key research areas in biology is development: how the intricate, densely interrelated genetic regulatory process is managed, and how cells "know" what to differentiate into, and when and where they do it. A prelude to these more complex topics is a discussion of what cells are made of, and what they do.

2.1 The Composition of Cells

Despite their differences, most cells have a great deal in common with each other. Every cell, whether a Archaea at the bottom of the ocean or a cell in a hair follicle on the top of your head has certain basic qualities: they contain cytoplasm and genetic material, are enclosed in a membrane and have the basic mechanisms for translating genetic messages into the main type of biological molecule, the protein. All eucaryotic cells share additional components. Each of these basic parts of a cell is described briefly below:

Membranes are the boundaries between the cell and the outside world. Although there is no one moment that one can say life came into being, the origin of the first cell membrane is a reasonable starting point. At that moment, self-reproducing systems of molecules were individuated, and cells came into being. All present day cells have a *phospholipid* cell membrane. Phospholipids are *lipids* (oils or fats) with a phosphate group attached. The end with the phosphate group is *hydrophillic (*attracted to water) and the lipid end is *hydrophobic* (repelled by water). Cell membranes consist of two layers of these molecules, with the hydrophobic ends facing in, and the hydrophillic ends facing out. This keeps water and other materials from getting through the membrane, except through special pores or channels.

A lot of the action in cells happens at the membrane. For single celled organisms, the membrane contains molecules that sense the environment, and in some cells it can surround and engulf food, or attach and detach parts of itself in order to move. In Bacteria and Archaea, the membrane plays a crucial

role in energy production by maintaining a large acidity difference between the inside and the outside of the cell. In multicellular organisms, the membranes contain all sorts of signal transduction mechanisms, adhesion molecules, and other machinery for working together with other cells.

Proteins are the molecules that accomplish most of the functions of the living cell. The number of different structures and functions that proteins take on in a single organism is staggering. They make possible all of the chemical reactions in the cell by acting as *enzymes* that promote specific chemical reactions, which would otherwise occur only so slowly as to be otherwise negligible. The action of promoting chemical reactions is called *catalysis*, and enzymes are sometimes refered to as *catalysts*, which is a more general term. Proteins also provide structural support, and are the keys to how the immune system distinguishes self from invaders. They provide the mechanism for acquiring and transforming energy, as well as translating it into physical work in the muscles. They underlie sensors and the transmission of information as well.

All proteins are constructed from linear sequences of smaller molecules called amino acids. There are twenty naturally occurring amino acids. Long proteins may contain as many as 4500 amino acids, so the space of possible proteins is very large: 20^{4500} or 10^{5850}. Proteins also fold up to form particular three dimensional shapes, which give them their specific chemical functionality. Although it is easily demonstrable that the linear amino acid sequence completely specifies the three dimensional structure of most proteins, the details of that mapping is one of the most important open questions of biology. In addition a protein's three dimensional structure is not fixed; many proteins move and flex in constrained ways, and that can have a significant role in their biochemical function. Also, some proteins bind to other groups of atoms that are required for them to function. These other structures are called *prosthetic groups.* An example of a prosthetic group is *heme,* which binds oxygen in the protein hemoglobin. I will discuss proteins in more detail again below.

Genetic material codes for all the other constituents of the the cell. This information is generally stored in long strands of DNA. In Bacteria, the DNA is generally circular. In Eucaryotes, it is linear. During cell division Eucaryotic DNA is grouped into X shaped structures called chromosomes. Some viruses (like the AIDS virus) store their genetic material in RNA. This genetic material contains the blueprint for all the proteins the cell can produce. I'll have much more to say about DNA below.

Nuclei are the defining feature of Eucaryotic cells. The nucleus contains the genetic material of the cell in the form of *chromatin*. Chromatin contains long stretches of DNA in a variety of conformations,* surrounded by *nuclear proteins.* The nucleus is separated from the rest of the cell by a *nuclear membrane.* Nuclei show up quite clearly under the light microscope; they are per-

haps the most visible feature of most cells.

Cytoplasm is the name for the gel-like collection of substances inside the cell. All cells have cytoplasm. The cytoplasm contains a wide variety of different substances and structures. In Bacteria and Archaea, the cytoplasm contains all of the materials in the cell. In Eucarya, the genetic material is segregated into the cell nucleus.

Ribosomes are large molecular complexes, composed of several proteins and RNA molecules. The function of ribosomes is to assemble proteins. All cells, including Bacteria and Archaea have ribosomes. The process of translating genetic information into proteins is described in detail below. Ribosomes are where that process occurs, and are a key part of the mechanism for accomplishing that most basic of tasks.

Mitochondria and Chroloplasts are cellular organelles involved in the production the energy that powers the cell. Mitochondria are found in all eucaryotic cells, and their job is respiration: using oxygen to efficiently turn food into energy the cell can use. Some bacteria and archaea get their energy by a process called *glycolysis*, from glyco- (sugar) and -lysis (cleavage or destruction). This process creates two energy-carrying molecules for every molecule of sugar consumed. As oxygen became more abundant†, some organisms found a method for using it (called *oxidative phosphorylation*) to make an order of magnitude increase in their ability to extract energy from food, getting 36 energy-carrying molecules for every sugar.

These originally free living organisms were engulfed by early eucaryotes. This symbiosis gradually became obligatory as eucaryotes came to depend on their mitochondria for energy, and the mitochondria came to depend on the surrounding cell for many vital functions and materials. Mitochondria still have their own genetic material however, and, in sexually reproducing organisms, are inherited only via the cytoplasm of the egg cell. As a consequence, all mitochondria are maternally inherited.

Like the mitochondria, chloroplasts appear to have originated as free-living bacteria that eventually became obligatory symbionts, and then parts of eucaryotic plant cells. Their task is to convert sunlight into energy-carrying molecules.

Other Parts of Cells. There are other organelles found in eucaryotic

**Conformation* means shape, connoting one of several possible shapes. DNA conformations include the traditional double helix, a *supercoiled* state where certain parts of the molecule are deeply hidden, a reverse coiled state called Z-DNA, and several others.

†There was very little oxygen in the early atmosphere. Oxygen is a waste product of glycolysis, and it eventually became a significant component of the atmosphere. Although many modern organisms depend on oxygen to live, it is a very corrosive substance, and living systems had to evolve quite complex biochemical processes for dealing with it.

cells. The *endoplasmic reticulum* (there are two kinds, rough and smooth) is involved in the production of the cell membrane itself, as well as in the production of materials that will eventually be exported from the cell. The *Golgi apparatus* are elongated sacs that are involved in the packaging of materials that will be exported from the cell, as well as segregating materials in the cell into the correct intracellular compartment. *Lysosomes* contain substances that are used to digest proteins; they are kept separate to prevent damage to other cellular components. Some cells have other structures, such as *vacuoles* of lipids for storage (like the ones often found around the abdomen of middle-aged men).

Now that you have a sense of the different components of the cell, we can proceed to examine the activities of these components. Life is a dynamical system, far from equilibrium. Biology is not only the study of living things, but living actions.

3. Life as a Biochemical Process

Beginning with the highest levels of taxonomy, we have taken a quick tour of the varieties of organisms, and have briefly seen some of their important parts. So far, this account has been entirely descriptive. Because of the tremendous diversity of living systems, descriptive accounts are a crucial underpinning to any more explanatory theories. In order to understand how biological systems work, one has to know what they are.

Knowledge of cells and tissues makes possible the functional accounts of physiology. For example, knowing that the cells in the bicep and in the heart are both kinds of muscle helps explain how the blood circulates. However, at this level of description, the work that individual cells are able to do remains mysterious. The revolution in biology over the last three decades resulted from the understanding cells in terms of their chemistry. These insights began with descriptions of the molecules involved in living processes, and now increasingly provides an understanding of the molecular structures and functions that are the fundamental objects and actions of living material.

More and more of the functions of life (e.g. cell division, immune reaction, neural transmission) are coming to be understood as the interactions of complicated, self-regulating networks of chemical reactions. The substances that carry out and regulate these activities are generally referred to as biomolecules. Biomolecules include proteins, carbohydrates, lipids—all called *macromolecules* because they are relatively large—and a variety of small molecules. The genetic material of the cell specifies how to create proteins, as well as when and how much to create. These proteins, in turn, control the flow of energy and materials through the cell, including the creation and transformation of carbohydrates, lipids and other molecules, ultimately accomplishing all of the functions that the cell carries out. The genetic material

itself is also now known to be a particular macromolecule: DNA.

In even the simplest cell, there are more than a thousand kinds of biomolecules interacting with each other; in human beings there are likely to be more than 100,000 different kinds of proteins specified in the genome (it is unlikely that all of them are present in any particular cell). Both the amount of each molecule and its concentration in various compartments of the cell determines what influence it will have. These concentrations vary over time, on scales of seconds to decades. Interactions among biomolecules are highly non-linear, as are the interactions between biomolecules and other molecules from outside the cell. All of these interactions take place in parallel among large numbers of instances of each particular type. Despite this daunting complexity, insights into the structure and function of these molecules, and into their interactions are emerging very rapidly.

One of the reasons for that progress is the conception of life as a kind of information processing. The processes that transform matter and energy in living systems do so under the direction of a set of symbolically encoded instructions. The "machine" language that describes the objects and processes of living systems contains four letters, and the text that describes a person has about as many characters as three years' worth of the *New York Times* (about $3x10^9$). In the next section, we will delve more deeply into the the chemistry of living systems.

4. The Molecular Building Blocks of Life

Living systems process matter, energy and information. The basic principle of life, reproduction, is the transformation of materials found in the environment of an organism into another organism. Raw materials from the local environment are broken down, and then reassembled following the instructions in the genome. The offspring will contain instructions similar to the parent. The matter, energy and information processing abilities of living systems are very general; one of the hallmarks of life is its adaptability to changing circumstances. Some aspects of living systems have, however, stayed the same over the years. Despite nearly 4 billion years of evolution, the basic molecular objects for carrying matter, energy and information have changed very little. The basic units of matter are proteins, which subserve all of the structural and many of the functional roles in the cell; the basic unit of energy is a phosphate bond in the molecule adenosine triphosphate (ATP); and the units of information are four nucleotides, which are assembled together into DNA and RNA.

The chemical composition of living things is fairly constant across the entire range of life forms. About 70% of any cell is water. About 4% are small molecules like sugars and inorganic *ions*[*]. One of these small molecules is ATP, the energy carrier. Proteins make up between 15% and 20% of the cell;

DNA and RNA range from 2% to 7% of the weight. The cell membranes, lipids and other, similar molecules make up the remaining 4% to 7% (Alberts, et al., 1989).

4.1 Energy

Living things obey all the laws of chemistry and physics, including the second law of thermodynamics, which states that the amount of entropy (disorder) in the universe is always increasing. The consumption of energy is the only way to create order in the face of entropy. Life doesn't violate the second law; living things capture energy in a variety of forms, use it to create internal order, and then transfer energy back to the environment as heat. An increase in organization within a cell is coupled with a greater increase in disorder outside the cell.

Living things must capture energy, either from sunlight through photosynthesis or from nutrients by respiration. The variety of chemicals that can be oxidized by various species to obtain energy through respiration is immense, ranging from simple sugars to complex oils and even sulfur compounds from deep sea vents (in the case of Archaea).

In many cases, the energy is first available to the cell as an electrochemical gradient across the cell membrane. The cell can tap into electrochemical gradient by coupling the energy that results from moving electrons across the membrane to other processes. There are many constraints on the flow of energy through a living system. Most of the chemical reactions that organisms need to survive require an input of a minimum amount of energy to take place at a reasonable rates; efficient use of energy dictates that this must be delivered in a quanta exceeding the minimum requirement only slightly.

The energy provided for biochemical reactions has to be useable by many different processes. It must be possible to provide energy where it is needed, and to store it until it is consumed. The uses of energy throughout living systems are very diverse. It is needed to synthesize and transport biomolecules, to create mechanical action through the muscle proteins actin and myosin, and to create and maintain electrical gradients, such as the ones that neurons use to communicate and compute.

Storing and transporting energy in complex biochemical systems runs the

*An inorganic ion is a charged atom, or a charged small group of atoms, not involving carbon. These substances, like iron and zinc, play small but vital role. For example, changing the balance of calcium and sodium ions across a cell membrane is the basic method for exciting of neurons.

The individual building blocks of the larger molecules, i.e. amino acids and nucleic acids, are also considered small molecules when not part of a larger structure. Some of these molecules play roles in the cell other than as components of large molecules. For example, the nucleic acid adenine is at the core of the energy carrying molecule adenosine triphosphate (ATP).

risk of disrupting chemical bonds other than the target ones, so the unit of energy has to be small enough not to do harm, but large enough to be useful. The most common carrier of energy for storage and transport is the outermost phosphate bond in the molecule *adenosine triphosphate*, or *ATP*. This molecule plays a central role in every living system: it is the carrier of energy. Energy is taken out of ATP by the process of *hydrolysis*, which removes the outermost phosphate group, producing the molecule adenosine diphosphate (ADP). This process generates about 12 kcal per mole* of ATP, a quantity appropriate for performing many cellular tasks. The energy "charge" of a cell is expressed in the ratio of ATP/ADP and the electrochemical difference between the inside and the outside of the cell (which is called the *transmembrane potential*). If ATP is depleted, the movement of ions caused by the transmembrane potential will result in the synthesis of additional ATP. If the transmembrane potential has been reduced (for example, after a neuron fires), ATP will be consumed to pump ions back across the gradient and restore the potential.

ATP is involved in most cellular processes, so it is sometimes called a *currency* metabolite. ATP can also be converted to other high energy phosphate compounds such as *creatine phosphate*, or other nucleotide triphosphates. In turn, these molecules provide the higher levels of energy necessary to transcribe genes and replicate chromosomes. Energy can also be stored in different chemical forms. Carbohydrates like glycogen provide a moderate density, moderately accessible form of energy storage. Fats have very high energy storage density, but the energy stored in them takes longer to retrieve.

4.2 Proteins

Proteins are the primary components of living things, and they play many roles. Proteins provide structural support and the infrastructure that holds a creature together; they are enzymes that make the chemical reactions necessary for life possible; they are the switches that control whether genes are turned on or off; they are the sensors that see and taste and smell, and the effectors that make muscles move; they are the detectors that distinguish self from nonself and create an immune response. Finding the proteins that make up a creature and understanding their function is the foundation of explanation in molecular biology.

Despite their radical differences in function, all proteins are made of the same basic constituents: the amino acids. Each amino acid shares a basic structure, consisting of a central carbon atom (C), an *amino* group (NH_3) at

*kcal is an abbreviation for kilocalorie, the amount of energy necessary to raise a liter of water one degree centigrade at standard temperature and pressure. It is equivalent to 1 dieter's calorie. A mole is an amount of a substance, measured in terms of the number of molecules, rather than by its mass. One mole is 6×10^{23} molecules.

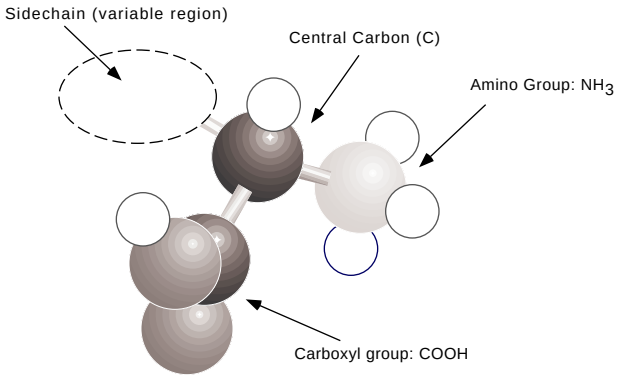

Figure 2: The basic chemical structure of an amino acid. Carbon atoms are black, Oxygen is dark grey, Nitrogen light grey, and hydrogen white.

one end, a *carboxyl* group (COOH) at the other, and a variable sidechain (R), as shown in Figure 2. These chemical groups determine how the molecule functions, as Mavrovouniotis's chapter in this volume explains. For example, under biological conditions the amino end of the molecule is positively charged, and the carboxyl end is negatively charged. Chains of amino acids are assembled by a reaction that occurs between the nitrogen atom at the amino end of one amino acid and the carbon atom at the carboxyl end of another, bonding the two amino acids and releasing a molecule of water. The linkage is called a *peptide bond*, and long chains of amino acids can be strung together into polymers*, called *polypeptides*, in this manner. All proteins are polypeptides, although the term polypeptide generally refers to chains that are shorter than whole proteins.

When a peptide bond is formed, the amino acid is changed (losing two hydrogen atoms and an oxygen atom), so the portion of the original molecule integrated into the polypeptide is often called a *residue*. The sequence of amino acid residues that make up a protein is called the protein's *primary structure.* The primary structure is directly coded for in the genetic material: The individual elements of a DNA molecule form triples which unambiguously specify an amino acid. A genetic sequence maps directly into a sequence of amino acids. This process is discussed in greater detail below.

It is interesting to note that only a small proportion of the very many possible polypeptide chains are naturally occurring proteins. Computationally, this is unsurprising. Many proteins contain more than 100 amino acids (some

*Polymers are long strings of similar elements; -mer means "element," as in monomer, dimer, etc. Homopolymer is a term that refers to polymers made up of all the same element; heteropolymers are made of several different units. Proteins and DNA are both heteropolymers. Glycogen, a substance used for the medium-term storage of excess energy, is an example of a homopolymer.

have more than 4000). The number of possible polypeptide chains of length 100 is 20^{100} or more than 10^{130}. Even if we take the high estimates of the number of species ($5x10^7$) and assume that they all have as many different proteins as there are in the most complex organism ($<10^7$) and that no two organisms share a single protein, the ratio of actual proteins to possible polypeptides is much less than $1:10^{100}$—a very small proportion, indeed.

The twenty naturally occuring amino acids all have the common elements shown in Figure 2. The varying parts are called *sidechains*; the two carbons and the nitrogen in the core are sometimes called the *backbone*. Peptide bonds link together the backbones of a sequence of amino acids. That link can be characterized as having two degrees of rotational freedom, the phi (ϕ) and psi (ψ) angles (although from the point of view of physics this is a drastic simplification, in most biological contexts it is valid). The conformation of a protein backbone (i.e. its shape when folded) can be adequately described as a series of ϕ/ψ angles, although it is also possible to represent the shape using the Cartesian coordinates of the central backbone atom (the alpha carbon, written Cα), or using various other representational schemes (see, e.g., Hunter or Zhang & Waltz in this volume).

The dimensions along which amino acids vary are quite important for a number of reasons. One of the major unsolved problems in molecular biology is to be able to predict the structure and function of a protein from its amino acid sequence. It was demonstrated more than two decades ago that the amino acid sequence of a protein determines ultimate conformation and, therefore, its biological activity and function. Exactly how the properties of the amino acids in the primary structure of a protein interact to determine the protein's ultimate conformation remains unknown. The chemical properties of the individual amino acids, however, are known with great precision. These properties form the basis for many representations of amino acids, e.g. in programs attempting to predict structure from sequence. Here is a brief summary of some of them.

Glycine is the simplest amino acid; its sidechain is a single hydrogen atom. It is nonpolar, and does not ionize easily. The *polarity* of a molecule refers to the degree that its electrons are distributed asymmetrically. A nonpolar molecule has a relatively even distribution of charge. *Ionization* is the process that causes a molecule to gain or lose an electron, and hence become charged overall. The distribution of charge has a strong effect on the behavior of a molecule (e.g. like charges repel). Another important characteristic of glycine is that as a result of having no heavy (i.e. non-hydrogen) atoms in its sidechain, it is very flexible. That flexibility can give rise to unusual kinks in the folded protein.

Alanine is also small and simple; its sidechain is just a *methyl* group (consisting of a carbon and three hydrogen atoms). Alanine is one of the most

commonly appearing amino acids. Glycine and alanine's sidechains are *aliphatic*, which means that they are straight chains (no loops) containing only carbon and hydrogen atoms. There are three other aliphatic amino acids: *valine, leucine* and *isoleucine.* The longer aliphatic sidechains are hydrophobic. Hydrophobicity is one of the key factors that determines how the chain of amino acids will fold up into an active protein. Hydrophobic residues tend to come together to form compact core that exclude water. Because the environment inside cells is *aqueous* (primarily water), these hydrophobic residues will tend to be on the inside of a protein, rather than on its surface.

In contrast to alanine and glycine, the sidechains of amino acids *phenylalanine*, *tyrosine* and *tryptophan* are quite large. Size matters in protein folding because atoms resist being too close to one another, so it is hard to pack many large sidechains closely. These sidechains are also *aromatic*, meaning that they form closed rings of carbon atoms with alternating double bonds (like the simple molecule benzene). These rings are large and inflexible. Phenylalanine and tryptophan are also hydrophobic. Tyrosine has a *hydroxyl* group (an OH at the end of the ring), and is therefore more reactive than the other sidechains mentioned so far, and less hydrophobic. These large amino acids appear less often than would be expected ifproteins were composed randomly. *Serine* and *threonine* also contain hydroxyl groups, but do not have rings.

Another feature of importance in amino acids is whether they ionize to form charged groups. Residues that ionize are characterized by their *pK*, which indicates at what *pH* (level of acidity) half of the molecules of that amino acid will have ionized. *Arginine* and *lysine* have high pK's (that is, they ionize in basic environments) and *histidine*, *gluatmic acid* and *aspartic acid* have low pK's (they ionize in acidic ones). Since like charges repel and opposites attract, charge is an important feature in predicting protein conformation. Most of the charged residues in a protein will be found at its surface, although some will form bonds with each other on the inside of the molecule (called *salt-bridges*) which can provide strong constraints on the ultimate folded form.

Cysteine and *methionine* have hydrophobic sidechains that contain a sulphur atom, and each plays an important role in protein structure. The sulphurs make the amino acids' sidechains very reactive. Cysteines can form *disulphide* bonds with each other; disulphide bonds often hold distant parts of a polypeptide chain near each other, constraining the folded conformation like salt bridges. For that reason, cysteines have a special role in determining the three dimensional structure of proteins. The chapter by Holbrook, Muskal and Kim in this volume discusses the prediction of this and other folding constraints. Methionine is also important because all eucaryotic proteins, when originally synthesized in the ribosome, start with a methionine. It is a kind of "start" signal in the genetic code. This methionine is generally re-

moved before the protein is released into the cell, however.

Histidine is a relatively rare amino acid, but often appears in the *active site* of an enzyme. The active site is the small portion of an enzyme that effects the target reaction, and it is the key to understanding the chemistry involved. The rest of the enzyme provides the necessary scaffolding to bring the active site to bear in the right place, and to keep it away from bonds that it might do harm to. Other regions of enzymes can also act as a switch, turning the active site on and off in a process called *allosteric* control. Because histidine's pK is near the typical pH of a cell, it is possible for small, local changes in the chemical environment to flip it back and forth between being charged and not charged. This ability to flip between states makes it useful for catalyzing chemical reactions. Other charged residues also sometimes play a similar role in catalysis.

With this background, it is now possible to understand the basics of the protein folding problem which is the target of many of the AI methods applied in this volume. The genetic code specifies only the amino acid sequence of a protein. As a new protein comes off the ribosome, it folds up into the shape that gives it its biochemical function, sometimes called its *active conformation* (the same protein unfolded into some other shape is said to be *denatured*, which is what happens, e.g. to the white of an egg when you cook it). In the cell, this process takes a few seconds, which is a very long time for a chemical reaction. The complex structure of the ribosome may play a role in protein folding, and a few proteins need helper molecules, termed *chaperones* to fold properly. However, these few seconds are a very short time compared to how long it takes people to figure out how a protein will fold. In raw terms, the folding problem involves finding the mapping from primary sequence (a sequence of from dozens to several thousand symbols, drawn from a 20 letter alphabet) to the real-numbered locations of the thousands of constituent atoms in three space.

Although all of the above features of amino acids play some role in protein folding, there are few absolute rules. The conformation a protein finally assumes will minimize the total "free" energy of the molecule. Going against the tendencies described above (e.g. packing several large sidechains near each other) increases the local free energy, but may reduce the energy elsewhere in the molecule. Each one of the tendencies described can be traded off against some other contribution to the total free energy of the folded protein. Given any conformation of atoms, it is possible in principle to compute its free energy. Ideally, one could examine all the possible conformations of a protein, calculate the free energy by applying quantum mechanical rules, and select the minimum energy conformation as a prediction of the folded structure. Unfortunately, there are very many possible conformations to test, and each energy calculation itself is prohibitively complex. A wide variety of approaches have been taken to making this problem tractable, and, given a few hours of super-

computer time, it is currently possible to evaluate several thousand possible conformations. These techniques are well surveyed in Karplus & Petsko (1990). An alternative to the pure physical simulations are the various AI approaches which a significant portion of this volume is dedicated to describing.

The position of the atoms in a folded protein is called its *tertiary* structure. The *primary* structure is the amino acid sequence. *Secondary* structure refers to local arrangements of a few to a few dozen amino acid residues that take on particular conformations that are seen repeatedly in many different proteins. These shapes are stabilized by *hydrogen bonds* (a hydrogen bond is a relatively weak bond that also plays a role in holding the two strands of the DNA molecule together). There are two main kinds of secondary structure: corkscrew-shaped conformations where the amino acids are packed tightly together, called *α-helices*, and long flat sheets made up of two or more adjacent strands of the molecule, extended so that the amino acids are stretched out as far from each other as they can be. Each extended chain is called a *β-strand*, and two or more β-strands held together by hydrogen bonds are called a *β-sheet*. β-sheets can be composed of strands running in the same direction (called a *parallel* β-sheet) or running in the opposite direction (*antiparallel*). Other kinds of secondary structure include structures that are even more tightly packed than α-helices called *3-10 helices*, and a variety of small structures that link other structures, called *β-turns*. Some local combinations of secondary structures have been observed in a variety of different proteins. For example, two α-helices linked by a turn with an approximately 60° angle have been observed in a variety of proteins that bind to DNA. This pattern is called the *helix-turn-helix* motif, and is an example of what is known as *super-secondary* structure. Finally, some proteins only become functional when assembled with other molecules. Some proteins bind to copies of themselves; for example, some DNA-binding proteins only function as dimers (linked pairs). Other proteins require prostehtic groups such as heme or chlorophyl. Additions necessary to make the folded protein active are termed the protein's *quaternary* structure.

4.3 Nucleic Acids

If proteins are the workhorses of the biochemical world, nucleic acids are their drivers; they control the action. All of the genetic information in any living creature is stored in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), which are polymers of four simple nucleic acid units, called *nucleotides*. There are four nucleotides found in DNA. Each nucleotide consists of three parts: one of two base molecules (a *purine* or a *pyrimidine*), plus a sugar (ribose in RNA and deoxyribose DNA), and one or more phosphate groups. The purine nucleotides are *adenine* (A) and *guanine* (G), and the pyrimidines are *cytosine* (C) and *thymine* (T). Nucleotides are sometimes called bases, and, since DNA consists of two complementary strands bonded

together, these units are often called base-pairs. The length of a DNA sequences is often measured in thousands of bases, abbreviated kb. Nucleotides are generally abbreviated by their first letter, and appended into sequences, written, e.g., CCTATAG. The nucleotides are linked to each other in the polymer by phosphodiester bonds. This bond is directional, a strand of DNA has a head (called the *5'* end) and a tail (the *3'* end).

One well known fact about DNA is that it forms a double helix; that is, two *helical* (spiral-shaped) strands of the polypeptide, running in opposite directions, held together by hydrogen bonds. Adenines bond exclusively with the thymines (A-T) and guanines bond exclusively with cytosines (G-C). Although the sequence in one strand of DNA is completely unrestricted, because of these bonding rules the sequence in the complementary strand is completely determined. It is this feature that makes it possible to make high fidelity copies of the information stored in the DNA. It is also exploited when DNA is transcribed into complementary strands of RNA, which direct the synthesis of protein. The only difference is that in RNA, uracil (U) takes the place of thymine; that is, it bonds to adenine.

DNA molecules take a variety of conformations (shapes) in living systems. In most biological circumstances, the DNA forms a classic double helix, called B-DNA; in certain circumstances, however, it can become supercoiled or even reverse the direction of its twist (this form is called Z-DNA). These alternative forms may play a role in turning particular genes on and off (see below). There is some evidence that the geometry of the B-DNA form (e.g for example, differing twist angles between adjacent base pairs) may also be exploited by cell mechanisms. The fact that the conformation of the DNA can have a biological effect over and above the sequence it encodes highlights an important lesson for computer scientists: *there is more information available to a cell than appears in the sequence databases.* This lesson also applies to protein sequences, as we will see in the discussion of post-translational modification.

Now that we have covered the basic structure and function of proteins and nucleic acids, we can begin to put together a picture of the molecular processing that goes on in every cell.

5. Genetic Expression: From Blueprint to Finished Product

5.1 Genes, the Genome and the Genetic Code

The genetic information of an organism can be stored in one or more distinct DNA molecules; each is called a *chromosome*. In some sexually reproducing organisms, called *diploids*, each chromosome contains two similar DNA molecules physically bound together, one from each parent. Sexually reproducing organisms with single DNA molecules in their chromosomes are

called haploid. Human beings are diploid with 23 pairs of linear chromosomes. In Bacteria, it is common for the ends of the DNA molecule to bind together, forming a circular chromosome. All of the genetic information of an organism, taken together as a whole, is refered to as its *genome*.

The primary role of nucleic acids is to carry the encoding of the primary structure of proteins. Each non-overlapping triplet of nucleotides, called a *codon*, corresponds to a particular amino acid (see table 1). Four nucleotides can form $4^3 = 64$ possible triplets, which is more than the 20 needed to code for each amino acid (pairs would provide only 16 codons). Three of these codons are used to designate the end of a protein sequence, and are called stop codons. The others all code for a particular amino acid. That means that most amino acids are encoded by more than one codon. For example, alanine is represented in DNA by the codons GCT, GCC, GCA and GCG. Notice that the first two nucleotides of these codons are all identical, and that the third is redundant. Although this is not true for all of the amino acids, most codon synonyms differ only in the last nucleotide. This phenomenon is called the *degeneracy* of the code. Whether it is an artifact of the evolution, or serves a purpose such as allowing general changes in the global composition of DNA (e.g. increasing the proportion of purines) without changing the coded amino acids is still unknown.

There are some small variations in the translation of codons into amino acids from organism to organism. Since the code is so central to the functioning of the cell, it is very strongly conserved over evolution. However, there are a few systems that use a slightly different code. An important example is found in mitochondria. Mitochondria have their own DNA, and probably represent previously free living organisms that were enveloped by eucaryotes. Mitochondrial DNA is translated using a slightly different code, which is more degenerate (has less information in the third nucleotide) than the standard code. Other organisms that diverged very early in evolution, such as the ciliates, also use different codes.

The basic process of synthesizing proteins maps from a sequence of codons to a sequence of amino acids. However, there are a variety of important complications. Since codons come in triples, there are three possible places to start parsing a segment of DNA. For example, the chain ...AATGCGATAAG... could be read ...AAT-GCG-ATA... or ...ATG-CGA-TAA... or ...TGC-GAT-AAG.... This problem is similar to decoding an asynchronous serial bit stream into bytes. Each of these parsings is called a *reading frame*. A parsing with a long enough string of codons with no intervening stop codons is called an *open reading frame*, or *ORF*; and could be translated into a protein. Organisms sometimes code different proteins with overlapping reading frames, so that if the reading process shifts by one character, a completely different, but still functional protein results! More often, frame shifts, which can be introduced by insertions and deletions in the DNA se-

quence or transcriptional "stuttering," produce nonsense.

Not only are there three possible reading frames in a DNA sequence, it is possible to read off either strand of the double helix. Recall that the second strand is the complement of the first, so that our example above (AATGCGATAAG) can also be read inverted and in the opposite direction, e.g. CTTATCGCATT. This is sometimes called reading from the *antisense* or *complementary* strand. An antisense message can also be parsed three ways, making a total of 6 possible reading frames for every DNA sequence. There are known examples of DNA sequences that code for proteins in both directions with several overlapping reading frames: quite a feat of compact encoding.

And there's more. DNA sequences coding for a single protein in most eucaryotes have noncoding sequences, called *introns*, inserted into them. These introns are spliced out before the sequence is mapped into amino acids. Different eucaryotes have a variety of different systems for recognizing and removing these introns. Most bacteria don't have introns. It is not known whether introns evolved only after the origin of eucaryotes, or whether selective pressure has caused bacteria to lose theirs. The segments of DNA that actually end up coding for a protein are called *exons*. You can keep these straight by remembering that **in**trons are **in**sertions, and that **ex**ons are **ex**pressed.

DNA contains a large amount of information in addition to the coding sequences of proteins. Every cell in the body has the same DNA, but each cell type has to generate a different set of proteins, and even within a single cell type, its needs change throughout its life. An increasing number of DNA signals that appear to play a role in the control of expression are being characterized. There are a variety of signals identifying where proteins begin and end, where splices should occur, and an exquisitely detailed set of mechanisms for controlling which proteins should be synthesized and in what quantities. Large scale features of a DNA molecule, such as a region rich in Cs and Gs can play a biologically important role, too.

Finally, some exceptions to the rules I mentioned above should be noted. DNA is sometimes found in single strands, particularly in some viruses. Viruses also play other tricks with nucleic acids, such as transcribing RNA into DNA, going against the normal flow of information in the cell. Even non-standard base-pairings sometimes play an important role, such as in the structure of transfer RNA (see below).

5.2 RNA: Transcription, Translation, Splicing & RNA Structure

The process of mapping from DNA sequence to folded protein in eucaryotes involves many steps (see Figure 3). The first step is the *transcription* of a portion of DNA into an RNA molecule, called a messenger RNA (mRNA). This process begins with the binding of a molecule called RNA polymerase

to a location on the DNA molecule. Exactly where that polymerase binds determines which strand of the DNA will be read and in which direction. Parts of the DNA near the beginning of a protein coding region contain signals which can be recognized by the polymerase; these regions are called *promoters*. (Promoters and other control signals are discussed further below.) The polymerase catalyzes a reaction which causes the DNA to be used as a template to create a complementary strand of RNA, called the *primary transcript*. This transcript contains introns as well as exons. At the end of the transcript, 250 or more extra adenosines, called a *poly-A tail*, are often added to the RNA. The role of these nucleotides is not known, but the distinctive signature is sometimes used to detect the presence of mRNAs.

The next step is the *splicing* the exons together. This operation takes takes place in a ribosome-like assembly called a *spliceosome*. The RNA remaining after the introns have been spliced out is called a *mature* mRNA. It is then transported out of the nucleus to the cytoplasm, where it then binds to a ribosome.

A ribosome is a very complex combination of RNA and protein, and its operation has yet to be completely understood. It is at the ribosome that the mRNA is used as a blueprint for the production of a protein; this process is called *translation*. The reading frame that the translation will use is determined by the ribosome. The translation process depends on the presence of molecules which make the mapping from codons in the mRNA to amino acids; these molecules are called *transfer-RNA* or *tRNAs*. tRNAs have an anti-codon (that binds to its corresponding codon) near one end and the corresponding amino acid on the other end. The anti-codon end of the tRNAs bind to the mRNA, bringing the amino acids corresponding the mRNA sequence into physical proximity, where they form peptide bonds with each other. How the tRNAs find only the correct amino acid was a mystery until quite recently. This process depends on the three dimensional structure of the RNA molecule, which is discussed in Steeg's chapter of this volume. As the protein comes off the ribosome, it folds up into its native conformation. This process may involve help from the ribosome itself or from chaperone molecules, as was described above.

Once the protein has folded, other transformations can occur. Various kinds of chemical groups can be bound to different places on the proteins, including sugars, phosphate, actyl or methyl groups. These additions can change the hyrogen bonding proclivity or shape of the protein, and may be necessary to make the protein active, or may keep it from having an effect before it is needed. The general term for these transformations is *post-translational modifications*. Once this process is complete, the protein is then transported to the part of the cell where it will accomplish its function. The transport process may be merely passive diffusion through the cytoplasm, or there may be an active transport mechanism that moves the protein across

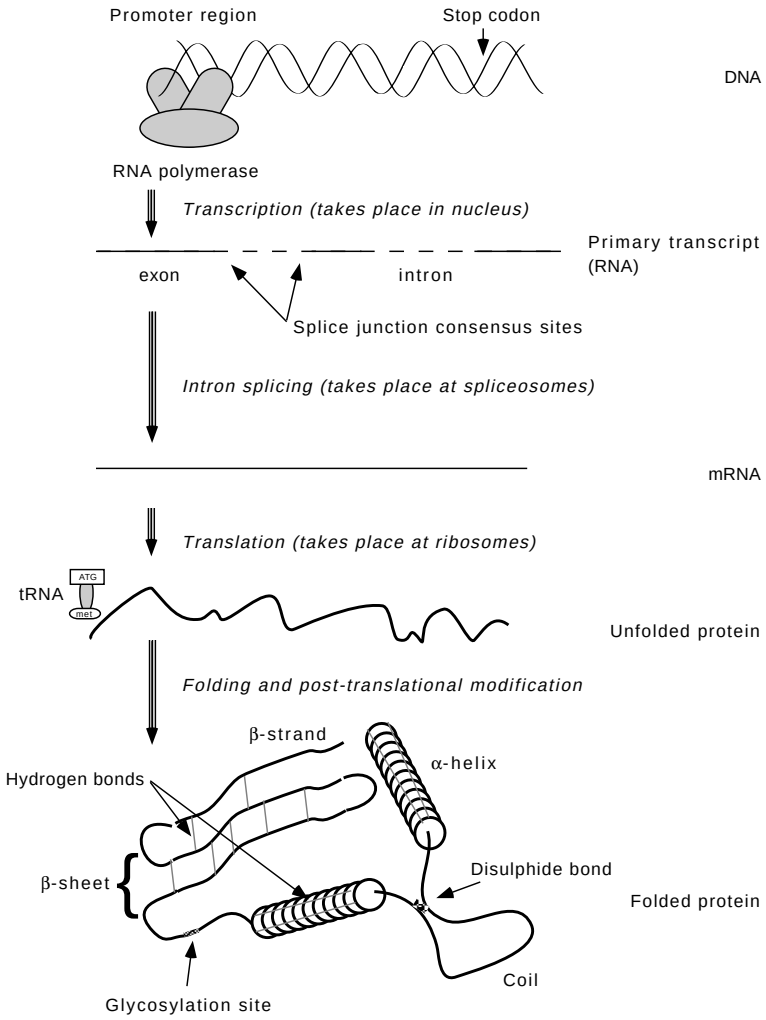

Figure 3. A schematic drawing of the entire process of protein synthesis. An RNA Polymerase binds to a promoter region of DNA, and begins the transcription process, which continues until a stop codon is reached. The product is an RNA molecule called the primary transcript, which contains regions that code for proteins (exons) and regions which do not (introns). The introns are spliced out at splicosomes, and the joined exons are transported to a ribosome. There, transfer RNAs match amino acids to the appropriate codons in the RNA; the amino acids form peptide bonds and become an unfolded protein. The protein then folds into local formations like helices and sheets, and forms internal bonds across longer distances. Post-translational processing can add additional substance; e.g., glycosylation adds sugar molecules to the protein.

membranes or into the appropriate cellular compartment.

5.3 Genetic Regulation

Every cell has the same DNA. Yet the DNA in some cells codes for the proteins needed to function as, say, a muscle, and other code for the proteins to make the lens of the eye. The difference lies in the regulation of the genetic machinery. At any particular time, a particular cell is producing only a small fraction of the proteins coded for in its DNA. And the amount of each protein produced must be precisely regulated in order for the cell to function properly. The cell will change the proteins it synthesizes in response to the environment or other cues. The mechanisms that regulate this process constitute a finely tuned, highly parallel system with extensive multifactoral feedback and elaborate control structure. It is also not yet well understood.

Genes are generally said to be on or off (or *expressed/not expressed*), although the amount of protein produced is also important. The production process is controlled by a complex collection of proteins in the nucleus of eucaryotic cells that influence which genes are expressed. Perhaps the most important of these proteins are the *histones*, which are tightly bound to the DNA in the chromosomes of eucaryotes. Histones are some of the most conserved proteins in all of life. There are almost no differences in the sequence of plant and mammalian histones, despite more than a billion years of divergence in their evolution. Other proteins swarm around the DNA, some influencing the production of a single gene (either encouraging or inhibiting it), while others can influence the production of large numbers of genes at once. An important group of these proteins are called *topoisomerases*; they rearrange and untangle the DNA in various ways, and are the next most prevalent proteins in the chromosome.

Many regulatory proteins recognize and bind to very specific sequences in the DNA. The sequences that these proteins recognize tend to border the protein coding regions of genes, and are known generally as *control regions*. Sequences that occur just upstream (towards the 5' end) of the coding region that encourage the production of the protein are called *promoters.* Similar regions either downstream of the coding region or relatively far upstream are called *enhancers.* Sequences that tend to prevent the production of a protein are called *repressors*. Karp's chapter in this volume discusses how this complex set of interactions can be modeled in knowledge-based systems.

Cells need to turn entire suites of genes on and off in response to many different events, ranging from normal development to trying to repair damage to the cell. The control mechanisms are responsive to the level of a product already in the cell (for homeostatic control) as well as to a tremendous variety of extracellular signals. Perhaps the most amazing activities in gene regulation occur during development; not only are genes turned on and off with precise timing, but the control can extend to producing alternative splic-

ings of the nascent primary transcripts (as is the case in the transition from fetal to normal hemoglobin).

5.4 Catalysis & Metabolic Pathways

The translation of genes into proteins, crucial as it is, is only a small portion of the biochemical activity in a cell. Proteins do most of the work of managing the flow of energy, synthesizing, degrading and transporting materials, sending and receiving signals, exerting forces on the world, and providing structural support. Systems of interacting proteins form the basis for nearly every process of living things, from moving around and digesting food to thinking and reproducing. Somewhat surprisingly, a large proportion of the chemical processes that underlie all of these activities are shared across a very wide range of organisms. These shared processes are collectively referred to as *intermediary metabolism.* These include the *catabolic* processes for breaking down proteins, fats and carbohydrates (such as those found in food) and the *anabolic* processes for building new materials. Similar collections of reactions that are more specialized to particular organisms are called *secondary metabolism.* The substances that these reactions produce and consume are called *metabolites.*

The biochemical processes in intermediary metabolism are almost all *catalyzed reactions*. That is, these reactions would barely take place at all at normal temperatures and pressures; they require special compounds that facilitate the reaction — these compounds are called *catalysts* or *enzymes*. (It is only partially in jest that many biochemistry courses open with the professor saying that the reactions that take place in living systems are ones you were taught were impossible in organic chemistry.) Catalysts are usually named after the reaction they facilitate, usually with the added suffix *-ase*. For example, alcohol dehydrogenase is the enzyme that turns ethyl alcohol into acetaldehyde by removing two hydrogen atoms. Common classes of enzymes include *dehydrogenases*, *synthetases*, *proteases* (for breaking down proteins), *decarboxylases* (removing carbon atoms), *transferases* (moving a chemical group from one place to another), *kinases*, *phosphatases* (adding or removing phosphate groups, respectively) and so on. The materials transformed by catalysts are called *substrates.* Unlike the substrates, catalysts themselves are not changed by the reactions they participate in. A final point to note about enzymatic reactions is that in many cases the reactions can proceed in either direction. That is, and enzyme that transforms substance A into substance B can often also facilitate the transformation of B into A. The direction of the transformation depends on the concentrations of the substrates and on the energetics of the reaction (see Mavrovouniotis' chapter in this volume for further discussion of this topic).

Even the basic transformations of intermediary metabolism can involve

dozens or hundreds of catalyzed reactions. These combinations of reactions, which accomplish tasks like turning foods into useable energy or compounds are called metabolic *pathways.* Because of the many steps in these pathways and the widespread presence of direct and indirect feedback loops, they can exhibit many counterintuitive behaviors. Also, all of these chemical reactions are going on in parallel. Mavrovouniotis's chapter in this volume describes an efficient system for making inferences about these complex systems.

In addition to the feedback loops among the substrates in the pathways, the presence or absence of substrates can affect the behavior of the enzymes themselves, through what is called *allosteric* regulation. These interactions occur when a substance binds to an enzyme someplace other than its usual *active site* (the atoms in the molecule that have the enzymatic effect). Binding at this other site changes the shape of the enzyme, thereby changing its activity. Another method of controlling enzymes is called *competitive inhibition.* In this form of regulation, substance other than the usual substrate of the enzyme binds to the active site of the enzyme, preventing it from having an effect on its substrate.

These are the basic mechanisms underlying eucaryotic cells (and much of this applies to bacterial and archaeal ones as well). Of course, each particular activity of a living system, from the capture of energy to immune response, has its own complex network of biochemical reactions that provides the mechanism underlying the function. Some of these mechanisms, such as the *secondary messenger system* involving cyclic adenosine monophosphate (cAMP) are widely shared by many different systems. Others are exquisitely specialized for a particular task in a single species: my favorite example of this is the evidence that perfect pitch in humans (being able to identify musical notes absolutely, rather than relative to each other) is mediated by a single protein. The functioning of these biochemical networks is being unravelled at an ever increasing rate, and the need for sophisticated methods to analyze relevant data and build suitable models is growing rapidly.

5.5 Genetic Mechanisms of Evolution

In the beginning of this chapter, I discussed the central role that evolution plays in understanding living systems. The mechanisms of evolution at the molecular level are increasingly well understood. The similarities and differences among molecules that are closely related provide important information about the structure and function of those molecules. Molecules (or their sequences) which are related to one another are said to be *homologous.* Although genes or proteins that have similar sequences are often assumed to be homologous, there are well known counterexamples due to *convergent evolution.* In these cases, aspects of very distantly related organisms come to resemble one another through very different evolutionary pathways. Unless there is evidence to the contrary, it is usually safe to assume that macromole-

cular sequences that are similar to each other are homologous.

The sources of variation at the molecular level are very important to understanding how molecules come to differ from each other (or *diverge).* Perhaps the best known mechanism of molecular evolution is the *point mutation,* or the change of a single nucleotide in a genetic sequence. The change can be to *insert* a new nucleotide, to *delete* an existing one, or to change one nucleotide into another. Other mechanisms include large scale chromosomal rearrangements and inversions. An important kind of rearrangement is the *gene duplication*; in which additional copies of a gene are inserted into the genome. These copies can then diverge, so that, for example, the original functionality may be preserved at the same time as related new genes evolve. These duplication events can lead to the presence of *pseudogenes,* which are quite similar to actual genes, but are not expressed. These pseudogenes present challenges for gene recognition algorithms, such as the one proposed in Searls chapter in this volume. Sexual reproduction adds another dimension to the exchange of genetic material. DNA from the two parents of a sexually reproducing organism undergoes a process called *crossover,* which forms a kind of mosaic that is passed on to the offspring.

Most mutations have relatively little effect. Mutations in the middle of introns generally have no effect at all (although mutations at the ends of an intron can affect the splicing process). Mutations in the third position of most codons have little effect at the protein level because of the redundancy of the genetic code. Even mutations that cause changes in the sequence of a protein are often neutral, as demonstrated by Sauer, *et al* (1989). Their experimental method involved *saturation mutagenesis* which explores are relatively large proportion of the space of possible mutations in parallel. Neutral mutations are the basis of genetic drift, which is the phenomena that accounts for the differences between the DNA that codes for functionally identical proteins in different organisms. This drift is also the basis for the molecular clock, described above. Of course, some point mutations are lethal, and others lead to diseases such as cystic fibrosis. Very rarely, a mutation will be advantageous; it will then rapidly get fixed in the population, as the organisms with the conferred advantage out reproduce the ones without it. Diploid sexually reproducing organisms have two copies of each gene (one from each parent), resulting in an added layer of complexity in the effect of mutations. Sometimes the extra copy can compensate (or partially compensate) for a mutation.

Molecular evolution also involves issues of selection and inheritance. Inheritance requires that the genes from the parent be passed to the offspring. DNA itself is replicated by splitting the double helix into two complimentary strands and then extending a primer by attaching complementary nucleotides. This process is modelled in detail Brutlag, *et al*'s chapter in this volume. The molecular mechanisms underlying the whole complex process of cell divi-

sion (i.e. the cell cycle) are strikingly conserved in eucaryotes, and knowledge about this process is growing rapidly (see, e.g., Hartwell (1991) for a review). Selection also occurs on factors that are only apparent on the molecular level, such as the efficiency of certain reaction pathways (see, e.g. Hochachka & Somero [1984]).

6. Sources of Biological Knowledge

The information in this chapter has been presented textbook style, with little discussion of how the knowledge arose, or where errors might have crept in. The purpose of this section is to describe some of the basic experimental methods of molecular biology. These methods are important not only in understanding the source of possible errors in the data, but also because computational methods for managing laboratory activities and analyzing raw data are another area where AI can play a role (see the chapters by Edwards, *et al* and Glasgow, *et al*, in this volume). I will also describe some of the many online information resources relevant to computational molecular biology that are available.

6.1 Model Organisms: Germs, Worms, Weeds, Bugs & Rodents

The investigation of the workings of even a single organism is so complex as to take many dedicated scientists many careers worth of time. Trying to study all organisms in great depth is simply beyond the abilities of modern biology. Furthermore, the techniques of biological experimentation are often complex, time consuming and difficult. Some of the most valuable methods in biological research are invasive, or require organisms to be sacrificed, or require many generations of observation, or observations on large populations. Much of this work is impractical or unethical to carry out on humans. For these reasons, biologists have selected a variety of model organisms for experimentation. These creatures have qualities that make possible controlled laboratory experiments at reasonable cost and difficulty with results that can often be extrapolated to people or other organisms of interest.

Of course, research involving humans can be done ethically, and in some areas of biomedical research, such as final drug testing, it is obligatory. Other research methods involve kinds of human cells can be grown successfully in the laboratory. Not many human cell types thrive outside of the body. Some kinds of human cancer cells do grow well in the laboratory, and these cells are an important vehicle for research.

Sometimes the selection of a new model organism can lead to great advances in a field. For example, the use of a particular kind of squid made possible the understanding of the functioning of neurons because it contained a motor neuron that is more than 10 times the size of most neural cells, and hence easy to find and use in experiments. There are experimentally useful

correlates of nearly every aspect of human biology found in some organism or another, but the following six organisms form the main collection of models used in molecular biology:

E. coli The ubiquitous intestinal bacterium *Escherichia coli* is a workhorse in biological laboratories. Because it is a relatively simple organism with fast reproduction time and is safe and easy to work with, *E. coli* has been the focus of a great deal of research in genetics and molecular biology of the cell. Although it is a Bacterium, many of the basic biochemical mechanisms of *E. coli* are shared by humans. For example, the first understanding of how genes can be turned on and off came from the study of a virus that infects these bacteria (Ptashne, 1987). *E. coli* is a common target for genetic engineering, where genes from other organisms are inserted into the bacterial genome then produced in quantity. *E. coli* is now the basis of the international biotechnology industry, churning out buckets full of human insulin, the heart attack drug TPA, and a wide variety of other substances.

Saccharomyces *Saccharomyces cervesiae* is better known as brewer's yeast, and it is another safe, easy to grow, short generation time organism. Other yeasts, such as *Schizosaccharomyces pombe*, are also used extensively. Surprisingly, yeasts are very much like people in many ways. Unlike the bacterium *E. coli*, yeasts are eucaryotes, with a cell nucleus, mitochondria, a eucaryotic cell membrane, and many of the other cellular components and processes found in most other eucaryotes, including people. Because these yeasts are so easy to grow and manipulate, and because they are so biochemically similar to people, many insights about the molecular processes involved in metabolism, biosynthesis, cell division, and other crucial areas of biology have come from the investigation of *Saccharomyces* (Saccharomyces is a genus name, which, when used alone, refers to all species that are within that genus). Yeasts play another important role in molecular biology. One of the crucial steps in sequencing large amounts of DNA is to be able to prepare many copies of moderate sized pieces of DNA. An widely used method for doing this is the *yeast artificial chromosome* (or YAC), which is discussed below.

Arabidopsis The most important application of increased biological understanding is generally thought to be in medicine, and increased understanding of human biology has indeed led to dramatic improvements in health care. However, in terms of effect on human life, agriculture is just as significant. A great deal of research into genetics and biochemistry has been motivated by the desire to better understand various aspects of plant biology. An important model organism for plants is *Arabidopsis thaliana*, a common weed. *Arabidopsis* makes a good model because it undergoes the same processes of growth, development, flowering and reproduction as most higher plants, but it's genome has 30 times less DNA than corn, and very little repetitive DNA. It also produces lots of seeds, and takes only about six weeks to grow to matu-

rity. There are several other model organisms used to investigate botanical questions, including tomatoes, tobacco, carrots and corn.

C. elegans One of the most exciting model organisms to emerge recently has been the nematode worm *Caenorhabditis elegans.* This tiny creature, thousands of which can be found in a spadeful of dirt, has already been used to generate tremendous insight about cellular development and physiology. The adult organism has exactly 959 cells, and every normal worm consists of exactly the same collection of cells in the same places doing the same thing. It is one of the simplest creatures with a nervous system (which involves about a third of its cells). Not only is the complete anatomy of the organism known, but a complete cell fate map has been generated, tracing the developmental lineage of each of each cell throughout the lifespan of the organism. This map allows researchers to relate behaviors to particular cells, to trace the effects of genetic mutations very specifically, and perhaps to gain insight into the mechanisms of aging as well as development. A large, highly integrated picture and text database of information about the cell fates, genetic maps and sequences, mutation effects and other relevant information about *C. elegans* is currently under construction at the University of Arizona.

D. melanogaster *Drosophila melanogaster,* a common fruit fly, has long been a staple of classical genetics research. These flies have short generation times, and many different genetically determined morphological characteristics (e.g. eye color) that can readily be determined by visual inspection. *Drosophila* were used for decades in exploring patterns of inheritance; now that molecular methods can be applied, they have proven invaluable for a variety of studies of genetic expression and control. An important class of genetic elements that regulate many other genes, in effect, specifying complex genetic programs, were first discovered in *Drosophila*; these areas are called *homeoboxes*. Molecular genetics in *Drosophila* is also providing great insights into how complex body plans are generated.

M. musculus *Mus musculus* is the basic laboratory mouse. Mice are mammals, and, as far as biochemistry is concerned, are practically identical to people. Many questions about physiology, reproduction, functioning of the immune and nervous systems and other areas of interest can only be addressed by examining creatures that are very similar to humans; mice nearly always fit the bill. The similarities between mice and people mean also that the mouse is a very complicated creature; it has a relatively large, complex genome, and mouse development and physiology is not as regular or consistent as that of *C. elegans* or *Drosophila.* Although our depth of understanding of the mouse will lag behind understanding of simpler organisms, the comparison of mouse genome to human is likely to be a key step, both in understanding their vast commonalities, and in seeing the aspects of our genes that make us uniquely human.

7. Experimental Methods

Molecular biologists have developed a tremendous variety of tools to address questions of biological function. This chapter can only touch briefly on a few of the most widely used methods, but the terminology and a sense of the kinds of efforts required to produce the data used by computer scientists can be important for understanding the strengths and limitations of various sources of data.

Imaging. The first understanding of the cellular nature of life came shortly after the invention of the light microscope, and microscopy remaines central to research in biology The tools for creating images have expanded tremendously. Not only are there computer controled light microscopes with a wide variety of imaging modalities, but there are now many other methods of generating images of the very small. The electon microscope offers extremely high resolution, although it requires exposing the imaged sample to high vacuum and other harsh treatments. New technologies including the Atomic Force Microscope (AFM) and the Scanning Tunnelling Microscope (STM) offer the potential to create images of individual molecules. Biologists use these tools extensively.

Gel Electrophoresis. A charged molecule, when placed in an electric field, will be accelerated; positively charged molecules will move toward negative electrodes and vice versa. By placing a mixture of molecules of interest in a medium and subjecting them to an electric charge, the molecules will migrate through the medium and separate from each other. How fast the molecules will move depends on their charge and their size—bigger molecules see more resistance from the medium. The procedure, called *electrophoresis* involves putting a spot of the mixture to be analyzed at the top of a polyacrylamide or agarose gel, and applying an electric field for a period of time. Then the gel is stained so that the molecules become visible; the stains appear as stripes along the gel, and are called *bands.* The location of the bands on the gel are proportional to the charge and size of the molecules in the mixture (see Figure 4 for an example). The intensity of the stain is an indication of the amount of a particular molecule in the mixture. If the molecules are all the same charge, or have charge proportional to their size (as, for example, DNA does) then electrophoresis separates them purely by size.

Often, several mixtures are run simultaneously on a single gel. This allows for easy calibration to standards, or comparison of the contents of different mixtures, showing, for example, the absence of a particular molecular component in one. The adjacent, parallel runs are sometimes called *lanes.* A variation on this technique allows the sorting of molecules by a chemical property called the *isoelectric point,* which is related to its pK. A combination of the two methods, called *2D electrophoresis* is capable of very fine

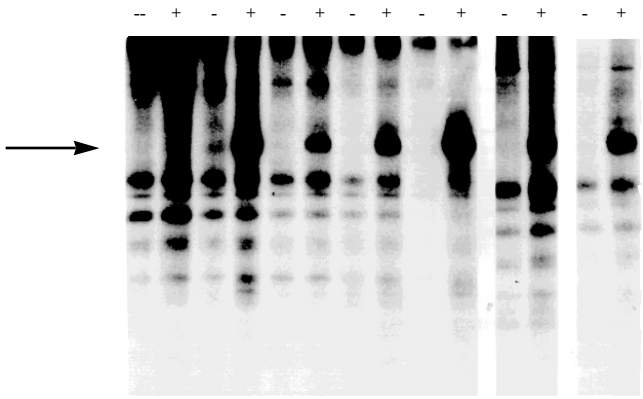

Figure 4. This is an example of a gel electrophoresis run.. Each column was loaded with a different mixture. The mixtures are then separated vertically by their charge and size. The gel is then stained, producing dark bands where a molecule of a given size or charge is present in a mixture. In this gel, the columns marked with a - are a control group. The band marked with an arrow is filled only in the + columns.

distinctions, for example, mapping each protein in a cell to a unique *spot* in two-space, the size of the spot indicating the amount of the protein. Although there are still some difficulties in calibration and repeatability, this method is potentially a very powerful tool for monitoring the activities of large biochemical systems. In addition, if a desired molecule can be separated from the mixture this way, individual spots or bands can be removed from the gel for further processing, in a procedure called *blotting.*

Cloning. A group of cells with identical genomes are said to be *clones* of one another. Unless there are mutations, a single cell that reproduces asexually will produce identical offspring; these clones are sometimes called a *cell line*, and certain standardized cell lines, for example the HeLa cell line, play an important role in biological research.

This concept has been generalize to cloning individual genes. In this case, a piece of DNA containing a gene of interest is inserted into the genome of a target cell line, and the cells are screened so that all of the resulting cells have an identical copy of the desired genetic sequence. The DNA in these cells is said to be *recombinant*, and the cell will produce the protein coded for by the inserted gene.

Cloning a gene requires some sophisticated technology. In order for a cloned gene to be expressed, it must contain the appropriate transcription signals for the target cell line. One way biologists ensure that this will happen is to put the new gene into a bacteriophage (a virus that infects bacteria), or a plasmid (a circular piece of DNA found outside of the chromosome of bacteria that replicates independently of the bacteria's chromosomal DNA). These

devices for inserting foreign DNA into cells are called *vectors.*

In order to cut and paste desired DNA fragments into vectors, biologists use *restriction enzymes*, which cut DNA at precisely specified points. These enzymes are produced naturally by bacteria as a way of attacking foreign DNA. For example, the commonly used enzyme *EcoRI* (from *E. coli)* cuts DNA between the G and the A in the sequence GAATTC; these target sequences are called *restriction sites*. Everywhere a restriction site occurs in a DNA molecule treated with *EcoRI*, the DNA will be broken. Restriction enzymes play many roles in biology in addition to making gene cloning possible; a few others will be described below.

Both the insertion of the desired gene into the vector and the uptake of the vector by the target cells are effective only a fraction of the time. Fortunately, cells and vectors are small and it is relatively easy to grow a lot of them. The process is applied to a population of target cells, and then the resulting population is screened to identify the cells where the gene was successfully inserted. This can be difficult, so many vectors are designed to facilitate screening. One popular vector, *pBR322*, contains a naturally occurring transcription start signal and some antibiotic resistance genes, designed with conveniently placed restriction sites. If this vector is taken up by the target cells, it will confer resistance to certain antibiotics to them. By applying the anitbiotic to the whole colony, the researcher can kill all the cells that did not get the cloned gene. More sophisticated manipulations involving multiple antibiotic resistances and carefully placed restriction sites can also be used to ensure that the gene was correctly taken up by the vector.

There are many variations on these techniques for inserting foreign genes. It is now possible to use simple bacteria to produce large amounts of almost any isolated protein, including, for example, human insulin. Although it is a more complex process, it is also possible to insert foreign genes into plants and animals, even people. A variety of efforts are underway to use these techniques to engineer organisms for agriculture, medicine and other applications. Not all of these applications are benign. One of the most successful early efforts was to increase the resistance of tobacco plants to pesticides, and there are clear military applications. On the other hand, these methods also promise new approaches to producing important rare biological compounds inexpensively (e.g. for novel cancer treatments or cleaning up toxic waste) and improving the nutritional value or hardiness of agricultural products. The entire field of genetic engineering is controversial, and there are a variety of controls on what experiments can be done and how they can be done.

Hybridization and Immunological Staining. Biological compounds can show remarkable specificity, for example, binding very selectively only to one particular compound. This ability plays an important role in the laboratory, where researchers can identify the presence or absence of a particular molecule (or even a region of a molecule) in vanishingly small amounts.

Antibodies are the molecules that the immune system uses to identify and fight off invaders. Antibodies are extremely specific, recognizing and binding to only one kind of molecule. Dyes can be attached to the antibody, forming a very specific system for identifying the presence (and possibly quantifying the amount) of a target molecule that is present in a system.

There is a conceptually related method for identifying very specifically the presence of a particular nucleotide sequence in a macromolecule. The complement to a single-stranded DNA sequence will bind quite specifically to that sequence. One technique measures how similar two related DNA sequences are by testing how strongly the single-stranded versions of the molecules stick to each other, or *hybridize*. The more easily they come apart, the more differences there are between their sequences. It is also possible to attach a dye or other marker to a specific piece of DNA (called a *probe*) and then hybridize it to a longer strand of DNA. The location along the strand that is complementary to the probe will then be marked. There are many variations on hybridization and immunological staining that are customized to the needs of a particular experiment.

Gene Mapping and Sequencing. The Human Genome Project is the effort to produce a map and then the sequence of the human genome. The purpose of a genetic map is to identify the location and size of all of the genes of an organism on its chromosomes. This information is important for a variety of reasons. First, because crossover is an important component of inheritance in sexually reproducing organisms, genes that are near each other on the chromosome will tend to be inherited together. In fact, this forms the basis for *linkage analysis*, which is a technique that looks at the relationships between genes (or phenotypes) in large numbers of matings (in this context, often called *crosses*) to identify which genes tend to be inherited together, and are therefore likely to be near each other. Second, it is possible to clone genes of known locations, opening up a wide range of possible experimental manipulations. Finally, it is currently possible to determine the sequence of moderate size pieces of DNA, so if an important gene has been mapped, it is possible to find the sequence of that area, and discover the protein that is responsible for the genetic characteristic. This is especially important for understanding the basis of inherited diseases.

The existence of several different kinds of restriction enzymes makes possible a molecular method of creating genetic maps. The application of each restriction enzyme (the process is called a *digest*) creates a different collection of *restriction fragments* (the cut up pieces of DNA). By using gel electrophoresis, it is possible to determine the size of these fragments. Using multiple enzymes, together and separately, results in sets of fragments which can be (partially) ordered with respect to each other, resulting in a genetic map. AI techniques for reasoning about partial orders have been effectively applied to the problem of assembling the fragments into a map (Letovsky &

Berlyn, 1992). These *physical maps* divide a large piece of DNA (like a chromosome) into parts, and and there is an associated method for obtaining any desired part.

Restriction fragment mapping becomes problematic when applied to large stretches of DNA, because the enzymes can produce many pieces of about the same size, making the map ambiguous. The use of different enzymes can help address this problem to a limited degree, but a variety of other techniques are now also used.

Being able to divide the genome into moderate sized chunks is a prerequisite to determining its sequence. Although there are several clever methods for determining the sequence of DNA molecule, all of them are limited to a resolution of well under a thousand basepairs at a time. In order to take this sequencing ability and determine the sequence of large pieces of DNA, many different overlapping chunks must be sequenced, and then these sequences must be assembled. In order to accomplish this task, it is necessary to break the DNA in an entire genome down into a set of more manageable sized pieces. The ordering of these pieces must be known (so they can be reassembled into a complete sequence), taken together the pieces must cover the entire genome, and the same set of pieces must be accessible to many different laboratories. This process is usually accomplished in several stages. The first stage generates relatively large pieces called *contigs*. Contigs are maintained in cloned cell lines so that they can be reproduced and distributed. Often, these pieces of DNA are made into *Yeast artificial chromosomes*, or *YACs*, which can hold up to about a million basepairs of sequence each, requiring on the order of 10,000 clones to adequately cover the entire human genome. Each of these is then broken down into sets of smaller pieces, often in the form of *cosmids*. A cosmid is a particular kind of bacteriophage (a virus that infects bacteria) that is capable of accepting inserts of 30,000 or so basepairs. The difficulties in generating and maintaining collections of clones that large have led to alternative technologies for large scale sequencing.

One alternative involves a new technology based on the *polymerase chain reaction*, or *PCR*. This mechanism was revolutionary becauase it made it possible to rapidly produce huge amounts of a specific region of DNA, simply by knowing a little bit of the sequence around the desired region. PCR exponentially *amplifies* (makes copies of) a segment of a DNA molecule, given a unique pair of sequences that bracket the desired piece. First, short sequences of DNA (called *oligonucleotides*, or *oligos*) complementary to

*There are many interesting uses of this technology. For example, it gives law enforcement the ability to generate enough DNA for identification from vanishing small samples of tissue. A more amusing application is the rumored use of PCR to spy on what academic competitors are doing in their research. Almost any correspondence from a competitor's lab will contain traces of DNA which can be amplified by PCR to identify the specific clones the lab is working with.

each of the bracketing sequences are synthesized. Creating short pieces of DNA with a specific sequence is routine technology, now often performed by laboratory robots. These pieces are called *primers.* The primers, the target DNA and the enzyme DNA polymerase are then combined. The mixture is heated, so that the hydrogen bonds in the DNA break and the molecule splits into two single strands. When the mixture cools sufficiently, the primers bond to the regions around the area of interest, and the DNA polymerase replicates the DNA downstream of the primers. By using a heat resistant polymerase from an Archaea species that lives at high temperatures, it is possible to rapidly cycle this process, doubling the amount of desired segment of DNA each time. This technology makes possible the exponential amplification of entire DNA molecules or any specific region of DNA for which bracketing primers can be generated.*

In order to use PCR for genome mapping and sequencing, a collection of unique (short) sequences spread throughout the genome must be identified for use as primers. The sequences must be unique in the genome so that the source of amplified DNA is unambiguous, and they have to be relatively short so that they are easy to synthesize. The sites in the genome that correspond to these sequences are called *sequence tagged sites* or *STSs.* The more STSs that are known, the finer grained the map of the genome they provide. Finding short, unique sequences even in $3x10^9$ bp of DNA is not that difficult; a simple calculation shows that most sequences of length 16 or so can reasonably be expected to be unique in a genome of that size. An early goal of the Human Genome Project is to generate a list of STSs spaced at approximately 100kbp intervals over the entire human genome. If it is possible to find STSs that adequately cover the genome, it will not be necessary to build and maintain libraries of 10,000 YACs and ten times as many cosmids. Any region of DNA of interest can be identified by two STSs that bracket it. Instead of having to maintain large clone collections, these STSs can be stored in a database, and any researcher who needs a particular section of DNA can synthesize the appropriate primers and use PCR to produce many copies of just of that section.

Another issue that has been raised about the project to sequence the genome is the need to know the sequences of all of the introns and other noncoding regions of DNA. One way to address this issue is to target only coding regions for sequencing. The ability to find the sequences that a particular cell is using to produce proteins at a particular point in time is also useful in a variety of other areas as well. This information can be gleaned by gathering the mRNAs present in the cytoplasm of the cell, and sequencing them. Instead of sequencing the mRNAs directly, biologists use an enzyme called *reverse transcriptase* to make DNA molecules complementary to the mRNAs (called *cDNA*s) and then sequence that DNA. Using PCR and other technology, it is possible to capture at least portions of most of the mRNAs a cell is

producing. By sequencing these cDNAs, researchers can focus their attention on just the parts of the genome that code for expressed proteins.

Large scale attempts to sequence at least part of all of the cDNAs that can be produced from brain tissue have resulted in partial sequences for more than 2500 new proteins in a very short period of time (Adams, *et al*, 1992). These sequences called ESTs, for *expressed sequence tags* can be used as PCR primers for future, more detailed experiments. This work has created controversy because of the ensuing attempt by the National Institutes of Health to patent the EST sequences.

Crystallography and NMR. Until the relationship between protein sequence and structure is more fully understood, the sequences produced by genome projects will provide only part of the biochemical story. Additional information about protein structure is necessary to understand how the proteins function. This structural information is at the present primarily gathered by *X-ray crystallography*. In order to determine the structure of a protein in this manner, a very large, pure crystal of the protein must be grown (this process can take years, and may never succeed for certain proteins). Then the X-ray diffraction pattern of the crystal is measured, and this information can be used indirectly to determine the positions of the atoms in the molecule. Glasgow, *et al's* chapter in this volume describes this process in more detail. Because of the difficulties in crystallography, relatively few structures are known, but the number of new structures is growing exponentially, with a doubling time of a bit over two years.

A promising alternative to crystallography for determining protein structure is multi-dimensional *nuclear magnetic resonance*, or *NMR*. Although this process does not require the crystallization of the protein, there are technical difficulties in analyzing the data associated with large molecules like proteins. *Edwards, et al*'s chapter in this volume describes some of the challenges. Both crystallography and NMR techniques result in static protein structures, which are to some degree misleading. Proteins are flexible, and the patterns of their movement are likely to play an important role in their function. Although NMR has the potential to provide information about this facet of protein activity, there is very little data available currently.

7.1 Computational Biology

In the last five years, biologists have come to understand that sharing the results of experiments now takes more than simple journal publication. In the 1980s, many journals were overwhelmed with papers reporting novel sequences and other biological data. Paper publications of sequences are hard to analyze, prone to typographical errors, and take up valuable journal space.

*Researchers without internet access can contact NCBI by writing to NCBI/National Library of Medicine/Bethesda, MD 20894 USA or calling +1 (301) 496-2475.

Databases were established, journals began to require deposition into the databases before publication, and various tools began to appear for managing and analyzing the databases.

When Doolittle, et al (1983) used the nascent genetic sequence database to prove that a cancer causing gene was a close relative of a normal growth factor, molecular biology labs all over the world began installing computers or linking up to networks to do database searches. Since then, a bewildering variety of computational resources for biology have arisen. These databases and other resources are a valuable service not only to the biological community, but also to the computer scientist in search of domain information.

There is a database of databases, listing these resources which is maintained at Los Alamos National Laboratory. It is called LiMB(Lawton, Burks & Martinez, 1989), and contains descriptions, contacts and access methods for more than 100 molecular biology related databases. It is a very valuable tool for tracking down information. Another general source for databases and information about them is the National Center for Biotechnology Information (NCBI), which is part of the National Library of Medicine. Many databases are available via anonymous ftp from the NCBI server, ncbi.nlm.nih.gov.*

A few of the databases that may be of particular interest to computer scientists are described here. There are several databases that maintain genetic sequences, and they are increasingly coordinated. They are Genbank (Moore, Benton & Burks, 1990), the European Molecular Biology Laboratory nucleotide sequence database (EMBL) (Hamm & Cameron, 1986), and the DNA Database, Japan (DDBJ) (Miyazawa, 1990). NCBI will also provide a sequence database beginning in 1992. The main protein sequence database is the Protein Identification Resource (PIR) (George, Barker & Hunt, 1986). NCBI also provides a non-redundant combination of protein sequences from various sources (including translations of genetic sequences) in its NRDB.

Several databases contain information about three dimensional structures of molecules. The Protein Data Bank (PDB) maintained by Brookhaven National Laboratory, contains protein structure data, primarily from crystallographic data. BioMagRes (BMR) is a database of NMR derived data about proteins, including three dimensional coordinates, that is maintained at the University of Wisconsin, Madison (Ulrich, Markley & Kyogoku, 1989). CARBBANK, contains structural information for complex carbohydrates (Doubet, Bock, Smith, Albersheim & Darvill, 1989). Chemical Abstracts Service (CAS) Online Registry File is a commercial database that contains more than 10 million chemical substances, many with three dimensional coordinates and other useful information. The Cambridge Structural Database contains small molecule structures, and is available to researchers at moderate charge.

Genetic map databases (GDB), as well as a database of inherited human diseases and characteristics (OMIM) are maintained at the Welch Medical

Library at Johns Hopkins University. To get access to these databases, send email to help@welch.jhu.edu. Other genetic map databases are available for many of the model organisms listed above; consult LiMB for more information about them.

There is a database of information about compounds involved in intermediary metabolism called CompoundKB, developed by Peter Karp that is available from NCBI. This database is available in KEE knowledge base form as well as several others, and there is associated LISP code which makes it attractive for artificial intelligence researchers; see Karp's and Mavrovouniotis's chapters in this volume for possible applications of the knowledge base.

Finally, one of the most important computer-based assets for a computer scientist interested in molecular biology information is the bulletin board system called *bionet.* This bboard is available through usenet as well as by electronic mail. The discussion groups include computational biology, information theory and software, as well as more than 40 other areas. Bionet is an excellent source for information and contacts with computationally sophisticated biologists.

8. Conclusion

AI researchers have often had unusual relationships with their collaborators. "Experts" are somehow "knowledge engineered" so that what they know can be put into programs. Biology has a long history of collaborative research, and it does not match this AI model. Computer scientists and biologists often have differing expectations about collaboration, education, conferences and many other seemingly mundane aspects of research. In order to work with biologists, AI researchers must understand a good deal about the domain and find ways to bridge the gap between these rather different scientific cultures.

This brief survey of biology is intended to help the computer scientist get oriented and understand some of the commonly used terms in the domain. Many more detailed, but still accessible books are listed in the references. I find this material fascinating. Not only is it interesting as a domain for AI research, but it provides a rich set of metaphors for thinking about intelligence: genetic algorithms, neural networks and Darwinian automata are but a few of the computational approaches to behavior based on biological ideas. There will, no doubt, be many more.

Acknowledgements

This chapter was written at the instigation of Harold Morowitz, who understood long before I did that such an introduction could indeed be accom-

plished in less than 1000 pages. He also taught the biochemistry course that I finally took, two years *after* finishing my Ph.D. David J. States deserves much of the credit as well. In the three years we have been working together, he greatly extended my understanding of not only what biologists know, but how they think. He has read several drafts of this chapter and made helpful suggestions. David Landsman, Mark Boguski, Kalí Tal and Jill Shirmer have also read the chapter and made suggestions. Angel Lee graciously supplied the gel used in Figure 4. Of course, all remaining mistakes are my responsibility.

References

Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. & Venter, J. C. (1992). Sequence Identification of 2,375 Brain Genes. *Nature, 355*(6361), 632-4.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. (1989). *The Molecular Biology of the Cell* (2nd. ed. ed.). New York, NY: Garland Publishing.

Buss, L. (1987). *The Evolution of Individuality*. Princeton, NJ: Princeton University Press.

Cherty, L. M., Case, S. M. & Wilson, A. C. (1978). Frog Perspectives on the Morphological difference between Humans and Chimpanzees. *Science, 200*, 209-211.

Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A., & Antoniades, H. N. (1983). Simian Sarcoma *Onc* Gene, *v-sis*, Is Derived from the Gene (or Genes) Encoding Platelet Derived Growth Factor. *Science*, 221, 275-277.

Doubet, S., Bock, K., Smith, D., Albersheim, P. & Darvill, A. (1989). The Complex Carbohydrate Structure Database. *Trends in Biochemical Sciences, 14*, 475.

George, D., Barker, W. & Hunt, L. (1986). The Protein Identification Resource. *Nucleic Acids Research, 14*, 11-15.

Hamm, G. & Cameron, G. (1986). The EMBL Data Library. *Nucleic Acids Research, 14*, 5-9.

Hartwell, L. (1991). Twenty-five Years of Cell Cycle Genetics. *Genetics, 129*(4), 975-980.

Hochachka, P. W. & Somero, G. N. (1984). *Biochemical Adaptation*. Princeton, NJ: Princeton University Press.

Karplus, M. & Petsko, G. A. (1990). Molecular Dynamics Simulations in Biology. *Nature, 347*(October), 631-639.

Kessel, R. G. & Kardon, R. H. (1979). *Tissues and Organs: A Text-Atlas of Scanning electron Microscopy*. San Francisco, CA: W.H. Freeman and Company.

Langton, C., eds. (1989). *Artificial Life* (VI.). Redwood City, CA: Addison Wesley.

Lawton, J., Burks, C. & Martinez, F. (1989). Overview of the LiMB Database. *Nucleic Acids Research, 17*, 5885-5899.

Letovsky, S. & Berlyn, M. (1992). CPROP: A Rule-based Program for Constructing a Genetic Map. *Genomics, 12*, 435-446.

Li, W.-H. & Graur, D. (1991). *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates, Inc.

Margolis, L. (1981). *Symbiosis and Cell Evolution*. San Francisco: Freeman.

May, R. M. (1988). How Many Species Are There on Earth? *Science, 241*(September 16), 1441-1450.

Miyazawa, S. (1990). DNA Data Bank of Japan: Present Status and Future Plans. *Computers and DNA, 7*, 47-61.

Moore, J., Benton, D. & Burks, C. (1990). The GenBank Nucleic Acid Data Bank. *Focus, 11*(4), 69-72.

Ptashne, M. (1987). *A Genetic Switch: Gene Control and the Phage Lambda*. Palo Alto, CA: Blackwell Scientific Publications.

Sauer, R. T. (1989). Genetic Analysis of Protein Stability and Function. *Annual Review of Genetics, 23*, 289-310.

Ulrich, E., Markley, J. & Kyogoku, Y. (1989). Creation of Nuclear Magnetic Resonance Data Repository and Literature Base. *Protein Sequence and Data Analysis, 2*, 23-37.

Woese, C. R., Kandler, O. & Wheelis, M. L. (1990). Towards a Natural System of Organisms: Proposal for the Domains Archaia, Bacteria, and Eucarya. (June), 4576-4579.