Mathematical Challenges from Genomics and Molecular Biology

Richard M. Karp

A fundamental goal of biology is to understand how living cells function. This understanding is the foundation for all higher levels of explanation, including physiology, anatomy, behavior, ecology, and the study of populations. The field of molecular biology analyzes the functioning of cells and the processes of inheritance principally in terms of interactions among three crucially important classes of macromolecules: DNA, RNA, and proteins. Proteins are the molecules that enable and execute most of the processes within a cell. DNA is the carrier of hereditary information in the form of genes and directs the production of proteins. RNA is a key intermediary between DNA and proteins.

Molecular biology and genetics are undergoing revolutionary changes. These changes are guided by a view of a cell as a collection of interrelated subsystems, each involving the interaction among many genes and proteins. Emphasis has shifted from the study of individual genes and proteins to the exploration of the entire genome of an organism and the study of networks of genes and proteins. As the level of aspiration rises and the amount of available data grows by orders of magnitude, the field becomes increasingly dependent on mathematical modeling, mathematical analysis, and computation. In the sections that follow we give an introduction to the mathematical and computational challenges that arise in this field, with an emphasis on discrete algorithms and the role of combinatorics, optimization, probability, statistics, pattern recognition, and machine learning.

We begin by presenting the minimal information about genes, genomes, and proteins required to understand some of the key problems in genomics. Next we describe some of the fundamental goals of the molecular life sciences and the role of genomics in attaining these goals. We then give a series of brief vignettes illustrating algorithmic and mathematical questions arising in a number of specific areas: sequence comparison, sequence assembly, gene finding, phylogeny construction, genome rearrangement, associations between polymorphisms and disease, classification and clustering of gene expression data, and the logic of transcriptional control. An annotated bibliography provides pointers to more detailed information.

Richard M. Karp is a member of the International Computer Science Institute (ICSI), Berkeley, and a University Professor at the University of California, Berkeley. His e-mail address is karp@icsi.berkeley.edu.

Genes, Genomes, and Proteins

The Double Helix

The field of genetics began with Gregor Mendel (1865), who postulated the existence of discrete units of information (which later came to be called genes) that govern the inheritance of individual characteristics in an organism. In the first half of the twentieth century it was determined that the genes are physically embodied within complex DNA macromolecules that lie within structures called chromosomes which occur in every living cell. This set the stage for the epochal discovery of the structure of DNA by Watson and Crick in 1953. They showed that a DNA molecule is a double helix consisting of two strands. Each helix

is a chain of *bases*, chemical units of four types: A, C, T, and G. Each base on one strand is joined by a hydrogen bond to a complementary base on the other strand, where A is complementary to T, and C is complementary to G. Thus the two strands contain the same information. Certain segments within these chromosomal DNA molecules contain genes, which are the carriers of the genetic information and, in a sense to be explained later, spell the names of the proteins. Thus the genetic information is encoded digitally, as strings over the four-letter alphabet {A, C, T, G}, much as information is encoded digitally in computers as strings of zeros and ones.

In humans there are forty-six chromosomes. All but two of these (the sex chromosomes) occur in pairs of "homologous" chromosomes. Two homologous chromosomes contain the same genes, but a gene may have several alternate forms called alleles, and the alleles of a gene on the two chromosomes may be different.

The total content of the DNA molecules within the chromosomes is called the *genome* of an organism. Within an organism, each cell contains a complete copy of the genome. The human genome contains about three billion base pairs and about 35,000 genes.

Proteins

Proteins are the workhorses of cells. They act as structural elements, catalyze chemical reactions, regulate cellular activities, and are responsible for communication between cells. A protein is a linear chain of chemical units called amino acids, of which there are twenty common types. The function of a protein is determined by the three-dimensional structure into which it folds. One of the premier problems in science is the *protein folding problem* of predicting the three-dimensional structure of a protein from its linear sequence of amino acids. This problem is far from being solved, although progress has been made by a variety of methods. These range from numerical simulation of the physical forces exerted by the amino acids on one another to pattern recognition techniques which correlate motifs within the linear amino acid sequence with structural features of a protein.

From Genes to Proteins

The fundamental dogma of molecular biology is that DNA codes for RNA and RNA codes for protein. Thus the production of a protein is a two-stage process, with RNA playing a key role in both stages. An RNA molecule is a single-stranded chain of chemical bases of four types: A, U, C, and G. In the first stage, called transcription, a gene within the chromosomal DNA is copied base-by-base into RNA according to the correspondence A → U, C → G, T → A, G → C. The resulting RNA transcript of the gene is then transported within the cell to a molecular machine called a ribosome which has the function of translating the RNA into protein. Translation takes place according to the genetic code, which maps successive triplets of RNA bases to amino acids. With minor exceptions, this many-to-one function from the sixty-four triplets of bases to the twenty amino acids is the same in all organisms on Earth.

Regulation of Gene Expression

All the cells within a living organism (with the exception of the sperm and egg cells) contain nearly identical copies of the entire genome of the organism. Thus every cell has the information needed to produce any protein that the organism can produce. Nevertheless, cells differ radically in the proteins that they actually produce. For example, there are more than 200 different human cell types, and most proteins are produced in only a subset of these cell types. Moreover, any given cell produces different proteins at different stages within its cycle of operation, and its protein production is influenced by its internal environment and by the signals impinging upon it from other cells.

It is clear, then, that the expression of a gene within a cell (as measured by the abundance and level of activity of the proteins it produces) is regulated by the environment of the cell. Transcription of a gene is typically regulated by proteins called transcription factors that bind to the DNA near the gene and enhance or inhibit the copying of the gene into RNA. Similarly, translation can be regulated by proteins that bind to the ribosome. Certain post-translational processes, such as the chemical modification of the protein or the transport of protein to a particular compartment in the cell can also be regulated so as to affect the activity of the protein. Thus gene expression can be viewed as a complex network of interactions involving genes, proteins, and RNA, as well as other factors such as temperature and the presence or absence of nutrients and drugs within the cell.

The Goals of Genomics

In this section we enumerate some of the goals of genomics.

1. Sequence and compare the genomes of different species. To sequence a genome means to determine its sequence of bases. This sequence will, of course, vary from individual to individual, and those individual differences are of paramount importance in determining each individual's genetic makeup, but there is enough agreement to justify the creation of a composite reference genome for a species. For example, any two humans will have the same complement of genes (but different alleles) and will agree in about 999 bases out of 1,000.

   The sequencing of the human genome has been a central goal of the world genomics community since 1990. Draft sequences were completed in February 2001, and the quest continues for a much more accurate sequence.

This achievement was preceded by the sequencing of many bacterial genomes, yeast, the nematode, and, in June 1999, the fruit fly *Drosophila melanogaster*. The sequencing of a new organism is often of value for medical, agricultural, or environmental studies. In addition, it may be useful for comparative studies with related organisms.

2. Identify the genes and determine the functions of the proteins they encode. This process is essential, since without it a sequenced genome is merely a meaningless jumble of A's, C's, T's, and G's. Genes can be identified by methods confined to a single genome or by comparative methods that use information about one organism to understand another related one.
3. Understand gene expression. How do genes and proteins act in concert to control cellular processes? Why do different cell types express different genes and do so at different times?
4. Trace the evolutionary relationships among existing species and their evolutionary ancestors.
5. Solve the protein folding problem: From the linear sequence of amino acids in a protein, determine the three-dimensional structure into which it folds.
6. Discover associations between gene mutations and disease. Some diseases, such as cystic fibrosis and Huntington's disease, are caused by a single mutation. Others, such as heart disease, cancer, and diabetes, are influenced by both genetic and environmental factors, and the genetic component involves a combination of influences from many genes. By studying the relation between genetic endowment and disease states in a population of individuals, it may be possible to sort out the genetic influences on such complex diseases.

Having completed our brief survey of the general goals of genomics, we now turn to a number of examples of specific problems in genomics. These typically involve the creation of a mathematical model, the development of an algorithm, and a mathematical analysis of the algorithm's performance.

Sequence Comparison

The similarity of a newly discovered gene or protein to known genes or proteins is often an indication of its importance and a clue to its function. Thus, whenever a biologist sequences a gene or protein, the next step is to search the sequence databases for similar sequences. The BLAST (Basic Local Alignment Search Tool) program and its successive refinements serve this purpose and are the most important single software tool for biologists.

In preparation for giving a measure of the similarity between two sequences of residues (i.e., bases or amino acids), we need a definition: An *alignment* of a pair of sequences x and y is a new pair of sequences x' and y' of equal length such that x' is obtained from x and y' is obtained from y by inserting occurrences of the special space symbol $(-)$. Thus, if $x = acbcdb$ and $y = abbdcdc$, then one alignment of x and y is as follows:

```
x′ = a - c b c - d b,
y′ = a b - b d c d c.
```

Given an alignment of two sequences x and y, it is natural to assess its quality (i.e., the extent to which it displays the similarity between the two sequences) as a score, which is the sum of scores associated with the individual columns of the alignment. The score of a column is given by a symmetric scoring function σ that maps pairs of symbols from the alphabet $\Sigma \cup \{-\}$ to the real numbers, where Σ is the set of residues. Normally we will choose $\sigma(a, a) > 0$ for all symbols $a \in \Sigma$, so that matched symbols increase the score of the alignment, and $\sigma(a, -) < 0$ for all $a \in \Sigma$, so that misalignments are penalized. In the case of the alphabet of amino acids, $\sigma(a, b)$ reflects the frequency with which amino acid a replaces amino acid b in evolutionarily related sequences.

The global alignment problem is to find the optimal alignment of two strings x and y with respect to a given scoring function σ. A dynamic programming algorithm called the Needleman-Wunsch algorithm solves this problem in a number of steps proportional to the product of the lengths of the two sequences. A straightforward implementation of this algorithm requires space proportional to the product of the lengths of the two sequences, but there is a refinement which, at the cost of doubling the execution time, reduces the space requirement to $m + \log n$, where m and n are the lengths of the shorter and the longer of the two sequences.

A related problem is that of local alignment, in which we seek the highest score of an alignment between consecutive subsequences of x and y, where these subsequences may be chosen as desired. Such an alignment is intended to reveal the extent of local similarity between sequences that may not be globally similar. This problem can be solved within the same time and space bounds as the global alignment problem, using a dynamic programming algorithm due to Smith and Waterman.

A *gap* is a sequence of consecutive columns in an alignment in which each symbol of one of the sequences (x' or y') is the space symbol $(-)$. Gaps typically correspond to insertions or deletions of residue sequences over the course of evolution. Because mutations causing such insertions and deletions may be considered a single evolutionary event (and may be nearly as likely as the insertion or deletion of a single residue), we may wish to assign a (negative) score to a gap which is greater

than the sum of the (negative) scores of its columns. The above dynamic programming algorithms can be adapted for this purpose.

One of the most commonly occurring tasks in computational genomics is to search a database for sequences similar to a given sequence. BLAST is a set of programs designed for this purpose. Ideally, it would be desirable to scan through the entire database for high-scoring local alignments, but this would require a prohibitive amount of computation. Instead, a filtering program is used to find regions of the database likely to have a high-scoring local alignment with the given sequences; the full local alignment algorithm is then used within these regions. One principle underlying the filtering program is that two sequences are likely to have a high-scoring local alignment only if there is a reasonably long exact match between them.

Multiple Alignment

The concept of an alignment can be extended to alignments of several sequences. A *multiple alignment* of the sequences $x_1, x_2, \ldots, x_n$ is an n-tuple $(x'_1, x'_2, \ldots, x'_n)$ of sequences of equal length where, for $i = 1, 2, \ldots, n$, the sequence x'_i is obtained from x_i by inserting occurrences of the space symbol.

Just as in the case of a pairwise alignment, the scoring of a multiple alignment is based on a symmetric score function σ from $(\Sigma \cup \{-\})^2$ into the real numbers; usually we take $\sigma(-, -)$ to be zero. The score of a multiple alignment is then computed column by column. The *sum-of-pairs scoring method*, in which the score of a multiple alignment is the sum of $\sigma(a, b)$ over all aligned pairs of symbol occurrences, is a natural choice, with the convenient property that the score of an alignment is the sum of the scores of its induced pairwise alignments.

For the commonly used scoring methods, the problem of finding a maximum-score multiple alignment of a set of sequences is NP-hard, and various heuristics are used in practice.

Hidden Markov Models

Multiple alignments are an important tool for exhibiting the similarities among a set of sequences. Hidden Markov models (HMMs) provide a more flexible probabilistic method of exhibiting such similarities. An HMM is a Markov chain that stochastically emits an output symbol in each state. It is specified by a finite set of states, a finite set of output symbols, an initial state, transition probabilities $p(q, q')$, and emission probabilities $e(q, b)$. Here $p(q, q')$ is the probability that the next state is q' given that the present state is q, and $e(q, b)$ is the probability of emitting output symbol b in state q. In typical biological applications the output symbols are residues (nucleotides or amino acids), and an HMM is used to represent the statistical features of a family of sequences, such as the family of globin proteins. A subsequent section describes the construction of an HMM representing the statistical features of human genes.

An HMM for a family of sequences should have the property that sequences in the family tend to be generated with higher probability than other sequences of the same length. In view of this property, one can judge whether any given sequence lies in the family by computing the probability that the HMM generates it, a task that can be performed efficiently by a simple dynamic programming algorithm.

In order to construct a hidden Markov model of a family of sequences, one needs a *training set* consisting of representative sequences from the family. The first step in constructing the HMM is to choose the set of states and the initial state and to specify which transition probabilities and which emission probabilities can be nonzero. These choices are guided by the modeler's knowledge of the family. Given these choices, one can use the *EM-algorithm* to choose the numerical values of the nonzero transition probabilities and emission probabilities in order to maximize the product of the emission probabilities of the sequences in the training set.

Sequence Assembly

The genomes of different organisms vary greatly in size. There are about 3 billion base pairs in the human genome, 120 million in the genome of the fruit fly *Drosophila melanogaster*, and 4.7 million in the genome of the bacterium *E. coli*. There is no magic microscope than can simply scan across a genome and read off the bases. Instead, genomes are sequenced by extracting many fragments called *reads* from the genome, sequencing each of these reads, and then computationally assembling the genome from these reads. The typical length of a read is about 500 bases, and the total length of all the reads is typically five to eight times the length of the genome. The reads come from initially unknown locations distributed more or less randomly across the genome. The process of sequencing a read is subject to error, but the error rate is usually low.

Shotgun sequencing is conceptually the simplest way to assemble a genome from a set of reads. In this method the reads are compared in pairs to identify those pairs that appear to have a significant overlap. Then the reads are aligned in a manner consistent with as many of these overlaps as possible. Finally, the most likely genomic sequence is derived from the alignment.

During the 1990s The Institute for Genomic Research (TIGR) used the shotgun method to sequence the genomes of many microorganisms, of size up to about five megabases. However, the method was not believed to be applicable to organisms, such as *Homo sapiens*, having much larger genomes containing many *repeat families*. Repeat families are sequences that are repeated with very little

variation throughout a genome. For example, the ALU repeat family consists of nearly exact repetitions of a sequence of about 280 bases covering about 10 percent of the human genome. Repeat families in a genome complicate the sequence assembly process, since matching sequences within two reads may come from distinct occurrences of a repeat sequence and therefore need not indicate that the reads overlap.

The Human Genome Project, an international effort coordinated by the U.S. Department of Energy and the National Institutes of Health, favors a divide and conquer approach over the shotgun sequencing approach. The basic idea is to reduce the sequencing of the entire genome to the sequencing of many fragments called *clones* of length about 130,000 bases whose approximate locations on the genome have been determined by a process called *physical mapping*.

In 1998 the biologist Craig Venter established Celera Genomics as a rival to the Human Genome Project and set out to sequence the human genome using the shotgun sequencing approach. Venter was joined by the computer scientist Gene Myers, who had conducted mathematical analyses and simulation studies indicating that the shotgun sequencing approach would work provided that most of the reads were obtained in pairs extracted from the ends of short clones. The advantage of using paired reads is that the approximate distance between the two reads is known. This added information reduces the danger of falsely inferring overlaps between reads that are incident with different members of the same repeat family. Celera demonstrated the feasibility of its approach by completing the sequencing of *Drosophila melanogaster* (the fruit fly) in March 2000.

In February 2001 Celera and the Human Genome Project independently reported on their efforts to sequence the human genome. Each group had obtained a rough draft sequence covering upwards of 90 percent of the genome but containing numerous gaps and inaccuracies. Both groups are continuing to refine their sequences. The relative merits of their contrasting approaches remain a topic of debate, but there is no doubt about the significance of their achievement.

Gene Finding

Although the sequencing of the human genome is a landmark achievement, it is not an end in itself. A string of three billion A's, C's, T's, and G's is of little value until the meaning hidden within it has been extracted. This requires finding the genes, determining how their expression is regulated, and determining the functions of the proteins they encode. These are among the goals of the field of *functional genomics*. In this section we discuss the first of these tasks, gene finding.

Living organisms divide into two main classes: prokaryotes, such as bacteria and blue-green algae, in which the cell does not have a distinct nucleus, and eukaryotes, in which the cells contain visibly evident nuclei and organelles. Gene finding within prokaryotes is relatively easy because each gene consists of a single contiguous sequence of bases. In higher eukaryotes, however, a gene typically consists of two or more segments called *exons* that code for parts of a protein, separated by noncoding intervening segments called *introns*. In the process of transcription the entire sequence of exons and introns is transcribed into a *pre-mRNA transcript*. Then the introns are removed and the exons are spliced together to form the mRNA transcript that goes to a ribosome to be translated into protein. Thus the task of gene identification involves parsing the genomic region of a gene into exons and introns. Often this parsing is not unique, so that the same gene can code for several different proteins. This phenomenon is called *alternative splicing*.

The identification of a gene and its parsing into exons and introns is based on signals in the genomic sequence that help to identify the beginning of the first exon of a gene, the end of the last exon, and the exon-intron boundaries in between. Some of these signals derive from the nature of the genetic code. Define a *codon* as a triplet of DNA bases. Sixty-one of the sixty-four codons code for specific amino acids. One of these (ATG) is also a start codon determining the start of translation. The other three codons (TAA, TAG, and TGA) are stop codons which terminate translation. It follows that the concatenation of all the exons starts with ATG (with occasional exceptions) and ends with one of the three stop codons. In addition, each intron must start with GT and end with AG. There are also important statistical tendencies concerning the distribution of codons within exons and introns and the distribution of bases in certain positions near the exon-intron boundaries. These deterministic signals and statistical tendencies can be incorporated into a hidden Markov model for generating genomic sequence. Given a genomic sequence, one can use a dynamic programming algorithm called the Viterbi algorithm to calculate the most likely sequence of states that would occur during the emission of the given sequence by the model. Each symbol is then identified as belonging to an exon, intron, regulatory region, etc., according to the state that the HMM resided in when the symbol was emitted.

Another approach to gene finding is based on the principle that functioning genes tend to be preserved in evolution. Two genes in different species are said to be *orthologous* if they are derived from the same gene in a common ancestral species. In a pair of species that diverged from

one another late in evolution, such as man and mouse, one can expect to find many orthologous pairs of genes which exhibit a high level of sequence similarity; hence the fact that a sequence from the human genome has a highly similar counterpart in the mouse increases the likelihood that both sequences are genes. Thus one can enhance gene finding in both man and mouse by aligning the two genomes to exhibit possible orthologous pairs of genes.

Phylogeny Construction

The evolutionary history of a genetically related group of organisms can be represented by a *phylogenetic tree*. The leaves of the tree represent extant species. Each internal node represents a postulated *speciation event* in which a species divides into two populations that follow separate evolutionary paths and become distinct species.

The construction of a phylogenetic tree for a group of species is typically based on observed properties of the species. Before the era of genomics these properties were usually morphological characteristics such as the presence or absence of hair, fur, or scales or the number and type of teeth. With the advent of genomics the trees are often constructed by computer programs based on comparison of related DNA sequences or protein sequences in the different species.

An instance of the phylogeny construction problem typically involves n species and m *characters*. For each species and each character a *character state* is given. If, for example, the data consists of protein sequences of a common length aligned without gap symbols, then there will be a character for each column of the alignment, and the character state will be the residue in that position. The output will be a rooted binary tree whose leaves are in one-to-one correspondence with the n species.

The informal principle underlying phylogenetic tree construction is that species with similar character states should be close together in the tree. Different interpretations of this principle yield different formulations of the tree construction problem as an optimization problem, leading to several classes of tree construction methods. In all cases, the resulting optimization problem is NP-hard. We shall discuss *parsimony methods*, *distance-based methods*, and *maximum-likelihood methods*.

Parsimony Methods

The internal nodes of a phylogenetic tree are intended to represent ancestral species whose character states cannot be observed. In *parsimony methods* of tree construction the task is to construct a tree T and an assignment A of character states to the internal nodes to minimize the sum, over all edges of the tree, of the number of changes in character state along the edge.

Distance-Based Methods

Distance-based methods are based on the concept of an *additive metric*. Define a *weighted phylogenetic tree* T as a phylogenetic tree in which a nonnegative *length* $\lambda(e)$ is associated with each edge e. Define the distance between two species as the sum of the lengths of the edges on the path between the two species in the tree. The resulting distance function is called the *additive metric* realized by T.

Distance-based methods for phylogeny construction are based on the following assumptions:

1. There is a well-defined evolutionary distance between each pair of species, and this distance function is an additive metric.
2. The "correct" phylogenetic tree, together with appropriately chosen edge distances, realizes this additive metric.
3. The evolutionary distances between the extant species can be estimated from the character state data for those species.

This suggests the following *additive metric reconstruction problem*: Given a distance function D defined on pairs of species, construct a tree and a set of edge distances such that the resulting additive metric approximates D as closely as possible (for a suitable measure of closeness of approximation).

The following is an example of a simple stochastic model of molecular evolution which implies that the distances between species form an additive metric. The models used in practice are similar, but more complex.

The model is specified by a weighted phylogenetic tree T with edge lengths $\lambda(e)$. The following assumptions are made:

1. All differences between the character states of species are due to random mutations.
2. Each edge e represents the transition from an ancestral species to a new species. Independently for each character, the number of mutations during this transition has a Poisson distribution with mean $\lambda(e)$.
3. Whenever a mutation occurs, the new character state is drawn uniformly from the set of all character states (not excluding the character state that existed before the mutation).

The *neighbor-joining algorithm* is a widely used linear-time algorithm for the additive metric reconstruction problem. Whenever the given distance function D is an additive metric, the neighbor-joining algorithm produces a weighted tree whose metric is D. The neighbor-joining algorithm also enjoys the property of *asymptotic consistency*. This means that, if the data is generated according to the stochastic model described above (or to certain generalizations of that model) then, as the number of characters tends to infinity, the tree and edge weights produced by the neighbor-joining algorithm will converge to the correct tree and edge weights with probability one.

Maximum Likelihood Methods

Maximum likelihood methods for phylogeny construction are based on a stochastic model such as the one described above. Let A_x be the observed assignment of states for character x to the extant species. For any model $M = (T, \lambda)$ specifying the tree structure and the edge lengths, let $L_x(M)$ be the probability of observing the assignment A_x, given the model M. Define $L(M)$, the likelihood of model M, as the product of $L_x(M)$ over all characters x. The goal is to maximize $L(M)$ over the set of all models. This problem is NP-hard, but near-optimal solutions can be found using a combination of the following three algorithms:

1. an efficient algorithm based on dynamic programming for computing the likelihood of a model $M = (T, \lambda)$;
2. an iterative numerical algorithm for optimizing the edge lengths for a given tree; i.e., computing $F(T) = \max_\lambda L(T, \lambda)$;
3. a heuristic algorithm for searching in the space of trees to determine $\max_T F(T)$.

Genome Rearrangement

In comparing closely related species such as cabbage and turnip or man and mouse, one often finds that individual genes are almost perfectly conserved, but their locations within the genome are radically different. These differences seem to arise from global rearrangements involving the duplication, reversal, or translocation of large regions within a genome. This suggests that the distance between genomes should be measured not only by counting mutations, but also by determining the number of large-scale rearrangements needed to transform one genome to another.

To study these problems mathematically we view a genome as a sequence of occurrences of genes, define a set of primitive rearrangement operations, and define the distance between two genomes as the number of such operations needed to pass from one genome to the other. As an example, consider the case of two genomes that contain the same n genes but in different orders, and in which the only primitive operation is the reversal of a sequence of consecutive genes. Each genome can be modeled as a permutation of $\{1, 2, \ldots, n\}$ (i.e., a sequence of length n containing each element of $\{1, 2, \ldots, n\}$ exactly once), and we are interested in the *reversal distance* between the two permutations, defined as the minimum number of reversal operations required to pass from one permutation to the other. To make the problem more realistic we can take into account that genes are oriented objects and that the reversal of a segment not only reverses the order of the genes within it, but also reverses the orientation of each gene within the segment. In this case each genome can be modeled as a *signed permutation*, i.e., a permutation with a sign (+ or −) attached to each of the n elements, and the reversal operation reverses the order of a sequence of consecutive elements and the sign of each of these elements. It turns out that the problem of computing the reversal distance between two (unsigned) permutations is NP-hard, but there is an elegant quadratic-time algorithm for computing the reversal distance between two signed permutations.

DNA Microarrays

In this section we describe a key technology for measuring the abundances of specific DNA or RNA molecules within a complex mixture, and we describe applications of this technology to the study of associations between polymorphisms and disease, to the classification and clustering of genes and biological samples, and to the analysis of genetic regulatory networks.

Two oriented DNA molecules x and y are called *complementary* if y can be obtained by reversing x and replacing each base by its complementary base, where the pairs (A,T) and (C,G) are complementary. There is a similar notion of complementarity between RNA and DNA. *Hybridization* is the tendency of complementary, or nearly complementary, molecules to bind together.

Specific molecules within a complex sample of DNA or RNA can be identified by detecting their hybridization to complementary *DNA probes*. A DNA microarray is a regular array of DNA probes deposited at discrete addressable spots on a solid surface; each probe is designed to measure the abundance of a specific DNA or RNA molecule such as the mRNA transcript of a gene. It is possible to manufacture DNA microarrays with tens of thousands of spots on a surface the size of a postage stamp.

Here we concentrate on the applications of DNA microarrays, omitting all technological details about the manufacture of the arrays, the application of DNA or RNA samples to the arrays, and the measurement of hybridization. It is important to note, however, that at the present state of the art the measurements are subject to large experimental error. Methods of experimental design and statistical analysis are being developed to extract meaningful results from the noisy measurements, but currently one can obtain only a rough estimate of the abundance of particular molecules in the sample.

Associations between Polymorphisms and Disease

The genomes of any two humans differ considerably. Each of us carries different alleles (commonly occurring variant forms) of genes and *polymorphisms* (local variations in the sequence, typically due to mutations). Of particular interest are *Single-Nucleotide Polymorphisms* (SNPs) caused

by mutations at a single position. Several million commonly occurring SNPs within the human genome have been identified. It is of great interest to find statistical associations between a genotype (the variations within an individual's genome) and phenotype (observable characteristics such as eye color or the presence of disease). Some genetic diseases result from a single polymorphism, but more commonly there are many genetic variations that influence susceptibility to a disease; this is the case for atherosclerosis, diabetes, and the many types of cancer. Microarrays are a fundamental tool for association studies because they enable an experimenter to apply DNA probes for thousands of different polymorphisms in a single experiment. The statistical problems of finding subtle associations between polymorphisms and complex diseases are currently being investigated intensively.

Classification and Clustering Based on Microarray Data

Microarrays can be used to identify the genetic changes associated with diseases, drug treatments, or stages in cellular processes such as apoptosis (programmed cell death) or the cycle of cell growth and division. In such applications a number of array experiments are performed, each of which produces noisy measurements of the abundances of many gene transcripts (mRNAs) under a given experimental condition. The process is repeated for many conditions, resulting in a *gene expression matrix* in which the rows represent experiments, the columns represent genes, and the entries represent the mRNA levels of the different gene products in the different experiments. A fundamental computational problem is to find significant structure within this data. The simplest kind of structure would be a partition of the experiments, or of the genes, into subclasses having distinct patterns of expression.

In the case of *supervised learning* one is given independent information assigning a *class label* to each experiment. For example, each experiment might measure the mRNA levels in a leukemia specimen, and a physician might label each specimen as either an acute lymphoblastic leukemia (ALL) or an acute myeloid leukemia (AML). The computational task is to construct a decision rule that correctly predicts the class labels and can be expected to generalize to unknown specimens. In the case of *unsupervised learning* the class labels are not available, and the computational task is to partition the experiments into homogeneous clusters on the basis of their expression data.

Typically the number of genes measured in microarray experiments is in the thousands, but the classes into which the experiments should be partitioned can be distinguished by the expression levels of a few dozens of critical genes, with the other genes being irrelevant, redundant, or of lesser significance. Thus there arises the *feature selection problem* of identifying the handful of genes that best distinguish the classes inherent in the data. Sometimes a gene expression matrix contains local patterns, in which a subset of the genes exhibit consistent expression patterns within a subset of experiments. These local patterns cannot be discerned through a global partitioning of the experiments or of the genes but require identification of the relevant subsets of genes and of experiments. Research on the feature selection problem and on the problem of identifying local patterns is in its infancy.

Supervised Learning

Machine learning theory casts the problem of supervised learning in the following terms. Given a set of *training examples* $\{x_1, x_2, \ldots, x_n\}$ drawn from a probability distribution over R^d, together with an assignment of a class label to each training example, find a rule for partitioning all of R^d into classes that is consistent with the class labels for the training examples and is likely to generalize correctly, i.e., to give correct class labels for other points in R^d.

There is a general principle of machine learning theory which, informally stated, says that, under some smoothness conditions on the probability distribution from which the training examples are drawn, a rule is likely to generalize correctly if

1. each training example lies within the region assigned to its class and is far from the boundary of that region;
2. the rule is drawn from a "simple" parametrized set of candidate rules. The notion of simplicity involves a concept called the Vapnik-Chervonenkis dimension, which we omit from this discussion.

One reasonably effective decision rule is the *nearest neighbor rule*, which assigns to each point the same class label as the training example at minimum Euclidean distance from it.

In the case of two classes (*positive examples* and *negative examples*) the *support vector machine* method is often used. It consists of the following two stages:

Mapping into feature space: Map each training example x to a point $\phi(x) = (\phi_1(x), \phi_2(x), \ldots)$, where the ϕ_i are called *features*. The point $\phi(x)$ is called the *image* of training example x.

Maximum-margin separation: Find a hyperplane that separates the images of the positive training examples from the images of the negative training examples and, among such separating hyperplanes, maximizes the smallest distance from the image of a training example to the hyperplane.

The problem of computing a maximum-margin separation is a quadratic programming problem. It turns out that the only information needed about the training examples is the set of inner products $\phi(x)^T\phi(y)$ between pairs (x, y) of training examples. It is possible for the images of the training examples to be points in an infinite-dimensional Hilbert space, as long as these inner products can be computed. *Mercer's Theorem* characterizes those functions $K(x, y)$ which can be expressed as inner products in finite- or infinite-dimensional feature spaces. These functions are called *kernels*, and the freedom to use infinite-dimensional feature spaces defined through their kernels is a major advantage of the support vector machine approach to supervised learning.

Clustering

Clustering is the process of partitioning a set of objects into subsets based on some measure of similarity (or dissimilarity) between pairs of objects. Ideally, objects in the same cluster should be similar, and objects in different clusters should be dissimilar.

Given a matrix of gene expression data, it is of interest to cluster the genes and to cluster the experiments. A cluster of genes could suggest either that the genes have a similar function in the cell or that they are regulated by the same transcription factors. A cluster of experiments might arise from tissues in the same disease state or experimental samples from the same stage of a cellular process. Such hypotheses about the biological origin of a cluster would, of course, have to be verified by further biochemical experiments.

Each gene or experiment can be viewed as an n-dimensional vector, with each coordinate derived from a measured expression level. The similarity between points can be defined as the inner product, after scaling each vector to Euclidean length 1. When experiments are being clustered the vectors are of very high dimension, and a preliminary feature selection step is required to exclude all but the most salient genes.

In the K-means algorithm the number of clusters is specified in advance, and the goal is to minimize the sum of the distances of points from the centers of gravity of their clusters. Locally optimal solutions can be obtained by an iterative computation which repeats the following step: Given a set of K clusters, compute the center of gravity of each cluster; then reassign each point to the cluster whose center of gravity is closest to the point.

Maximum likelihood methods assume a given number of clusters and also assume that the points in each cluster have a multidimensional Gaussian distribution. The object is to choose the parameters of the Gaussian distributions so as to maximize the likelihood of the observed data. In these methods a point is not definitely assigned to a cluster, but is assigned a probability of lying in each cluster.

Merging methods start with each object in a cluster by itself and repeatedly combine the two clusters that are closest together as measured, for example, by the distance between their centers of gravity. Most merging and splitting methods that have been proposed are heuristic in nature, since they do not aim to optimize a clearly defined objective function.

The Logic of Transcriptional Control

Cellular processes such as cell division, programmed cell death, and responses to drugs, nutrients, and hormones are regulated by complex interactions among large numbers of genes, proteins, and other molecules. A fundamental problem of molecular life science is to understand the nature of this regulation. This is a very formidable problem whose complete solution seems to entail detailed mathematical modeling of the abundances and spatial distributions within a cell of thousands of chemical species and of the interactions among them. It is unlikely that this problem will be solved within this century.

One aspect of the problem that seems amenable to mathematical methods is the logic of transcriptional control. As Eric Davidson has stated, "A large part of the answer lies in the gene control circuitry encoded in the DNA, its structure and its functional organization. The regulatory interactions mandated in the circuitry determine whether each gene is expressed in every cell, throughout developmental space and time, and if so, at what amplitude. In physical terms the control circuitry encoded in the DNA is comprised of *cis*-regulatory elements, i.e., the regions in the vicinity of each gene which contain the specific sequence motifs at which those regulatory proteins which affect its expression bind; plus the set of genes which encode these specific regulatory proteins (i.e., transcription factors)."

It appears that the transcriptional control of a gene can be described by a discrete-valued function of several discrete-valued variables. The value of the function represents the level of transcription of the gene, and each input variable represents the extent to which a transcription factor has attached to binding sites in the vicinity of the gene. The genes that code for transcription factors are themselves subject to transcriptional control and also need to be characterized by discrete-valued functions. A regulatory network, consisting of many interacting genes and transcription factors, can be described as a collection of interrelated discrete functions and depicted by a "wiring diagram" similar to the diagram of a digital logic circuit.

The analysis of this control circuitry involves biochemical analysis and genomic sequence analysis to identify the transcription factors and the

sequence motifs characteristic of the sites at which they bind, together with microarray experiments which measure the transcriptional response of many genes to selected perturbations of the cell. These perturbations may involve changes in environmental factors such as temperature or the presence of a nutrient or drug, or interventions that either disable selected genes or enhance their transcription rates. The major mathematical challenges in this area are the design of informative perturbations and the inference of the transcriptional logic from information about transcription factors, their binding sites, and the results of microarray experiments under perturbed conditions.

Bibliography

In this section we provide references for the principal topics introduced in this article.

Sequence Comparison

[1] D. GUSFIELD, *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, 1997.

Hidden Markov Models; Phylogeny Construction

[2] R. DURBIN, S. EDDY, A. KROGH, and G. MITCHISON, *Biological Sequence Analysis*, Cambridge University Press, 1998.

Sequence Assembly

[3] EUGENE W. MYERS et al., A whole-genome assembly of Drosophila, *Science* **287** (2000), 2196–2204.

Gene Finding

[4] S. SALZBERG, D. SEARLS, and S. KASIF (Eds.), *Computational Methods in Molecular Biology*, Elsevier Science, 1998.

Genome Rearrangement

[5] P. PEVZNER, *Computational Molecular Biology*, MIT Press, 2000.

DNA Microarrays

[6] M. SCHENA (Ed.), *DNA Microarrays: A Practical Approach*, Oxford University Press, 1999.

Supervised Learning

[7] N. CRISTIANINI and J. SHAWE-TAYLOR, *An Introduction to Support Vector Machines*, Cambridge University Press, 1999.

Clustering

[8] B. EVERITT, S. LANDAU, and M. LEESE, *Cluster Analysis*, Edward Arnold, fourth edition, 2001.

The Logic of Transcriptional Control

[9] E. DAVIDSON, *Genomic Regulatory Systems*, Academic Press, 2001.