Bioinformatics: Microarrays Analyses and Beyond

We have witnessed in the past years the rapid progresses in the human genome project and biotechnologies. These advances result in many complex datasets associated with in-depth scientific knowledge, e.g., genome sequences of many species, microarray expression profiles of different cell lines, single nucleotide polymorphisms (SNPs) in the human genome, etc. These data together with their underlying scientific challenges spawn the new field of Bioinformatics, which sprawls many academic disciplines as well as the pharmaceutical industry, and create one of the most exciting times for all quantitative researchers. There is no doubt that statistics will be pivotal in this new field, but it remains a challenge to us statisticians whether we can play a leading role in this biology and informatics revolution. This is not just a challenge, in fact, but also a golden opportunity for our discipline.

The recent developments of two high throughput biological data generation technologies help foster the bioinformatics hype in statistics: the genome sequencing technology and the DNA chip technology. The word “genome” refers to the collection of all the chromosomes (chains of DNA bases; human has 23 pairs of these) in a cell. Certain segments of the genome, called “genes” (or coding regions), encode the information needed to make proteins, which are action molecules of the cell, responsible for nearly all cellular processes. It is estimated that the human genome has about 30,000 genes, which, surprisingly, only account for ~3% of the genome. The expression of these genes, i.e., the amount of protein products to be made in a cell, is tightly regulated so as to meet the requirements of specific cells and for cells to respond to changes in their environment. A central goal of molecular biology is to understand the regulation of protein synthesis.

In order to make a protein molecule, a gene is first transcribed to messenger RNA (mRNA), an easily degradable molecule, which then carries the information to a cellular machinery (ribosome) for protein production (Figure 1). While
1. Transcription

Transcriptions take place inside nuclei, during which the RNA polymerase uses one strand of the DNA double helix as a template to synthesize an mRNA. This mRNA then migrates from the nucleus to the cytoplasm after going through several maturation steps, including splicing. The coding mRNA sequence, which can be described as units of three nucleotides called codons, is bound by ribosome to start the translation stage (i.e., protein production). During this stage, the amino acids are added one by one with the help of tRNAs as ribosome moves from codon to codon along the mRNA.

Figure 1: The process of eukaryotic protein synthesis. (Picture taken from the Graphics Gallery Web site of Access Excellence at the National Health Museum www.accessexcellence.org/AB/GG/protein_synthesis.html. Transcriptions take place inside nuclei, during which the RNA polymerase uses one strand of the DNA double helix as a template to synthesize an mRNA. This mRNA then migrates from the nucleus to the cytoplasm after going through several maturation steps, including splicing. The coding mRNA sequence, which can be described as units of three nucleotides called codons, is bound by ribosome to start the translation stage (i.e., protein production). During this stage, the amino acids are added one by one with the help of tRNAs as ribosome moves from codon to codon along the mRNA.)
there are several levels of gene regulation, the dominant form is transcriptional regulation. Specific sequence signals upstream of each gene provide a target, called the promoters, for RNA polymerase (a machinery for transcription) to bind so as to initiate the transcription. When transcription factors (TFs, proteins specialized in regulating gene expressions) bind near the promoter region of a gene, they interfere with the function of RNA polymerase, thus, either repressing or enhancing the production of mRNA. The amount of a certain mRNA copies in a cell reflect, albeit imperfectly, the expression level of the corresponding gene.

**What is the microarray?**

By orderly arranging samples, the microarray provides a large-scale medium for matching known and unknown DNA segments based on base-pairing rules. There are two classes of microarrays. The cDNA arrays apply to glass slides (or nylon membranes) spots of *complimentary* DNAs (cDNAs), which are generated in biological labs by reverse transcription (so that they only include the protein-coding part of the genome). The oligonucleotide arrays (often referred to as the Affymetrix arrays) place many thousands of gene-specific oligonucleotides (called probes) synthesized directly on a silicon chip. The probes are about 25 base pairs long, and 20 probe-pairs (one perfect match and one mismatch) are often used to represent each gene (like a 20 digit barcode).

In order to compare two types of cells (e.g., a cancer cell versus a normal cell), for example, the biologist first extracts the DNA materials from all the cells and labels those from one cell type (say, cancer cell) by fluorescence cy5 (red) and the other cell type by cy3 (green). The microarray is then exposed to the mixture of the two DNA samples for hybridization. When mRNA for a gene is more abundant in the cancer cell than in the normal cell, for example, the array spot corresponding to that gene will show a red color. Numerically, a vector of length $G$ is reported, where $G$ is the number of spots (genes) in the array, and each entry of the vector records the ratios of the fluorescence intensities (cy5/cy3). When more than two types of cells are in consideration, the microarray data often takes the form of a Gxp matrix, where each column corresponds to a cell type (e.g., lymphoma cell, leukemia cell, normal cell, etc.) or a treatment, and each row corresponds to a gene. Thus, through the use of DNA microarrays, one can monitor simultaneously the expression levels of thousands of genes in different types of cells.

**The role of statistics**

The amount of data produced by microarray experiments is daunting even to statisticians. An important pre-processing step, often termed as "low-level" normalization method is to subtract a constant from the expression measurements of all the genes. But as demonstrated by Li and Wong (2001), Schadt et al. (2000), Tseng et al. (2001), and Yang et al. (2002), such an approach can be problematic due to certain expression intensity-dependent biases. More sophisticated statistical approaches using "rank invariant" genes or robust curve estimation (e.g., "LOESS") are often more appropriate.

A central task intended for the microarray experiment is to find genes that are differentially expressed in the two samples (or types of cells). Suppose that the identical microarray experiment is repeated $p$ times (e.g.,
become as prominent as "Eisen clustering." There is also no definitive conclusion as to which method is the optimal choice. With gene clusters available, one may be able to use motif-searching tools (Liu et al. 1995) to help infer groups of co-regulated genes (Roth et al. 1998, Liu et al. 2002b). A further and much more difficult challenge is to infer gene regulatory pathways (i.e., the cascade of genes that lead to cellular function).

Closely related to clustering is the classification or supervised learning problem. For example, Golub et al. (1999) were interested in predicting the two subtypes of leukemia based on the gene expression profile of each sample. In such problems, one has a "training dataset" (usually of very small size) in which the class indicator for each sample is known, and wants to generate a good "rule" for predicting a future sample. This is where various "statistical learning" techniques come into the play. For example, Fisher's linear discriminant analysis, the nearest neighbor classification, support vector machines, Bayesian networks, classification and regression trees, boosting, bagging, logistic regressions, independent component analysis, etc., have all been applied to the array data. New techniques are still being developed.

Since typically thousands to tens of thousands of genes are surveyed in a microarray study, it is of interest to select a small subset of genes that can best characterize the two groups. This is of great value to the pharmaceutical industry because of their need to find effective biomarkers for monitoring treatments and for defining a subpopulation that response to a certain drug. Sometimes one may measure a have time series measurements of gene expressions (e.g., cell cycle data). The current techniques (e.g., hierarchical clustering, k-means, SOM, etc.) treat these time points as exchangeable. Although the singular-value decomposition method has been used for understanding the yeast cell cycle data, a time-series model-based clustering technique is also valuable.

Integration with other array data

More biological data of similar nature to DNA microarrays are becoming available. Among the many array technologies, Chromatin Immunoprecipitation (ChIP) combined with microarrays (Ren et al. 2000), the so-called "ChIP-chip" data, has recently become popular for studying in vivo interactions between transcription factors (TFs) and their target binding sites in the genome (Ren et al. 2000, Lieb et al. 2001). In this procedure, the expressions of those DNA segments that are bound by the TF of interest are enhanced. Thus, when mixed with a normal cell extract and hybridized to microarrays, the spots corresponding to those TF binding sites will light up. By combining ChIP-chip data with the gene expression data, scientists can often gain more insights on how the regulatory network should be mapped out (Simon et al. 2001). The ChIP-chip data can also be combined with the questions (Efron et al. 2001, Chen et al. 2002, Newton et al. 2001, West et al. 2001).

Another set of important and related tasks, often termed as unsupervised learning, is to find genes that behave similarly in various conditions (i.e., clustering the row vectors), and to find subgroups of samples (or patients' tissues) that are similar to each other (i.e., clustering the column vectors). While the first task can lead the biologists to novel discovery of genes in related biological pathways or having related functions (Spellman et al. 1998), the second task can result in clinically important subgroups of patients. Due to the influential Eisen et al. (1998) article and its associated software, the clustering method of choice for biologists has been hierarchical clustering (Alizadeh et al. 2000). Other methods such as the k-means method, self-organized maps (Tamayo et al. 1998), Gaussian mixture models (Yeung et al. 2001, Liu et al. 2002a), plaids models (Lazzeroni and Owen, 2002), etc., have later been applied to microarrays, although none of them... (continued on page 64)
genome sequence information for discovering the exact regulatory motif sites—patterns (Liu et al. 2002b).

Many important cellular tasks are achieved by interactions between proteins—they may interact to pass on signals (part of a signal transduction pathway) or to form a complex for tackling a difficult job (e.g., transcription or translation). In conjunction with high throughput expression and purification of recombinant proteins, biologists can prepare microarrays of functionally active proteins on glass slides. These arrays can then be used to identify protein-protein interactions, to identify the substrates of protein kinases, or to identify the targets of biologically active small molecules (MacBeath and Schreiber 2000). Another technology, the yeast two-hybrid system, has also been used successfully to investigate protein-protein interactions (Walhout and Vidal 2001). The integration of the protein interaction data with the DNA microarray data has revealed some interesting connections between expression profiles of the genes and the interactions among their product proteins (Ge et al. 2000).

Other promising array technologies are also under development. For example, the small molecule microarrays can be used to screen large libraries of compounds to identify new ligands for proteins of interest (MacBeath et al. 1999), antibody arrays can be used to study regulation at the protein level, and SNP arrays can be used to sample molecular variability in natural populations, diagnose genetic defects, and genotype rapidly a large number of SNP markers.) It is desirable, yet challenging, to develop a systematic approach to integrate these different types of data.

A Broader Array of Bioinformatics Problems

Microarray analysis provides for statisticians an excellent entry point to bioinformatics/computational biology. Here I give a brief personal account on other bioinformatics challenges that await statisticians’ contributions.

The protein folding problem, i.e., the prediction of the three-dimensional fold of a protein molecule based only on its primary sequence information, is often regarded as the crown jewel of the biopolymer research. Knowledge on the structures of target proteins and on how

(continued on page 66)
they interact with ligands is of paramount importance to drug designers. Although the 3-D structures of many proteins have been worked out by X-ray crystallographers, these structures only account for a small part of the protein universe and scientists are still not capable of predicting protein tertiary structures ab initio. Recently, theoreticians have turned their attentions to much simpler black-white bead model for understanding the design principles of protein structures (Dill et al. 1995, Zhang and Liu 2002). Practitioners have opted to use more statistically based threading method (Xu et al. 2002). This method "threads" the given protein sequence into a set of known structural templates and finds the most suitable sequence-template fit. Many structural templates are constructed by combining the known protein structures with statistical model-based protein sequence analysis.

Multiple sequence alignment is still the main tool for protein sequence analysis, which has been at the center of computational biology for about 30 years. With the completion of the human genome and genomes of many other species, the task of organizing and understanding the generated sequence and structural data becomes even more pressing and challenging. Many statistical and computational methods for sequence alignment have been proposed over the years, among which the most popular ones include Clustal W. (Thompson et al. 1994), PSI-BLAST (Altschul et al. 1997), SAM (www.cse.ucsc.edu/research/compbio/sam.html), and HMMER (http://hmmer.wustl.edu), etc. In particular, the application of hidden Markov models (Baldi, et al. 1994; Krogh et al. 1994; Durbin et al. 1996) and the Gibbs sampler (Lawrence et al. 1993, Neuwald et al. 1995, 1997) to biopolymer sequence analysis has revolutionized the field. Pfam database (Bateman et al. 2002) contains a large collection of annotated protein family profiles built based on hidden Markov models and is becoming very influential in protein research. An emerging challenge is the analysis of aligned protein sequences in order to gain further insights on protein functions (Neuwald et al. 2002).

There have been some recent interests in incorporating gene ontology (GO) in microarray analyses. Gene ontology refers to a dynamically controlled vocabulary that can be applied to (the genes of) all organisms. Each gene product can be described by its molecular function (e.g., transcription factor), its involvement in biological processes (e.g., mitosis), and its cellular location (e.g., nucleus). Bringing GO into the analysis of high throughput biological data such as microarrays can be extremely insightful. Recently, in the analysis of circadian gene regulation, Storch et al. (2002) mapped various clusters of genes based on their microarray experiments to GO hierarchies and found that clock-regulated genes in heart and liver participate in many related processes even though the two sets of genes have almost no overlap.
TFs identify the genes they are needed to regulate by recognizing via energetic interactions specific binding sites, often located upstream of the genes. It has been realized early on that these sites are often conserved. For example, the binding sites of STE12 of yeast look like "TGAACA." If the genome were indeed a "novel," then these patterns are like key words (with typos) in the novel. Thus, techniques for discovering new "words" in a text have been developed (Bussemaeker et al. 2000, Liu et al. 1995, Liu et al. 2002b) and applied to discover the TF motif sites and pattern.

Statistics are likely to play an important role in other bioinformatics problems, including evolutionary analysis and the analyses of SNPs in the human genome, both of which could help us understand our origins and shed light on protein functions and cellular processes. The SNPs have recently attracted much attention from scientists because of the SNPs' great potential in mapping genes responsible for complex diseases and the availability throughout SNP detection and analysis tools. Statistical modeling and computation are crucial to these developments (Daly et al. 2001, Niu et al. 2002).

Concluding remarks

Many have said that this century is the century for biology. No matter what this means to each of us, we can clearly see that many biological data have been generated and many biological facts are known, yet general principles are still lacking. As a statistician with an interest in biology, I feel blessed because I can now taste the great biological fruits (i.e., analyzing their data) without having to sweat to "grow" them by myself! I feel that our field is also blessed by the high-throughput biological data generation—statistics has never been in so much demand from biologists. But it is also a challenge to all statisticians. Indeed, if we statisticians do not proactively participate in the biotechnology revolution, other scientists (e.g., computer scientists) will learn and do statistics whether we approve it or not. We clearly have an advantage, for now, and we still can control our own fate if we try. ■

References continued


