

Bioinformatics

by

Alan J. Izenman

Introduction

The term *bioinformatics* was coined in the 1990's to describe an emerging field which combined computer science, information technology, and biological science. The catalyst of this new discipline was the competitive race to sequence the human genome. The Human Genome Project, a government-sponsored consortium of universities and laboratories, and Celera, a private company, simultaneously published draft accounts of the human genome in *Nature* and *Science* on 15th and 16th February 2001, respectively; see Dennis and Gallagher (2001). This was a momentous event in the history of science.

Recent advances in computation and sequencing technology has provided the discipline of bioinformatics with exciting new possibilities. These include: creating a growing number of databases to store, manage, and manipulate huge quantities of biological data; constructing efficient algorithms to analyze DNA, RNA, and protein sequence data; studying gene expression data through microarray technology; and helping the pharmaceutical industry to develop new drugs to treat diseases. These advances in biotechnology first led to *genomics* (the study of gene activity and expression), and then *proteomics* (the study of proteins); in the future, we expect the field of *metabolomics* (the study of metabolic pathways) to increase in importance. One of the most important tools of bioinformatics is the Internet, where researchers can now access sequence data or microarray data from any of a number of databases located around the world.

Today, statisticians, computer scientists, and biologists collaborate on bioinformatics research. The tools needed to analyze microarray data and to visualize relationships between sequence data include many of the multivariate statistical techniques described in this book.

The Flow of Genetic Information

The original and simplest version of the “central dogma” of molecular biology (Crick, 1970) says that, in all cells, genetic information flows in the direction DNA \rightarrow RNA \rightarrow proteins, where every cell from the same individual carries essentially identical DNA (deoxyribonucleic acid) in its nucleus. There are violations of this dogma, as we will see below. Cells, though similar in general structure, possess different properties, depending upon function and location in the body.

DNA is a double-stranded (*duplex*) polymer looking much like a ladder twisted into a helix. The two strands confer chemical stability on the DNA molecule and each strand can act to correct possible damage should it occur in the other strand. Each DNA *strand* is composed of an extremely-long sequence of *nucleotides*. Each nucleotide consists of a phosphate molecule and a deoxyribose sugar molecule (forming the *sugar-phosphate backbone* of DNA), and one of four nitrogen *bases*: adenine (A), cytosine (C), guanine (G), and thymine (T). A DNA strand can, therefore, be thought of as a “word” formed by letters repeatedly selected from the 4-letter alphabet $\{A, C, G, T\}$. The ‘rungs’ of the ladder are hydrogen bonds joining each letter in one strand of DNA to a letter in the other (complementary) strand so that *base pairs* (bps) are formed: A always pairs with T and C always pairs with G. Human DNA has about 3 billion (3×10^9) bps. The sequence of bases on each strand is read from the 5′ end to the 3′ end, so that the two strands are read in opposite directions. The terminology 3′ and 5′ derives from the chemical structure of DNA: 5′ refers to the start (or left-hand-side) of the stand, while 3′ refers to the end (or right-hand-side) of the strand.

A *genome* is the complete set of genetic instructions for creating an organism. In the case of the human genome, the 3 billion bps in human DNA are organized into 24 distinct, physically-separated, microscopic, rod-like structures called *chromosomes*, which live inside the cell nucleus, plus an additional 16,569 bps which make up mtDNA (*mitochondrial DNA*) and live outside the cell nucleus. Mitochondrial DNA is maternally inherited and is studied for ancestral inference (see, e.g., Griffiths and Tavaré, 1994).

Each chromosome consists of a single, continuous, tightly-coiled stretch of DNA wrapped around protein structures called *histones*. There are a total of 46 chromosomes in the nucleus of most human cells, and these are divided into two sets of 23, one set inherited from each parent. Each set of chromosomes contains 22 *autosomes* and an *X* or *Y* sex chromosome. A normal female will have a pair of *X* chromosomes and a normal male will have one *X* and one *Y* chromosome. Human autosomes are numbered by size, with chromosome 1, the largest, having 285 million bps, and chromosomes 21 and 22, the two smallest, having 44 million bps and 47 million bps, respectively. The *X* chromosome has 168 million bps and the *Y* chromosome has 51 million bps. Different species have different numbers of chromosomes.

Along each chromosome are genes ordered in a linear fashion. Some of these genes are *protein-coding genes* (i.e., specific segments of a DNA sequence, where each segment carries all the necessary instructions for manufacturing a particular protein), while others produce a type of RNA (ribonucleic acid) which does not code for proteins (*noncoding-RNA* or ncRNA). The human genome is currently estimated to consist of about 24,500 protein-coding genes (Pennisi, 2003), but even that number is still not considered to be a final count.

Gene Expression

These protein-coding genes are *expressed* through the two stages of transcription and translation. The *transcription* stage synthesizes the DNA segment coding the gene into a single strand of mRNA (*messenger RNA*). This is accomplished in several steps. First, an enzyme, *RNA polymerase*, breaks the weak hydrogen bonds forming the base-pairs, thereby partially unzipping the double-helix DNA to expose the gene. Second, the RNA polymerase operates on the unzipped portion of the complementary DNA strand, working in the $3' \rightarrow 5'$ direction, and assembles *ribonucleotides* into a strand of *pre-mRNA* using the rules $A \rightarrow U$, $T \rightarrow A$, and $C \leftrightarrow G$, where the base U (uracil) is substituted for T (thymine). Thus, a pre-mRNA word is composed from the 4-letter alphabet $\{A, C, G, U\}$.

In most sophisticated organisms, genes actually occur in pieces along the DNA sequence: regions of protein-coding DNA (or *exons*) alternate with regions of seemingly-irrelevant, noncoding DNA (or *introns*). So far, little is known of the origins and functions of these introns, even though in many cases they take up huge portions of the entire DNA. For example, less than 2% of human DNA is used for coding proteins. However, there is an increasing body of evidence which shows that introns carry instructions for controlling protein expression.

In the third step of the transcription process, the introns are “spliced out” of the pre-mRNA and discarded. Such processing produces *mature mRNA* consisting of an uninterrupted sequence of exons. By splicing together different combinations of exons within a single gene, the average human gene can create several different proteins. Any errors in this editing process would probably result in the loss of functionality of the protein. Because the DNA molecule cannot leave the cell nucleus, the genetic information in the DNA is carried from the cell nucleus out to the cytoplasm (where proteins are synthesized) by the mature mRNA acting purely as an intermediary.

In the *translation* stage, the mature mRNA is used as a template by a *ribosome* (a combination of proteins and ncRNA) to manufacture the *primary structure* of a specific protein. In this scenario, ncRNA is referred to as rRNA (*ribosomal RNA*). Although ribosomes exist in abundance in human cells, they usually attach themselves (in groups of at least 10) to an mRNA molecule, spacing themselves out like beads on a thread. Each ribosome moves along the mRNA molecule, reading in the $5' \rightarrow 3'$ direction. As it moves, the ribosome reads the mRNA in nonoverlapping 3-letter chunks (called *codons*) and translates each codon into an amino acid. In this manner, the amino acids are assembled into a gradually-lengthening protein chain.

Protein Structure

The 20 amino acids (viewed as a 20-letter alphabet) are given in Table 1.3. The *genetic code*, which specifies the codons by relating the 20-letter alphabet of a protein to mRNA’s 4-letter alphabet, is given in Table 1.4. Protein translation

TABLE 1. *The 20 amino acids (and their 3-letter and 1-letter abbreviations).*

Alanine (ala , A), Arginine (arg , R), Asparagine (asn , N), Aspartic acid (asp , D), Cysteine (cys , C), Glutamine (gln , Q), Glutamic acid (glu , E), Glycine (gly , G), Histidine (his , H), Isoleucine (ile , I), Leucine (leu , L), Lysine (lys , K), Methionine (met , M), Phenylalanine (phe , F), Proline (pro , P), Serine (ser , S), Threonine (thr , T), Tryptophan (trp , W), Tyrosine (tyr , Y), Valine (val , V)
--

is initiated by the codon AUG (when it appears at the beginning of a gene) and terminated by any of the three codons UAA, UAG, and UGA. When translation stops, the ribosome detaches itself from the mRNA, and the completed protein is released.

As an example, the *hemoglobin alpha chain* (Swiss Protein [Swiss-Prot] database code HBA_HUMAN, AC# P01922) is a human protein composed of a linear (polypeptide) chain of 141 amino acid molecules bound together by peptide bonds. Its amino acid sequence is given by:

```
VLSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSH  
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSADLHAHKLKLRVDPVNFKL  
LSHCLLVTLAAHLPAEFTPAVHASLDFKFLASVSTVLTSKYR
```

This protein transports oxygen from the lungs to the various peripheral tissues and gives blood its red color.

The number of all possible polypeptide chains of a particular length can be huge (e.g., for a 10-chain sequence, there would be $10^{20} = 10,240$ billion different sequences). Typically, there are 100–500 amino acids molecules in a chain, although some chains are much shorter (e.g., hormones with chain lengths of 25–100) and others much longer (e.g., fibrous proteins with chain lengths of 3000 or more).

The *secondary structures* of the protein are produced when the polypeptide chain is organized into regular repetitive patterns over short segments of 5–15 amino acids. The most important of the secondary structures are α -*helices* and β -*sheets*. Of all the amino acids in proteins, about one-third form α -helices, while about a quarter form β -sheets, proportions which vary from protein to protein. The β -sheets, formed by the hydrogen-bonding of a number of β -*strands*, are classified as either *parallel* or *antiparallel*, depending upon whether the strands run in the same or opposite directions, respectively. Mixtures of parallel and antiparallel β -sheets can also occur.

Proteins are usually classified by their secondary structures: α proteins are structures dominated by α -helices; β proteins are predominantly β -sheet struc-

TABLE 2. *The genetic code. There are $4^3 = 64$ codons whose first, second, and third letters define the amino acid. The codon AUG signals the start of and the three codons, UAA, UAG, and UGA, stop protein translation.*

2nd 1st letter	2nd letter				3rd letter
	U	C	A	G	
U	phe	ser	val	cys	U
	phe	ser	val	cys	C
	leu	ser	stop	stop	A
	leu	ser	stop	trp	G
C	leu	pro	his	arg	U
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	ile	thr	asn	ser	U
	ile	thr	asn	ser	C
	ile	thr	lys	arg	A
	met/ start	thr	lys	arg	G
G	val	ala	asp	gly	U
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G

tures; α/β proteins are characterized by the regular alternation of α -helices and β -sheets; and $\alpha + \beta$ proteins are characterized by the irregular alternation of α -helices and β -sheets.

For a protein to do its job properly, the polypeptide chain self-assembles (or “folds”) in a very specific and characteristic way, which yields the three-dimensional structure of the protein (also called *tertiary structure* or *conformation*). Folding is achieved very quickly; in some instances, on the order of a millionth of a second. If a chain is not folded appropriately (possibly caused by a mutation in the gene encoding the protein), its function is altered, and the result could lead, for example, to Alzheimer’s disease, cystic fibrosis, or bovine spongiform encephalopathy (BSE or mad-cow disease). Special proteins called *molecular chaperones* have been discovered which oversee the correct folding and assembly of other proteins without themselves being components of the final folded structure. An extremely challenging problem is the “protein folding problem”: predict the tertiary (3-D) structure and function of a protein from its primary (linear) structure, the amino acid sequence (Neumaier, 1997).

DNA Microarrays

DNA microarray technology has revolutionized the way molecular biologists and others in related biomedical, pharmaceutical, and clinical fields carry out gene analysis. The microarray has been described as “one of the great unintended consequences of the Human Genome Project” (Baker, 2003) and has been compared to the invention of the telescope and the microscope. Furthermore, the statistical community has become dazzled by a vision of “large and highly structured [microarray] data sets waiting to be mined for valuable information” (Churchill, 2003).

In recent years, we have witnessed the remarkable achievement whereby entire genomes have been sequenced. Given information on a particular genome, we would like to know, for example, whether gene expression is any different for cancerous tissue as opposed to healthy tissue. Instead of having to focus on a single gene at a time to answer such questions, we can now study an organism’s entire genome in a single experiment. Microarray technology has enabled the *expression levels* of a huge number of genes within a specific cell culture or tissue to be monitored, simultaneously and efficiently. This is important because differences in gene expression determine differences in protein abundance, which, in turn, determine different cell functions. Although protein abundance is difficult to determine, molecular biologists have discovered that gene expression can be measured indirectly through microarray experiments.

The keys to the microarray process are those of reverse transcription and hybridization. *Reverse transcription* provides one of the exceptions to the central dogma of molecular biology: it uses the enzyme *reverse transcriptase* to produce single-stranded cDNA (*complementary DNA*) from mRNA, where the cDNA sequence is the complement of the mRNA sequence. *Hybridization* (or *annealing*) refers to the process by which two complementary strands of nucleotides are base-paired with each other to form a double-stranded molecule.

Microarray analysis takes place on a glass or plastic microscope slide, usually referred to as a *microarray slide*. The two most popular types of microarray technologies are: cDNA microarrays (developed at Stanford University) and high-density, synthetic, oligonucleotide microarrays (developed by Affymetrix, Inc., under the GENECHIP® trademark). Both technologies use the idea of hybridizing a “target” (which is usually either a single-stranded DNA or RNA sequence, extracted from biological tissue of interest) to a “probe” (all or part of a single-stranded DNA sequence printed onto the microarray slide), and then measuring the intensities of the resulting gene expression. A microarray can generate intensity values for many thousands of genes.

The most useful application of microarray technology resides in the simultaneous comparison of expression levels of many thousands of genes, where each gene may be examined over a number of different conditions. Such “conditions” include different experimental conditions (treatment vs. control samples), different tissue samples (healthy vs. cancerous tumors), and different time points (which may incorporate environmental changes). In certain microarray exper-

iments, these conditions can be compared using a set of different microarray slides (see, e.g., Efron, Tibshirani, Storey, and Tusher, 2001).

Spotted cDNA Microarrays

A DNA *probe* is a long subsequence (500-5,000 bps) of a gene sequence, or possibly even the entire gene sequence. Although the identity of the sequence will be known, its function may be unknown. Thousands of individual DNA probes are printed as *spots* onto a two-channel microarray slide using a high-speed robotic *arrayer*; the spots, each of which corresponds to a specific gene, are deposited onto a two-way grid of dimples in a microarray slide. The latest microarray technology allows arrays of up to 50,000 spots to be produced. For reliability and technical reasons, it is standard procedure to distribute replicated spots for each gene randomly over the microarray slide.

The microarray slide is then exposed to a set of *targets*. Two biological mRNA samples, one obtained from cancerous tissue (the *experimental sample*), the other from healthy tissue (the *reference sample*), are reverse-transcribed into cDNA; then, the reference cDNA is labelled with a green fluorescent dye (e.g., Cy3) and the experimental cDNA is labelled with a red fluorescent dye (e.g., Cy5). The labelled cDNA are mixed in equal proportions to provide a two-color target, which is then hybridized to the probe DNA on the microarray slide. The hybridization process becomes extremely competitive when the same gene is present in both types of samples; the red and green cDNA both compete with each other to hybridize with each of the complementary spots printed on the array.

Following this competitive hybridization, the microarray is washed and dried. Then, a scanning microscope produces two high-resolution black-and-white digital images, one for the red channel and another for the green channel, of the fluorescence intensities at each pixel. For the technical details, see Yang, Buckley, Dudoit, and Speed (2002).

A composite array image, which is created by combining the two images (either by overlaying one image on the other or by merging the two images into a single image), is a grid-like pattern of spots. If we assume that a tumor sample is labelled red and a healthy tissue sample is labelled green, then each spot is synthetically colored as green (if a gene is expressed in the healthy tissue, but not in the tumor), red (if a gene is expressed in the tumor, but not in the healthy tissue), yellow (if a gene is expressed in both tissues), or grey (if a gene is expressed in neither tissue). Fluorescence measurements are taken of each dye separately at each spot on the array. High gene expression in the tissue sample yields large quantities of hybridized cDNA, which means a high intensity value. Low intensity values derive from low gene expression.

The primary goal is to compare the intensity values, R and G, of the red and green channels, respectively, at each spot on the array. The most popular statistic is the *intensity log-ratio*, $M = \log(R/G) = \log(R) - \log(G)$. Other

such functions include the *probe value*, $PV = \log(R - G)$. Another statistic is the *average log intensity*, $A = \frac{1}{2}(\log R + \log G)$. An *MA-plot* is a scatterplot of (A, M) , which is a 45-degree rotation and scaling of the *log-intensity plot* of $(\log G, \log R)$. One useful feature of the MA-plot is in exposing unusual spot-intensity values (Dudoit, Yang, Callow, and Speed, 2002). The logarithm in each case is taken to base 2 because intensity values are usually integers ranging from 0 to $2^{16} - 1$.

Oligonucleotide Microarrays

An *oligonucleotide sequence* (or *oligo*) is a short subsequence (25 bp) of a cDNA sequence, which matches a selected fragment of a known gene. Oligos are arranged as *probe-pairs*: the first oligo of the pair, which is the exact sequence of the particular gene fragment, is called a *perfect match* (PM) *probe*, while the second oligo, which is a control created from the PM sequence by replacing the middle (13th) nucleotide with its complement, is called a *mismatch* (MM) *probe*.

Replication plays a big role here in order to compensate for the short fragments of cDNA sequences used in this process. Each gene is represented by a *probe set* composed of 11 to 20 different oligo probe-pairs. In total, the probe set spans a region of at most 500 bases from the entire gene sequence. These oligo probe-pairs are placed on a small silicon chip using a photolithographic fabrication device (similar to that used to construct very-large-scale integrated (VLSI) computer circuits). The probes are arranged on the microarray in a grid-like pattern, which is used to identify and locate probe sets. To reduce potential bias, probe-pairs from the same probe set are randomly sprinkled all over the microarray, with the PM cell placed directly above its associated MM cell.

Once the microarray has been completed, RNA is extracted from a single tissue sample and the mRNA is reverse-transcribed into cDNA; the cDNA is made double-stranded and transcribed (*in vitro*) into cRNA (*complementary RNA*). The target cRNA is then labelled with a fluorescent dye and hybridized to the array. Only the PM probe should hybridize; because of the single mismatching nucleotide, the MM probe should not hybridize. However, we do see *nonspecific hybridization* taking place, in which the cRNA hybridizes either to background material or to an MM probe which is not its exact complement.

Following hybridization, the microarray is washed, stained with a fluorescent molecule, and scanned with a confocal laser. The result is a high-resolution digital image showing fluorescence intensity values at the pixels making up the image.

Two vectors of intensity values are obtained as output, one for PMs and one for MMs. One measure included in the GENECHIP software is the **detection call**, which focusses on the number of probes for a particular gene where the PM-value is greater than the MM-value. If many PM-values are greater than their MM-values, the gene is declared to be “present,” while if the reverse holds,

then the gene is declared to be “absent.” If the numbers are roughly equal, the gene is declared to be “marginal.”

Quantitative measures of expression level of each gene are also provided in the GENECHIP software. They include: **AvDiff**, the average of the PM – MM differences for all probes in the probe set corresponding to that gene; and **Signal**, a robust estimator, which applies Tukey’s biweight function to the probe-set values of $\log(\text{PM} - \text{MM}^*)$, where MM^* is constructed from MM so that $\text{PM} - \text{MM}^* > 0$.

Statistical Issues for Analyzing Microarray Data

Microarray data are multivariate data which take the form of a matrix of genes (rows) by samples (columns). The genes play the role of variables, while the samples are the observations. Given the enormous amount of data obtained from microarray experiments, interest typically focusses on reducing those data to a few well-chosen summary statistics. Visualization of the data is also important for detecting unusual patterns and anomalies.

We are generally interested in analyzing gene expression data in which samples are studied under a set of different “conditions” (e.g., healthy vs. diseased tissue, various treatment groups, environmental factors). Multivariate statistical techniques applied to microarray data, where the samples are divided into classes corresponding to those known conditions, include the supervised learning methods of discriminant analysis and classification (including tree-based classifiers, random forests, support vector machines). If we ignore the class designations, we can use the unsupervised methods of cluster analysis for grouping together either genes or samples; cluster analysis happens to be the most popular multivariate tool for analyzing microarray data. We may also be interested in drawing gene-expression maps using multidimensional scaling.

Variable selection is an especially important statistical tool when dealing with situations in which there are more variables than observations. Translated to the microarray context, where the typical microarray has many thousands of genes and fewer than 100 samples, the problem is to find those genes which are differentially expressed in the various classes and are also the most important in discriminating between those classes.