1 Bioinformatics and Mathematics

Traditionally, the study of biology is from morphology to cytology and then to the atomic and molecular level, from physiology to microscopic regulation, and from phenotype to genotype. The recent development of bioinformatics begins with research on genes and moves to the molecular sequence, then to molecular conformation, from structure to function, from systems biology to network biology, and further investigates the interactions and relationships among, genes, proteins, and structures. This new reverse paradigm sets a theoretical starting point for a biological investigation. It sets a new line of investigation with a unifying principle and uses mathematical tools extensively to clarify the ever-changing phenomena of life quantitatively and analytically.

It is well known that there is more to life than the genomic blueprint of each organism. Life functions within the natural laws that we know and those that we do not know. Life is founded on mathematical patterns of the physical world. Genetics exploits and organizes these patterns. Mathematical regularities are exploited by the organic world at every level of form, structure, pattern, behavior, interaction, and evolution. Essentially all knowledge is intrinsically unified and relies on a small number of natural laws. Mathematics helps us understand how monomers become polymers necessary for the assembly of cells. Mathematics can be used to understand life from the molecular to the biosphere levels, including the origin and evolution of organisms, the nature of genomic blueprints, and the universal genetic code as well as ecological relationships.

Mathematics and biological data have a synergistic relationship. Biological information creates interesting problems, mathematical theory and methods provide models for understanding them, and biology validates the mathematical models. A model is a representation of a real system. Real systems are too complicated, and observation may change the real system. A good system model should be simple, yet powerful enough to capture the behavior of the real system. Models are especially useful in bioinformatics. In this chapter we provide an overview of bioinformatics history, genetic code and mathematics, background mathematics for bioinformatics, and the big picture of bioinformatics–informatics.

Mathematics of Bioinformatics: Theory, Practice, and Applications, By Matthew He and Sergey Petoukhov

1.1 INTRODUCTION

Mendel's Genetic Experiments and Laws of Heredity The discovery of genetic inheritance by Gregor Mendel back in 1865 was considered as the start of bioinformatics history. He did experiments on the cross-fertilization of different colors of the same species. Mendel's genetic experiments with pea plants took him eight years (1856–1863). During this time, Mendel grew over 10,000 pea plants, keeping track of progeny number and type. He recorded the data carefully and performed mathematical analysis of the data. Mendel illustrated that the process of inheritance of traits could be explained more easily if it was controlled by factors passed down from generation to generation. He concluded that genes come in pairs. Genes are inherited as distinct units, one from each parent. He also recorded the segregation of parental genes and their appearance in the offspring as dominant or recessive traits. He published his results in 1865. He recognized the mathematical patterns of inheritance from one generation to the next. Mendel's laws of heredity are usually stated as follows:

- *The law of segregation.* A gene pair defines each inherited trait. Parental genes are randomly separated by the sex cells, so that sex cells contain only one gene of the pair. Offspring therefore inherit one genetic allele from each parent.
- *The law of independent assortment.* Genes for different traits are sorted from one another in such a way that the inheritance of one trait is not dependent on the inheritance of another.
- *The law of dominance.* An organism with alternate forms of a gene will express the form that is dominant.

In 1900, Mendel's work was rediscovered independently by DeVries, Correns, and Tschermak, each of whom confirmed Mendel's discoveries. Mendel's own method of research is based on the identification of significant variables, isolating their effects, measuring these meticulously, and eventually subjecting the resulting data to mathematical analysis. Thus, his work is connected directly to contemporary theories of mathematics, statistics, and physics.

Origin of Species Charles Darwin published *On the Origin of Species by Means of Natural Selection* (Darwin, 1859) or "The Preservation of Favored Races in the Struggle for Life." His key work was that evolution occurs through the selection of inheritance and involves transmissible rather than acquired characteristics between individual members of a species. Darwin's landmark theory did not specify the means by which characteristics are inherited. The mechanism of heredity had not been determined at that time.

First Genetic Map In 1910, after the rediscovery of Mendel's work, Thomas Hunt Morgan at Columbia University carried out crossing experiments with

the fruit fly (*Drosophila melanogaster*). He proved that the genes responsible for the appearance of a specific phenotype were located on chromosomes. He also found that genes on the same chromosome do not always assort independently. Furthermore, he suggested that the strength of linkage between genes depended on the distance between them on the chromosome. That is, the closer two genes lie to each other on a chromosome, the greater the chance that they will be inherited together. Similarly, the farther away they are from each other, the greater the chance of that they will be separated in the process of crossing over. The genes are separated when a crossover takes place in the distance between the two genes during cell division. Morgan's experiments also lead to *Drosophila*'s unusual position as, to this day, one of the best studied organisms and most useful tools in genetic research. In 1911, Alfred Sturtevant, then an undergraduate researcher in the laboratory of Thomas Hunt Morgan, mapped the locations of the fruit fly genes, creating the first genetic map ever made.

Transposable Genetic Elements In 1944, Barbara McClintock discovered that genes can move on a chromosome and can jump from one chromosome to another. She studied the inheritance of color and pigment distribution in corn kernels at the Carnegie Institution Department of Genetics in Cold Spring Harbor, New York. At age 81 she was awarded a Nobel prize. It is believed that transposons may be linked to such genetic disorders as hemophilia, leukemia, and breast cancer; and transposons may have played a crucial role in evolution.

DNA Double Helix In 1953, James Watson and Francis Crick proposed a double-helix model of DNA. DNA is made of three basic components: a sugar, an acid, and an organic "base." The base was always one of the four nucleotides: adenine (A), cytosine (C), guanine (G), or thymine (T). These four different bases are categorized in two groups: purines (adenine and guanine) and pyrimidines (thymine and cytosine). In 1950, Erwin Chargaff found that the amounts of adenine (A) and thymine (T) in DNA are about the same, as are the amounts of guanine (G) and cytosine (C). These relationships later became known as "Chargaff's rules" and led to much speculation about the three-dimensional structure that DNA would have. Rosalind Franklin, a British chemist, used the x-ray diffraction technique to capture the first high-quality images of the DNA molecule. Franklin's colleague Maurice Wilkins showed the pictures to James Watson, an American zoologist, who had been working with Francis Crick, a British biophysicist, on the structure of the DNA molecule. These pictures gave Watson and Crick enough information to propose in 1953 a double-stranded, helical, complementary, antiparallel model for DNA. Crick, Watson, and Wilkins shared the 1962 Nobel Prize in Physiology or Medicine for the discovery that the DNA molecule has a double-helical structure. Rosalind Franklin, whose images of DNA helped lead to the discovery, died of cancer in 1958 and, under Nobel rules, was not eligible for the prize.

In 1957, Francis Crick and George Gamov worked out the "central dogma," explaining how DNA functions to make protein. Their *sequence hypothesis* posited that the DNA sequence specifies the amino acid sequence in a protein. They also suggested that genetic information flows only in one direction, from DNA to messenger RNA to protein, the central concept of the central dogma.

Genetic Code *(see Appendix A)* The genetic code was finally "cracked" in 1966. Marshall Nirenberg, Heinrich Mathaei, and Severo Ochoa demonstrated that a sequence of three nucleotide bases, a codon or triplet, determines each of the 20 amino acids found in nature. This means that there are 64 possible combinations ($4^3 = 64$) for 20 amino acids. They formed synthetic messenger ribonucleic acid (mRNA) by mixing the nucleotides of RNA with a special enzyme called polynucleotide phosphorylase. This resulted in the formation of a single-stranded RNA in this reaction. The question was how these 64 genetic codes could code for 20 different amino acids. Nirenberg and Matthaei synthesized poly(U) by reacting only uracil nucleotides with the RNA-synthesizing enzyme, producing –UUUU–. They mixed this poly(U) with the protein-synthesizing machinery of *Escherichia coli* in vitro and observed the formation of a protein. This protein turned out to be a polypeptide of phenylalanine. They showed that a triplet of uracil must code for phenylalanine. Philip Leder and Nirenberg found an even better experimental protocol to solve this fundamental problem. By 1965 the genetic code was solved almost completely. They found that the "extra" codons are merely redundant: Some amino acids have one or two codons, some have four, and some have six. Three codons (called *stop codons*) serve as stop signs for RNA-synthesizing proteins.

First Recombinant DNA Molecules In 1972, Paul Berg of Stanford University created the first recombinant DNA molecules by combining the DNA of two different organisms. He used a restriction enzyme to isolate a gene from a human-cancer-causing monkey virus. Then he used lipase to join the section of virus DNA with a molecule of DNA from the bacterial virus lambda, creating the first recombinant DNA molecule. He realized the risks of his experiment and terminated it temporarily before the recombinant DNA molecule was added to *E. coli*, where it would have quickly been reproduced. He proposed a one-year moratorium on recombinant DNA studies while safety issues were addressed. Berg later resumed his studies of recombinant DNA techniques and was awarded the 1980 Nobel Prize in Chemistry. His experiments paved the road for the field of genetic engineering and the modern biotechnology industry.

DNA Sequencing and Database In early 1974, Frederick Sanger from the UK Medical Research Council was first to invent DNA-sequencing techniques. During his experiments to uncover the amino acids in bovine insulin, he developed the basics of modern sequencing methods. Sanger's approach

involved copying DNA strands, which would show the location of the nucleotides in the strands. To apply Sanger's approach, scientists had to analyze the composite collections of DNA pieces detected from four test tubes, one for each of the nucleotides found in DNA (adenosine, cytosine, thymidine, guanine). Then they needed to be arranged in the correct order. This technique is very slow and tedious. It takes many years to sequence only a few million letters in a string of DNA. Almost simultaneously, the American scientists Alan Maxam and Walter Gilbert were creating a different method called the *cleavage method*. The base for virtually all DNA sequencing was the dideoxy-chain-terminating reaction developed by Sanger.

In 1978, David Botstein developed restriction-fragment-length polymorphisms. Individual human beings differ one base pair in every 500 nucleotides or so. The most interesting variations for geneticists are those that are recognized by certain enzymes called *restriction enzymes*. Each of these enzymes cuts DNA only in the presence of a specific sequence (e.g., GAATTC in the case of the restriction enzyme EcoR1). This sequence is called a *restriction site*. The enzyme will bypass the region if it has mutated to GACTTC. Thus, when a specific restriction enzyme cuts the DNA of different people, it may produce fragments of different lengths. These DNA fragments can be separated according to size by making them move through a porous gel in an electric field. Since the smaller fragments move more rapidly than the larger ones, their sizes can be determined by examining their positions in the gel. Variations in their lengths are called *restriction-fragment-length polymorphisms*.

In 1980, Kary Mullis invented polymerase chain reaction (PCR), a method for multiplying DNA sequences in vitro. The purpose of PCR is to make a huge number of copies of a specific DNA fragment, such as a gene. Use of thermostable polymerase allows the dissociation of newly formed complementary DNA and subsequent annealing or hybridization of the primers to the target sequence with a minimal loss of enzymatic activity. PCR may be necessary to receive enough starting template for instance sequencing.

In 1986, scientists presented a means of detecting ddNTPs with fluorescent tags, which required only a single test tube instead of four. As a result of this discovery, the time required to process a given batch of DNA was reduced by one-fourth. The amount of sequenced base pairs increased rapidly from there on.

Established in 1988 as a national resource for molecular biology information, the National Center for Biotechnology Information (NCBI) carries out diverse responsibilities. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information: all for a better understanding of molecular processes affecting human health and disease. NCBI conducts research on fundamental biomedical problems at the molecular level using mathematical and computational methods.

The European Bioinformatics Institute (EBI) is a nonprofit academic organization that forms part of the European Molecular Biology Laboratory

(EMBL). The roots of the EBI lie in the EMBL Nucleotide Sequence Data Library, which was established in 1980 at the EMBL laboratories in Heidelberg, Germany and was the world's first nucleotide sequence database. The original goal was to establish a central computer database of DNA sequences rather than having scientists submit sequences to journals. What began as a modest task of abstracting information from literature soon became a major database activity with direct electronic submissions of data and the need for a highly skilled informatics staff. The task grew in scale with the start of the genome projects, and grew in visibility as the data became relevant to research in the commercial sector. It became apparent that the EMBL Nucleotide Sequence Data Library needed better financial security to ensure its long-term viability and to cope with the sheer scale of the task.

Human Genome Project In 1990, the U.S. Human Genome Project started as a 15-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the expected completion date to 2003. Project goals were to:

- Identify all the genes in human DNA
- Determine the sequences of the 3 billion chemical base pairs that make up human DNA
- Store this information in databases
- Improve tools for data analysis
- Transfer related technologies to the private sector
- Address the ethical, legal, and social issues (ELSIs) that may arise from the project

In 1991, working with Nobel laureate Hamilton Smith, Venter's genomic research project (TIGR) created the *shotgunning method*. At first the method was controversial among Venter's colleagues, who called it crude and inaccurate. However, Venter cross-checked his results by sequencing the genes in both directions, achieving a level of accuracy that greatly impressed his initial sceptical rivals. Within a year, TIGR published the entire genome of *Haemophilus influenzae*, a bacterium with nearly 2 million nucleotides.

The draft human genome sequence was published on February 15, 2001, in the journals *Nature* (publically funded Human Genome Project) and *Science* (Craig Venter's firm Celera).

1.2 GENETIC CODE AND MATHEMATICS

It is known that the secrets of life are more complex than DNA and the genetic code. One secret of life is the self-assembly of the first cell with a genetic

blueprint that allowed it to grow and divide. Another secret of life may be the mathematical control of life as we know it and the logical organization of the genetic code and the use of math in understanding life.

Mathematics has a fundamental role in understanding the complexities of living organisms. For example, the genetic code triplets of three bases in messenger ribonucleic acid (mRNA) that encode for specific amino acids during the translation process (synthesis of proteins using the genetic code in mRNA as the template) have some interesting mathematical logic in their organization (Cullman and Labouygues, 1984). An examination of this logical organization may allow us to better understand the logical assembly of the genetic code and life.

The genetic code in mRNA is composed of U for uracil, C for cytosine, A for adenine, and G for guanine. The genetic code triplets of three bases in messenger ribonucleic acid (mRNA) that encode for specific amino acids during the translation process (synthesis of proteins using the genetic code in mRNA as the template) have some interesting and mathematical logic in their organization.

In the first stage there was an investigation of the *standard genetic code*. In the past few decades, some other variants of the genetic code were revealed, which are described at the Web site http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi and which differ from the standard genetic code in some correspondences among 64 triplets, 20 amino acids, and stop codons. One noticeable feature of the genetic code is that some amino acids are encoded by several different but related base codons or triplets. There are 64 triplets or codons. In the case of the standard genetic code, three triplets (UAA, UAG, and UGA) are nonsense codons—no amino acid corresponds to their code. The remaining 61 codons represent 20 different amino acids. The genetic code is encoded in combinations of the four nucleotides found in DNA and then RNA. There are 16 possible combinations (4^2) of the four nucleotides of nucleotide pairs. This would not be sufficient to code for 20 amino acids (Prescott et al., 1993). The solution is mathematically simple. During the self-assembly and evolution of life, a code word (codon or triplet) evolved that provides for 64 (4^3) possible combinations. This simple code determines all the proteins necessary for life.

The genetic code is also degenerate. For example, up to six different codons are available for some amino acid. Another noteworthy aspect of biological messages is that minimal information is necessary to encode the messages (Peusner, 1974), and the messages can be encoded and decoded and put to work in amazingly short periods of time. A bacterial *E. coli* cell can grow and divide in half an hour, depending on the growth conditions. Mathematically, it could not be simpler.

Selenocysteine (twenty-first amino acid encoded by the genetic code) codon is UGA, normally a stop codon. Selenocysteine is a derivative of cysteine in which the sulfur atom is replaced by a selenium atom that is an essential atom in a small number of proteins, notably glutathione peroxidase. These proteins

are found in prokaryotes and eukaryotes, ranging from *E. coli* to humans. The selenocysteine is incorporated into proteins during translation in response to the UGA codon. This amino acid is readily oxidized by oxygen. Enzymes containing this amino acid must be protected from oxygen. As the oxygen concentration increased, the selenocysteine may gradually have been replaced by cysteine with the codons UGU and UGC (Madigan et al., 1997). The three-base code sometimes differs only in the third base position. For example, the genetic code for glycine is GGU, GGC, GGA, or GGG. Only the third base is variable. A similar third-base-change pattern exists for the amino acids lysine, asparagine, proline, leucine, and phenylalanine. These relationships are not random. For example, UUU codes for the same amino acid (phenylalanine) as UUC. In some codons the third base determines the amino acid. The second base is also important. For example, when the second base is C, the amino acid specified comes from a family of four codons for one amino acid, except for valine. Biological expression is in the form of coded messages—messages that contain the information on shapes of bimolecular structure and biochemical reactions necessary for life function. The coded message determines the protein, which folds into a shape that requires the minimal amount of energy. Therefore, the total energy of attraction and repulsion between atoms is minimal. How did this genetic code come to be the code of life as we know it? Nature had billions of years to experiment with different coding schemes, and eventually adopted the genetic code we have today.

It is simple in terms of mathematics. It is also conserved but can be mutated at the DNA level and also repaired. The code is thermodynamically possible and consistent with the origin, evolution, and diversity of life. Math as applied to understanding biology has countless uses. It is used to elucidate trends, patterns, connections, and relationships in a quantitative manner that can lead to important discoveries in biology. How can math be used to understand living organisms? One way to explore this relationship is to use examples from the bacterial world. The reader is also referred to an excellent text by Stewart (1998) that illustrates how math can be used to elucidate a fuller understanding of the natural world. For example, the exponential growth of bacterial cells (1 cell $\rightarrow$ 2 cells $\rightarrow$ 4 cells $\rightarrow$ 8 cells $\rightarrow$ 16 cells, and so on) is essential information that is one of the foundations of microbiology research. Exponential growth over known periods of time is essential in the understanding of bacterial growth in countless areas of research. The ability to use math to describe growth per unit of time is an excellent example of the interrelationship between math and the capability to understand this aspect of life. For example, the basic unit of life is the cell, an entity of 1. Bacteria also multiply by dividing. Remember that life is composed of matter, and matter is composed of atoms, and that atoms, especially in solids, are arranged in an efficient manner into molecules that minimize the energy needed to take on specific configurations. Often, these arrangements or configurations are repeating units of monomers that make up polymers. Stewart (1998) described it very well in his excellent book when he posed the question: "What could be more mathematical than

DNA?" The ability of DNA to replicate itself exactly and at the same time change ever so slightly allows evolutionary changes to occur. The mathematical sequences of four different bases (adenine, thymine, guanine, and cytosine) in DNA are the blueprint of life. Again, the order of the four bases determines the mRNA sequence, and then the protein that is synthesized. DNA in a cell is also capable of replicating itself precisely in a cell. The replicated DNA can then partition into each new cell when one cell divides and becomes two cells. The DNA can only replicate with the assistance of enzymes that unwind the DNA and allow the DNA strands to act as templates for synthesis of the second strand. The ability of a cell to unwind its DNA, replicate or copy new strands, and then partition them between two new cells has a mathematical basis. The four bases are paired in a specific manner: A (adenine) with T (thymine), C (cytosine) with G (guanine) on the opposite strands along a sugar phosphate backbone. Each strand can contain all four bases in any order. However, A must bond with T and C with G on opposite strands. This precise mathematical pairing must be obeyed.

Living organisms also have amazing mathematical order and symmetry. The repeating units of fatty acids, glycerol, and phosphate that make up a phospholipid membrane bilayer are one example. An excellent example of mathematical symmetry is the S-layer in many Archaea bacterial (prokaryotes consisting of methanogens, most extreme halophiles and hyperthermophiles, and *Thermoplasma*) cell walls that exhibit a hexagonal configuration. A cell that can assemble the same repeating units countless times is efficient and reduces the numbers of errors incorporated into the assembly. This is exactly the characteristic that is needed for a living cell to grow and divide. Yet a little bit of change can occur over time.

Biochemical reactions in cells are accompanied by gains or losses in energy during the reactions. Some of the energy is lost as heat and is not available to do work. In humans, heat is used to maintain a normal body temperature. The energy available to the cell is expressed as free energy and can be expressed as kJ/mol. Without the use of math and units of measurement, it would be impossible to describe energy metabolism in cells. Nor would we be able to describe the rates of enzyme reactions necessary for the self-assembly and functioning of life. Without units of temperature, we would not be able to describe the lower, upper, and optimum growth temperatures of specific microorganisms. The pH ranges for bacterial growth and the optimum pH values for enzyme reactions would be unknown without math to describe the values. Water availability values and oxygen concentrations would not be able to be described for growth of specific organisms. The examples are numerous. Without the use of math and scientific units to express values, our understanding of life would be minimal, and biology would not have made the great advances that it has made in the past decades. One central characteristic of living organisms is reproduction. From nutrients in their environment, they can self-assemble new cells in virtually exact copies. Second, living organisms are interdependent on each other and their activities. The Earth's biosphere,

with its abundance of oxygen and living organisms, was self-assembled by living organisms.

From a chaotic lifeless environment on the early Earth, life self-assembled with the cell as the basic unit, with mathematically precise order, symmetry, and base pairing in DNA as the genetic blueprint and with triplet codons as the genetic code for protein synthesis.

It is well known that all knowledge is intrinsically unified and lies in a small number of natural laws. Math can be used to understand life from the molecular level to the level of the biosphere. For example, this includes the origin and evolution of organisms, the nature of the genomic blueprints, and the universal genetic code as well as ecological relationships. Math helps us look for trends, patterns, and relationships that may or may not be obvious to scientists. Math allows us to describe the dimensions of genes and the sizes of organelles, cells, organs, and whole organisms. Without this knowledge, a paucity of information would still exist on many aspects of life.

1.3 MATHEMATICAL BACKGROUND

In this section we provide a general background of major branches of mathematics that we discuss in relation to bioinformatics throughout the book.

Algebra *Algebra* is the study of structure, relation, and quantity through symbolic operations for the systematic solution of equations and inequalities. In addition to working directly with numbers, algebra works with symbols, variables, and set elements. Addition and multiplication are viewed as general operations, and their precise definitions lead to advance structures such as groups, rings, and fields in which algebraic structures are defined and investigated axiomatically. Linear algebra studies the specific properties of vector spaces, including matrices. The properties common to all algebraic structures are studied in universal algebra. Axiomatic algebraic systems such as groups, rings, fields, and algebras over a field are investigated in the presence of a geometric structure (a metric or a topology) which is compatible with the algebraic structure. In recent years, algebraic structures have been discovered within the genetic codes, biological sequences, and biological structures. Matrices, polynomials, and other algebraic elements have been applied to studies of sequence alignments and protein structures and classifications.

Abstract Algebra Abstract algebra extends the familiar concepts from basic algebra to more general concepts. *Abstract algebra* deals with the more general concept of *sets*: a collection of all objects selected by property, specific for the set under binary operations. Binary operations are the keystone of algebraic structures studied in abstract algebra: They form a part of groups, rings, fields, and more. A *binary operation* is a rule for combining two objects of a given type to obtain another object of that type. More precisely, a binary operation

on a set S is a binary relation that maps elements of the Cartesian product $S \times S$ to S:

$$f: S \times S \rightarrow S$$

Addition (+), subtraction (–), multiplication (×), and division (÷) can be binary operations when defined on different sets, as is addition and multiplication of matrices, vectors, and polynomials. Groups, rings, and fields are fundamental structures in abstract algebra.

A *group* is a combination of a set S and a single binary operation "$*$" with the following properties:

- An *identity* element e exists such that for every member a of S, $e * a$ and $a * e$ are both identical to a.
- Every element has an *inverse*: For every member a of S, there exists a member a^{-1} such that $a * a^{-1}$ and $a^{-1} * a$ are both identical to the identity element.
- The operation is *associative*: If a, b, and c are members of S, then $(a * b) * c$ is identical to $a * (b * c)$.
- The set S is *closed* under the binary operation $*$.

For example, the set of integers under the operation of addition is a group. In this group, the identity element is 0 and the inverse of any element a is its negation, $-a$. The associativity requirement is met because for any integers a, b, and c, $(a + b) + c = a + (b + c)$. The integers under the multiplication operation, however, do not form a group. This is because, in general, the multiplicative inverse of an integer is not an integer. For example, 4 is an integer, but its multiplicative inverse is 1/4, which is not an integer.

The structures and classifications of groups are studied in group theory. A major result in this theory is the classification of finite simple groups, which is thought to classify all of the finite simple groups into roughly 30 basic types.

Semigroups, monoids, and quasigroups are structures similar to groups, but more general. They comprise a set and a closed binary operation, but do not necessarily satisfy the other conditions. A *semigroup* has an *associative* binary operation but might not have an identity element. A *monoid* is a semigroup that does have an identity but might not have an inverse for every element. A *quasigroup* satisfies a requirement that any element can be turned into any other by a unique pre- or postoperation; however, the binary operation might not be associative. All are instances of *groupoids*, structures with a binary operation upon which no further conditions are imposed. All groups are monoids, and all monoids are semigroups.

Groups have only one binary operation. Rings and fields explain the behavior of the various types of numbers; they are structures with two operators. A *ring* has two binary operations, + and ×, with × distributive over +.

Distributive property generalized the *distributive law* for numbers and specifies the order in which the operators should be applied. For the integers $(a + b) \times c = a \times c + b \times c$ and $c \times (a + b) = c \times a + c \times b$, and $\times$ is said to be *distributive* over +. Under the first operator (+), it is commutative (i.e., $a + b = b + a$). Under the second operator ($\times$) it is associative, but it does not need to have the identity or inverse property, so division is not allowed. The additive (+) identity element is written as 0 and the additive inverse of a is written as $-a$. Integers with both binary operations + and $\times$ are an example of a ring.

A *field* is a ring with the additional property that all the elements, excluding 0, form an *Abelian group* (have a commutative property) under $\times$. The multiplicative ($\times$) identity is written as 1, and the multiplicative inverse of a is written as a^{-1}. The rational numbers, the real numbers, and the complex numbers are all examples of fields.

These algebraic structures have been used in the study of genetic codes. Group theory has many applications in physics and chemistry, and it is potentially applicable in any situation characterized by symmetry. In chemistry, groups are used to classify crystal structures, regular polyhedrals, and the symmetries of molecules. The assigned point groups can then be used to determine physical properties (such as polarity and chirality) and spectroscopic properties (particularly useful for Raman spectroscopy and infrared spectroscopy), and to construct molecular orbitals.

Probability *Probability* is the language of uncertainty. It is the likelihood or chance that something is the case or will happen. Probability theory is used extensively in areas such as statistics, mathematics, science, philosophy, psychology, and in the financial markets to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems. The probability of an event E is represented by a real number in the range 0 to 1 and is denoted by $P(E)$, $p(E)$, or $\Pr(E)$. An impossible event has a probability of 0, and a certain event has a probability of 1.

Statistics *Statistics* is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. Statistical methods can be used to summarize or describe a collection of data; this is called *descriptive statistics*. Descriptive statistics can be used to summarize the data, either numerically or graphically, to describe the sample. Basic examples of numerical descriptors include the mean and standard deviation. Graphical summarizations include various types of charts and graphs. In addition, patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, and then used to draw inferences about the process or population being studied; this is called *inferential statistics*. Inferential statistics is used to model patterns in the data, accounting for randomness and drawing inferences about the larger population. These inferences may take the form of answers to yes/no questions (hypothesis testing), estimates

of numerical characteristics (estimation), descriptions of association (correlation), or modeling of relationships (regression). Other modeling techniques include ANOVA, time series, and data mining. Both descriptive and inferential statistics comprise applied statistics.

Probability and statistics have been used successfully to investigate sequence analysis, alignments, profile searches and phylogenetic trees, and many problems in bioinformatics.

Differential Geometry *Differential geometry* is a mathematical discipline that uses the methods of differential and integral calculus to study problems in geometry. The theory of plane and space curves and of surfaces in three-dimensional Euclidean space formed the basis for its initial development. Differential geometry has grown into a field concerned more generally with geometric structures on differentiable manifolds. It is closely related to differential topology and to the geometric aspects of the theory of differential equations. In physics, differential geometry is the language in which Einstein's general theory of relativity is expressed. According to the theory, the universe is a smooth manifold equipped with a pseudo-Riemannian metric, which describes the curvature of space–time. Understanding this curvature is essential for the positioning of satellites into orbit around the Earth. In the biological and medical sciences, differential geometry has been used to study protein confirmation and the elasticity of nonrigid objects such as human hearts and human faces.

Topology *Topology* is the mathematical study of the properties that are preserved through deformations, twistings, and stretchings of objects; however, tearing is not allowed. A circle is topologically equivalent to an ellipse (into which it can be deformed by stretching), and a sphere is equivalent to an ellipsoid. Similarly, the set of all possible positions of the hour hand of a clock is topologically equivalent to a circle (i.e., a one-dimensional closed curve with no intersections that can be embedded in two-dimensional space), the set of all possible positions of the hour and minute hands taken together is topologically equivalent to the surface of a torus (i.e., a two-dimensional surface that can be embedded in three-dimensional space), and the set of all possible positions of the hour, minute, and second hands taken together are topologically equivalent to a three-dimensional object. Topology can be used to abstract the inherent connectivity of objects while ignoring their detailed form. The mathematical definition of topology is described here briefly.

Let $\mathbf{X}$ be any set and let $\boldsymbol{T}$ be a family of subsets of $\mathbf{X}$. Then $\boldsymbol{T}$ is a topology on $\mathbf{X}$ if:

- Both the empty set and $\mathbf{X}$ are elements of $\boldsymbol{T}$.
- Any union of arbitrarily many elements of $\boldsymbol{T}$ is an element of $\boldsymbol{T}$.
- Any intersection of finitely many elements of $\boldsymbol{T}$ is an element of $\boldsymbol{T}$.

If $\boldsymbol{T}$ is a topology on $\mathbf{X}$, then $\mathbf{X}$ together with $\boldsymbol{T}$ is called a *topological space*.

All sets in $\boldsymbol{T}$ are called *open*; note that, in general, not all subsets of **X** need be in $\boldsymbol{T}$. A subset of **X** is said to be *closed* if its complement is in $\boldsymbol{T}$ (i.e., it is open). A subset of **X** may be open, closed, both, or neither.

A function or map from one topological space to another is called *continuous* if the inverse image of any open set is open. If the function maps the real numbers to the real numbers (both spaces with the standard topology), this definition of continuous is equivalent to the definition of continuous in calculus. If a continuous function is one-to-one and onto and if the inverse of the function is also continuous, the function is called a *homeomorphism*, and the domain of the function is said to be homeomorphic to the range. Another way of saying this is that the function has a natural extension to the topology. If two spaces are homeomorphic, they have identical topological properties and are considered to be topologically the same. The cube and the sphere are homeomorphic, as are the coffee cup and the doughnut. But the circle is not homeomorphic to the doughnut. DNA topology and protein topology are active research areas.

Knot Theory *Knot theory* is the mathematical branch of topology that studies mathematical *knots*, which are defined as embeddings of a circle in three-dimensional Euclidean space, R^3. This is basically equivalent to a conventional knotted string with the ends joined together to prevent it from becoming undone. Two mathematical knots are equivalent if one can be transformed into the other via a deformation of R^3 upon itself (known as an *ambient isotopy*); these transformations correspond to manipulations of a knotted string that do not involve cutting the string or passing the string through itself.

Knots can be described in various ways. Given a method of description, however, there may be more than one description that represents the same knot. For example, a common method of describing a knot is a planar diagram. But any given knot can be drawn in many different ways using a planar diagram. Therefore, a fundamental problem in knot theory is determining when two descriptions represent the same knot. One way of distinguishing knots is by using a *knot invariant*, a "quantity" that remains the same even with different descriptions of a knot. The concept of a knot has been extended to higher dimensions by considering n-dimensional spheres in m-dimensional Euclidean space.

The discovery of the Jones polynomial by Vaughan Jones in 1984 revealed deep connections between knot theory and mathematical methods in statistical mechanics and quantum field theory. In the last 30 years, knot theory has also become a tool in applied mathematics. Chemists and biologists use knot theory to understand, for example, the chirality of molecules and the actions of enzymes on DNA. In the last several decades of the twentieth century, scientists and mathematicians began finding applications of knot theory to problems in biology and chemistry. Knot theory can be used to determine whether or not a molecule is *chiral* (has "handedness"). Chemical compounds

of different handedness can have drastically differing properties, thalidomide being a notable example. More generally, knot theoretic methods have been used in studying *topoisomers*, topologically different arrangements of the same chemical formula. The closely related theory of *tangles* has been used effectively in studying the action of certain enzymes on DNA.

Graph Theory *Graph theory* is the study of *graphs*, mathematical structures used to model pairwise relations between objects from a certain collection. In this context a graph is a collection of vertices or *nodes* and a collection of *edges* that connect pairs of vertices. A graph may be *undirected*, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be *directed* from one vertex to another. A graph structure can be extended by assigning a weight to each edge of the graph. Graphs with weights, *weighted graphs*, are used to represent structures in which pairwise connections have some numerical values. For example, if a graph represents a road network, the weights could represent the length of each road. A digraph with weighted edges in the context of graph theory is called a *network*.

Many applications of graph theory exist in the form of network analysis. These split broadly into three categories:

1. Analysis to determine structural properties of a network, such as the distribution of vertex degrees and the diameter of the graph. A vast number of graph measures exist, and the production of useful ones for various domains remains an active area of research.
2. Analysis to find a measurable quantity within the network: for example, for a transportation network, the level of vehicular flow within any portion of it.
3. Analysis of dynamical properties of networks.

Graph theory is also used to study molecules in chemistry and biology. In chemistry a graph makes a natural model for a molecule, where vertices represent atoms and edge bonds. This approach is used especially in computer processing of molecular structures, ranging from chemical editors to database searching.

Fractals A *fractal* is generally "a rough or fragmented geometric shape that can be split into parts, each of which is (at least approximately) a reduced-size copy of the whole," a property called *self-similarity*. Because they appear similar at all levels of magnification, fractals are often considered to be infinitely complex (in informal terms). Natural objects that approximate fractals to a degree include clouds, mountain ranges, lightning bolts, coastlines, and snowflakes.

Fractals can also be classified according to their self-similarity. Three types of self-similarity are found in fractals:

1. *Exact self-similarity.* This is the strongest type of self-similarity; the fractal appears identical at different scales. Fractals defined by iterated function systems often display exact self-similarity.
2. *Quasi-self-similarity.* This is a loose form of self-similarity; the fractal appears approximately (but not exactly) identical at different scales. Quasi-self-similar fractals contain small copies of the entire fractal in distorted and degenerate forms. Fractals defined by recurrence relations are usually quasi-self-similar but not exactly self-similar.
3. *Statistical self-similarity.* This is the weakest type of self-similarity; the fractal has numerical or statistical measures that are preserved across scales. Most reasonable definitions of *fractal* trivially imply some form of statistical self-similarity. (A fractal dimension itself is a numerical measure that is preserved across scales.) Random fractals are examples of fractals that are statistically self-similar, but neither exactly self-similar nor quasi-self-similar.

Approximate fractals are easily found in nature. These objects display a self-similar structure over an extended but finite scale range. Examples include clouds, snowflakes, crystals, mountain ranges, lightning, river networks, cauliflower and broccoli, and systems of blood vessels and pulmonary vessels. Coastlines may be loosely considered fractal in nature.

Complexities *Complexity theory* and *chaos theory* study systems that are too complex to predict their future accurately, but nevertheless, exhibit underlying patterns that can help us cope in an increasingly complex world. Science usually examines the world by breaking it into smaller and smaller pieces until the pieces can be understood. When we use this approach, we often miss the bigger picture. Knowing all we can about an individual ant will not teach us about how an entire ant colony works. Dissecting a rat will never tell us all that we need to know about living rats. Sometimes the way that the parts interact is critical to how the entire system works. This is what complexity studies. Complexity is relevant to an enormous range of areas of study, including traffic flows, earthquakes, the stock market, and systems biology.

Rademacher and Walsh Functions Digital communication uses nonsinusoidal orthogonal functions, Rademacher and Walsh functions being among the best known. They are described extentively in the literature (Ahmed and Rao, 1975; Geadah and Corinthios, 1977; Goldberg, 1989a,b; Peterson and Weldon, 1972; Sklar, 2001; Trahtman and Trahtman, 1975; Vose and Wright, 1998; Waterman, 1999; Yarlagadda and Hershey, 1997; Zalmanzon, 1989).

Rademacher functions are an incomplete set of orthogonal functions introduced by Rademacher in 1922. A Rademacher function of index m, denoted by $\text{rad}(m, t)$, is a train of rectangular pulses with 2^{m-1} cycles in the half-open interval [0, 1), taking the values +1 or −1 (Figure 1.1). The exception is

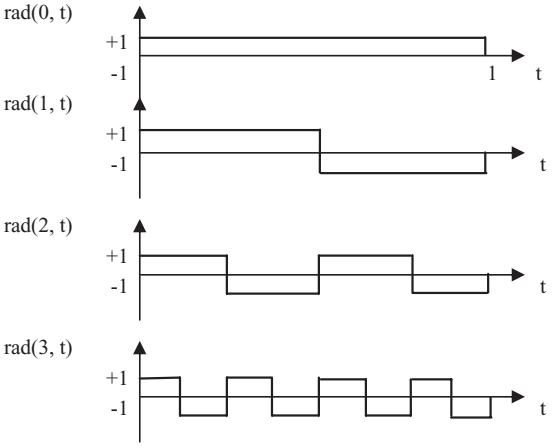

FIGURE 1.1 Rademacher functions.

rad(0, t), which is equal to +1 along the entire interval. Rademacher functions can be generated using the recurrence relation:

$$\text{rad}(m, t) = \text{rad}(1, 2^{m-1}t)$$

$$\text{rad}(1, t) = +1 \quad \text{if } t \text{ from } \left[0, \frac{1}{2}\right) \quad \text{and} \quad \text{rad}(1, t) = -1 \quad \text{if } t \text{ from } \left[\frac{1}{2}, 1\right)$$

The incomplete set of Rademacher functions was completed by Walsh in 1923 to form a complete orthogonal set of rectangular functions now known as *Walsh functions*. In the field of digital communication, sets of Walsh functions are generally classified into three groups, which differ from one another by the order in which individual functions appear:

1. Walsh ordering
2. Dyadic or Paley ordering
3. Natural or Hadamard ordering

All these variants of the sets of Walsh functions can be presented in connection with relevant Hadamard matrices (see Chapter 8). Peculiarities of these variants are related closely to the famous Gray code (Ahmed and Rao, 1975, pp. 88–93).

The complete set of Walsh functions defined on the unit interval [0, 1) can be divided into two groups of even and odd functions about the point $t = 0.5$. These even and odd functions are analogous to the sine and cosine functions, respectively. The class of nonsinusoidal orthogonal functions described plays an important role in the spectral analysis of signals and in relevant transforms of digital signals to provide effective transfer of information.

1.4 CONVERTING DATA TO KNOWLEDGE

The biological information we gain allows us to learn about ourselves, about our origins, and about our place in the world. We have learned that we are quantitatively strongly related to other primates, mice, zebrafish, fruit flies, roundworms, and even yeast. The findings should induce in us some modesty: in learning and seeing how much we share with all living organisms. The information we are gaining is not just of philosophical interest but is also intended to help humanity to lead healthy lives. Knowledge about primitive organisms provides much information about shared metabolic features and hints about diseases that affect humans in an economically and ethically acceptable manner.

Knowledge from many scientific disciplines and their subfields has to be integrated to achieve the goals of bioinformatics. It was believed (Wilson, 1998) that all knowledge is intrinsically unified, and that behind disciplines as diverse as physics and biology, and anthropology and the arts, lie a small number of natural laws. Applying the knowledge can lead to new scientific methods, new diagnostics, and new therapeutics.

At the beginning of the "genomic revolution," a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide, amino acid, and protein sequences. Development of this type of database involved not only design issues but also the development of complex interfaces whereby researchers could both access existing data and submit new or revised data. Ultimately, all of this information must be combined to form a comprehensive picture of normal cellular activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide, amino acid sequences, protein domains, and protein structures and interactions. Important research branches within bioinformatics include the development and implementation of tools that enable efficient access to, and use and management of, various types of information and new algorithms and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences. The process of converting data to knowledge may be illustrated as shown in Figure 1.2.

1.5 THE BIG PICTURE: INFORMATICS

Informatics is the study of the structure, behaviors, and interactions of natural and artificial computational systems. Informatics studies the representation, processing, and communication of information in natural and artificial systems. It has computational, cognitive, and social aspects. The central notion is the transformation of information: whether by computation or communication,

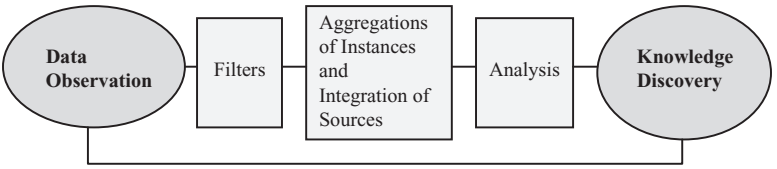

FIGURE 1.2 Process of converting data to knowledge.

TABLE 1.1 Information Building Blocks (Monomer to Polymer)

Monomer	Polymer
Nucleotides	DNA:
Adenine (A)	ACTGGTAGCCTTAGA …
Cytosine (C)	RNA:
Guanine (G)	ACUGGUAGCCUUAGA …
Thymine/ uracil (T/U)	
Amino acids	Protein:
Cysteine (Cys)	Met–Cys–Gly–Pro–Pro–Arg …
Alanine (Ala)	
Proline (Pro)	
Letters: A, B, C, …	Words: CAT, GO, FRIEND, …
Symbols: 0, 1	Binary code: 1001011100101 …
Monomial: $1, x, x^2, \ldots$	Polynomial: $P(x)$, …
Line: $l_1, l_2, l_3, \ldots$	Polygons: triangle, rectangle, …

whether by organisms or artifacts. Information building blocks are illustrated conceptually in Table 1.1.

Understanding informational phenomena such as computation, cognition, and communication enables technological advances. In turn, technological progress prompts scientific enquiry. The science of information and the engineering of information systems develop hand-in-hand. Informatics is the emerging discipline that combines the two. In natural and artificial systems, information is carried at many levels, ranging, for example, from biological molecules and electronic devices, through nervous systems and computers, and on to societies and large-scale distributed systems. It is characteristic that information carried at higher levels is represented by informational processes at lower levels. Each of these levels is the proper object of study for some discipline of science or engineering. Informatics aims to develop and apply firm theoretical and mathematical foundations for the features that are common to all computational systems.

In its attempts to account for phenomena, science progresses by defining, developing, criticizing, and refining new concepts. Informatics is developing its own fundamental concepts of communication, knowledge, data, interaction,

and information, and relating them to such phenomena as computation, thought, and language.

Informatics has many aspects and encompasses a number of existing academic disciplines: artificial intelligence, cognitive science, and computer science. Each takes part of informatics as its natural domain: In broad terms, cognitive science concerns the study of natural systems; computer science concerns the analysis of computation and the design of computing systems; and artificial intelligence plays a connecting role, designing systems that emulate those found in nature. Informatics also informs and is informed by other disciplines, such as mathematics, electronics, biology, linguistics, and psychology. Thus, informatics provides a link between disciplines with their own methodologies and perspectives, bringing together a common scientific paradigm, common engineering methods, and a pervasive stimulus from technological development and practical application.

Computational Systems Computational systems, whether natural or artificial, are distinguished by their great complexity with regard to both their internal structure and behavior, and their rich interaction with the environment. Informatics seeks to understand and to construct (or reconstruct) such systems using analytic, experimental, and engineering methodologies. The mixture of observation, theory, and practice will vary between natural and artificial systems.

In natural systems, the object is to understand the structure and behavior of a given computational system. Ultimately, the theoretical concepts underlying natural systems are built on observation and are themselves used to predict new observations. For artificial systems, the object is to build a system that performs a given informational function. The theoretical concepts underlying artificial systems are intended to secure their correct and efficient design and operation. Computer language systems have been evolving and communicating with biological data as part of computational systems. The computer languages and their interfaces with various data types are illustrated in Table 1.2.

TABLE 1.2 Communications Between Computer Languages and Data Types and BioModules[a]

Computer Languages	Design Goals
FORTRAN	Numerical analysis
LISP	Symbolic computation
C	System programming
C++	Objects, speed, compatibility with C
Java	Objects, Internet
Perl	System administration
Python	General programming

[a]BioModules = bio + languages.

Informatics provides an enormous range of problems and opportunities. One challenge is to determine how far, and in what circumstances, theories of information processing in artificial devices can be applied to natural systems. A second challenge is to determine how far principles derived from natural systems are applicable to the development of new types of artificial systems. A third challenge is to explore the many ways in which artificial information systems can help to solve problems facing humankind and help to improve the quality of life for all living things. One can also consider systems of mixed character; a question of longer-term interest may be to what extent it is helpful to maintain the distinction between natural and artificial systems. In Chapter 10 we present the evolution, future trends, and the central dogma of informatics.

1.6 CHALLENGES AND PERSPECTIVES

The interaction between biology and mathematics has been a rich area of research for more than a century. The interface between them presents challenges and opportunities for both mathematicians and biologists. Due to the explosion of biological data with the advent of new technologies that can organize the plethora of data, unique opportunities for research and new challenges have surfaced within the last 10 to 20 years. For biology, the possibilities range from the level of the cell and molecule to the level of the biosphere. For mathematics, the potential is great in traditional and nontraditional areas such as statistics and differential equations, knot theory, and topology. Stochastic processes and Markov chains in statistics have their origins in biological questions. Galton invented the correlation method based on questions in evolutionary biology. The analysis of variance was derived from R. A. Fisher's work in agriculture. Modeling the success (survival) over many generations of a family name led to the development of the subject of branching processes. The compilation of DNA sequence data led to Kingman's coalescence model and Ewens' sampling formula. Furthermore, biological applications have stimulated the study of ordinary and partial differential equations, especially regarding problems in chaos, fractal geometry, and bifurcation theory. Further interactions between mathematics and biology have presented new opportunities and challenges. A number of fundamental mathematical and biological issues cut across all these challenges.

- How do we incorporate variation among individual units in nonlinear systems and biological systems?
- How do we explain the interactions among phenomena that occur on a wide range of scales and molecular levels, of space, time, and organizational complexity?
- What is the relation between pattern and process both in mathematical and biological systems?

It is in the analysis of these issues that mathematics is most essential and holds the greatest potential. These challenges, such as aggregation of components to elucidate the behavior of ensembles, integration across scales, and inverse problems, are basic to all sciences, in particular to biological sciences, and a variety of techniques exist to deal with them and to begin to solve the biological problems that generate them. However, the uniqueness of biological systems shaped by evolutionary forces will pose new difficulties, mandate new perspectives, and lead to the development of new mathematics. Algebraic biology and matrix genetics for genetic language are presented in Chapters 2 and 8, and a denotational mathematics for cognitive informatics is introduced in Chapter 9. The excitement of this area of science is already evident, and is sure to grow in the years to come.

REFERENCES

Ahmed, N., and Rao, K. (1975). *Orthogonal Transforms for Digital Signal Processing*. New York: Springer-Verlag.

Cullman, G., and Labouygues, J. M. (1984). The mathematical logic of life. In: K. Dose, A. W. Schwartz, and W. H.-P. Thiemann (Eds.), *Proceedings of the 7th International Conference on the Origins of Life*. Dordrecht, The Netherlands: D. Reidel.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London: John Murray.

Geadah, Y. A., and Corinthios, M. J. (1977). Natural, dyadic and sequency order algorithms and processors for the Walsh–Hadamard transform. *IEEE Trans. Comput.*, **C-26**, 435–442.

Goldberg, D. E. (1989a). Genetic algorithms and Walsh functions: I. A gentle introduction. *Complex Syst.*, **2**(2), 129–152.

Goldberg, D. E. (1989b). Genetic algorithms and Walsh functions: II. Deception and its analysis. *Complex Syst.*, **3**(2), 153–171.

Madigan, M. T., Martinko, J. M., and Parker, J. (1997). *Brock Biology of Microorganisms*. Upper Saddle River, NJ: Prentice Hall.

Peterson, W. W., and Weldon, E. J. (1972). *Error-Correcting Codes*. Cambridge, MA: MIT Press.

Peusner, L. (1974). *Concepts in Bioenergetics*. Englewood Cliffs, NJ: Prentice-Hall.

Prescott, L. M., Harley, J. P., and Klein, D. A. (1993). *Microbiology*. Dubuque, IA: Wm. C. Brown.

Sklar, B. (2001). *Digital Communication: Fundamentals and Applications*. Upper Saddle River, NJ: Prentice Hall.

Stewart, I. (1998). *Life's Other Secret*. New York: Wiley.

Trahtman, A. M., and Trahtman, V. A. (1975). *The Foundations of the Theory of Discrete Signals on Finite Intervals*. Moscow: Sovetskoie Radio (in Russian).

Vose, M., and Wright, A. (1998). The simple genetic algorithm and the Walsh transform: I. Theory. *J. Evol. Comput.*, **6**(3), 253–274.

Waterman, M. S. (Ed.) (1999). *Mathematical Methods for DNA Sequences*. Boca Raton, FL: CRC Press.

Wilson, E. (1998). *Consilience: The Unity of Knowledge*. New York: Random House.

Yarlagadda, R., and Hershey, J. (1997). *Hadamard Matrix Analysis and Synthesis with Applications to Communications and Signal/Image Processing*. New York: Kluwer Academic.

Zalmanzon, L. A. (1989). *Fourier, Walsh and Haar Transformations and Their Application in Control, Communication and Other Systems*. Moscow: Nauka (in Russian).