**Study of Gene Expression: Statistics, Biology, and Microarrays**

**Ker-chau Li, Statistics Department, UCLA**

It is a great honor to share my enthusiasm about this fascinating subject with you. My talk is about a leading role that statisticians can play in the rising area of computational biology. As the title suggests, it has three parts. Let me begin with the Biology part. In particular, I will give a quick overview on the three basic macromolecules in the cellular biology: DNA, RNA, and protein. But as we know, statistics has many applications. So you may ask why it is so special about biology. I hope the reasons will get clearer as we move on. For now, let's look at a broader spectrum by going over a few slides (courtesy of U.S. department of Energy) about Human Genome Project. The human genome project has received a lot of media coverage, which is well-deserved because of its profound impact on many disciplines, including the research on global carbon cycles, industrial resources, agriculture, medicine and health, and eventually our daily life.

**1. Human genome project.** Begun in 1990, the U.S. Human Genome Project is a 13-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 effective resource and technological advances have accelerated the expected completion date to 2003. Project goals are to identify all the approximate 30,000 genes in human DNA, to determine the sequences of the 3 billion chemical base pairs that make up human DNA, to store this information in databases, to improve tools for data analysis, to transfer related technologies to the private sector, and to address the ethical, legal, and social issues that may arise from the project. In June 2000, a working draft of the entire human genome was completed and two papers were published in February 2001 separately by an international team ( Nature, Feb, 2001) and the private firm Celera Genomics ( Science, Feb, 2001).

In spite of this milestone achievement, the future challenge is even greater. Here is a list of some unknown questions:
• The exact gene number, exact locations, and functions
• Gene regulation
• DNA sequence organization
• Chromosomal structure and organization
• Noncoding DNA types, amount, distribution, information content, and functions
• Coordination of gene expression, protein synthesis, and post-translational events
• Interaction of proteins in complex molecular machines
• Predicted vs experimentally determined gene function
• Evolutionary conservation among organisms
• Protein conservation (structure and function)
• Proteomes (total protein content and function) in organisms
• Correlation of SNPs (single-base DNA variations among individuals) with health and disease
• Disease-susceptibility prediction based on gene sequence variation
• Genes involved in complex traits and multigene diseases

• Complex systems biology including microbial consortia useful for environmental restoration
• Developmental genetics, genomics

The anticipated economic benefits of human genome project are manifold. Some of them are listed here.
In molecular Medicine :
• improved diagnosis of disease
• earlier detection of genetic predispositions to disease
• rational drug design
• gene therapy and control systems for drugs
• pharmacogenomics "custom drugs" ;
In microbial Genomics :
• rapid detection and treatment of pathogens (disease-causing microbes) in medicine
• new energy sources (biofuels)
• environmental monitoring to detect pollutants
• protection from biological and chemical warfare
• safe, efficient toxic waste cleanup
In agriculture, livestock Breeding, and bioprocessing:
• disease-, insect-, and drought-resistant crops
• healthier, more productive, disease-resistant farm animals
• more nutritious produce
• biopesticides
• edible vaccines incorporated into food products
• new environmental cleanup uses for plants like tobacco
In this connection, I would like to mention the remarkable achievement of the completion of two drafts of rice genome, independently by a publicly funded group led Beijing Genomics Institute, and by the private firm Syngenta. (Science, 2002, 296, April).

**2. Central Dogma**. DNA, a long string of four different small molecules (A,T,G,C, nucleotides), contains the genetic information about all kinds of cellular activities. It is present in all cells. Actually there are two types of cells in the biological world. The eukaryotic cells have several compartments and one of them is the nucleus in which DNA is stored. The prokaryotic cells lack a nucleus and have no internal compartments. Although there are substantial differences between eukaryotic and prokaryotic cells, the central dogma of biology applies to both types of cells: DNA−>RNA ->protein.

• Nucleotide. A small molecule forming the base of DNA and RNA. There are four types, represented by A,G, C,T(or U).
• DNA. Long linear polymer (sequence) of deoxyribose nucleotides.
• Chromosome. A structural unit of genetic material consisting of either a single, circular double-strained DNA molecule (for prokaryotes) or a single, linear double-stranded DNA molecule and associated proteins.
• Gene. A short segment of (single-stranded) DNA from a chromosome that is necessary for the synthesis of a functional protein or RNA molecule.

• mRNA.  A copy of a gene the cell uses for protein synthesis. In eukaryotes, mRNA has to be transported from nucleus to cytoplasm where ribosomes are located.
• Transcription. The process of copying a gene into mRNA.
• Translation. The process of making a protein by ribosomes, using the information encoded in an mRNA.
• Protein.   A linear polymer composed of a twenty-letter amino acid alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}.  Each amino acid is encoded by three nucleotides in a row.

Proteins  often form  a  multi-unit complex as a functional identity. For example, mitochondrial  ATP synthase is  a multimeric protein complex bounded to inner mitochondrial membranes that catalyzes synthesis of ATP (the molecule required to fuel various energy-consuming cellular processes ) coupled to proton movement down the electrochemical gradient.  It is also called the $F_0F_1$ complex. Other complexes include proteasome, ribosome (part of it being r-RNA),  DNA and RNA polymerase, and so on.

**3. Gene expression and  Microarray**.  There are many perplexing questions surrounding the central dogma of biology. If  all cells in our body contain identical DNA, then why the liver is  different from the kidney? Why  eyes can see but not smell?
The key to answering such questions is the differential expression of genes in different tissues under different signal stimulation.  Although the encoded genetic program is the same for each cell, the varieties  and the amounts of  proteins that characterize a particular cell type are determined by the concentration of each protein's corresponding mRNA, the rate of translation, and the degradation of protein. The concentration of various mRNAs is determined by the rate of transcription which varies in cell types and in response to signals received from the environment.

The process of transcription is complex. Although the fundamental molecular mechanism has been under intensive study, much of the process is still unknown, especially in eukaryotes. The central question of how to predict what genes are turned on/off under what conditions remains elusive.  One major underlying hurdle is that  all cellular processes are interlocked. For example, the transcription factors, proteins that participate in transcription, are themselves subject to controls such as nucleus/cytoplasm localization, protein modification, and of course, transcription and translation. The interference of histone acetylation further complicates the matter. To resolve any kind of chicken-and-egg type of looping dilemma,  inevitably  we need to have a genome-wide assessment about the amount of each variety of protein under any given condition.
But protein quantification, two-D gels for example, is hard to conduct at the full  genome scale.  As an alternative, researchers are now turning to  mRNA, using the microarray technology.

DNA microarray is popularized by Pat Brown's Lab at Stanford.  Thousands of genes/open-reading frames are printed onto a glass slide first, using a robotic printing device.  Then the mRNAs   extracted from the cells under study are used to prepare cDNA  by reverse transcription in presence of Cy3(green) or Cy5(red) labeled deoxyuridine triphosphate(dUTP). Finally  by hybridization with the DNA on the glass

slide, the fluorescently-labeled cDNA at each spot can be quantitated with a laser scanner. There are several competing microarray techniques available. For example, Affymetrix uses selected oligonucleotides instead of the entire open reading frame.

## 4. Analysis of microarray data.

Microarray data are extremely noisy. There are three levels of analysis. The lowest level is to convert an image feature into a number that reflects the amount of mRNA. This image processing procedure is already a non-trivial step and is doomed to be error-prone. The major source of errors may come from the chemical processing of DNA chips and mRNAs. Signals are often unevenly distributed inside the printing spot, making background noises hard to be separated from the signals.

Although the ultimate goal of array quantitation is to measure the exact number of copies of mRNA present in the cell samples, it has not been realized with the current technology yet. Recognizing that differential expression is often the aim of the study, Pat Brown's lab uses the clever idea of two color scheme for obtaining a relative measure about the abundance of mRNA in the cells of interest (colored by Cy5) as compared to a control sample(colored by Cy3). How to convert these numbers into a meaningful measure of expression level becomes the focus of the second level analysis. Many studies have shown that it is not enough to take a simple ratio of CY5 over CY3 reading without chip to chip adjustment. Several statistical models are proposed in exploring the error patterns. But which one is most appropriate? While this will continue to remain debatable, it forces biologists to realize the sophistication involved in interpreting the seemingly straightforward measurement of microarrays.

After the expression level of each gene is quantitated, we can represent the result by a matrix of dimension n by p, where n is the number of genes and p is the number of conditions or arrays. Depending on the scope of study, for most experiments, n is in the order of several thousands or more, and p can vary from less than 10 to a couple of hundreds. How to extract biological information from this matrix of expression profiles constitutes the third level of analysis. This is the part that I will describe next using some examples.

**4.1 ALL versus AML**. E.S. Lander's group at MIT have published a report in Science (1999, vol 286, Oct 15, page 531-536)) demonstrating the feasibility of classifying cancers solely based on gene expression monitoring. Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery). The group chose acute leukemias as a test case. Classification of acute leukemias began with the observation of variability in clinical outcome and subtle differences in nuclear morphology (*4*). Enzyme-based histochemical analyses were introduced in the 1960s, providing the first basis for classification of acute leukemias into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL) or from myeloid precursors (acute myeloid leukemia, AML). In the 1970s, antibodies recognizing either lymphoid or myeloid cell surface molecules are developed.

Although the distinction between AML and ALL has been well established, the diagnosis procedure is complex. Common clinical practice involves an hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis. Genome-wide expression comparison between the ALL and AML patients promises a simple one step diagnostic procedure which is more cost-effective. The data contain 38 bone marrow samples (27 ALL, 11 AML) taken from acute leukemia patients at the time of diagnosis. They used Affymetrix oligonucleotide chips with 6817 human genes printed. From the expression data, they chose 50 genes as the class predictors and constructed a voting scheme for combining the results of prediction by each of these genes. They then apply the derived prediction rule to 34 independent new leukemia samples. Out of them, 29 are predicted with100 percent accuracy. The remaining 5 samples are deemed as weak prediction, meaning that there are more disagreements among the 50 predictor genes.

My comment on this otherwise first-rate scientific work is that there is the lack of optimality consideration in their construction of the classification rule. Most disturbing is the way they find the 50 genes. For each gene, they uses the ratio of the difference between sample mean of ALL and the sample mean of AML over the sum of the sample standard deviation of ALL and the sample standard deviation of AML. They further interpreted this measure as something similar to the Pearson's correlation between the gene expression and the class distinction. It seems to me that what they suggest is at odds with the basic statistical argument based on sample size and variance of the mean. In fact, this is the first time I saw an "application" of adding two standard deviations up. Does the analysis of microarray data need such new subjective invention of data analysis? It is your call.

### 4.2 Cell cycle regulated genes.
Different eukaryotic cells grow and divide at quite different rates and some of them do not even divide. Yeast, for examples, many divide every 120 min under suitable conditions; the first divisions of fertilized eggs in embryonic cells of sea urchins and insects take only 15-30 min; most growing plant and animal cells take 10-20h to double in number; nerve cells do not divide at all; fibroblasts are quiescent if there is no demand of growth such as assisting in healing wounds. Regulation of cell division is critical for the normal development of multicellular organisms and the lack of it ultimately leads to cancer. The budding yeast S. Cerevisiae and the fission yeast Schizosacchoromyces probe are especially useful models for the study of eukaryotic cell cycle.

A cell cycle can be divided into four phases, S, G2, M, G1, although the exact cutoff is often not a clean one. In S. Cerevisiae, bud emergence signifies the ending of G1 phase and the beginning of S phase. The major event in the S (synthesis) phase is chromosome replication. The bud gets bigger as the cell continues to grow (G2 phase) and a portion of nucleus gradually migrates into the daughter cell. Then at the M (mitosis) phase, a series of events happen: spindle formation, chromosome segregation, nuclear division,, and eventually cytokinesis. The daughter cell is separated from the mother cell. The daughter cell is smaller than its mother must grow in size considerably before it attempts to divide. Both mother and daughter cells remain in the G1 phase while growing, although it takes

mother cells a shorter time to reach a size compatible with cell division. The point of no return in the late G1 phase when the cell becomes irrevocably committed to entering the S phase and traversing the rest of the cycle is called START. On the other hand, if the environment is not right, yeast cells may stop growing and enter into the stationary phase, Go, and survive for a very long period of time.

Over years, the extensive study on the molecular basis for each major event in the cell cycle has led to the identification of the cyclic expression patterns for over 100 genes. The discovery is largely made on a one by one basis. Now with microarrays, all genes can be inspected together. A typical experiment will produce over 6000 expression curves, one for each gene. Four such datasets are available at the web site of Stanford Cell Cycle project. One major huddle for such experiments is how to synchronize the metabolic clocks for millions of cells to make sure the harvested cells are at the same point of cell cycle. The four datasets are obtained using four different methods of synchronization.

The analysis of such large data sets is by itself a very interesting problem for statisticians. The initial approach by Spellman et al used Fourier series for finding cell-cycle regulated genes. They consider cell cycle to be a periodic phenomenon, so Fourier analysis appears to be the right tool to use. For practical reasons, the number of cycles is small, typically one, two or at most three. Thus in essence, their methods amount to approximate a curve with a sine or cosine function. But the perfect geometric symmetry of sine, cosine curves contradicts the biological reality that the four phases are not of equal duration. In view of this, Li, Yuan and Yang (2002) proposed a different approach which allows the data to speak for themselves in choosing more appropriate basis functions. Their approach is a combination of principal component analysis with nested regression modeling. For the cdc15 experiment, while the re-analysis has a good agreement with the original analysis, it also reveals several non-cyclic curve patterns from the list of Spellman et al's 800 cell-cycle regulated genes. More interestingly, the first principal component shows an oscillation pattern going up and down alternatively throughout the curve. A biochemical explanation of this finding is still lacking.

**4.3. High dimensional data analysis tools.**
The rich information hidden underneath large scale microarray datasets can only be unearthed by systematic computational methods. Many popular multivariate data analytic tools has been applied. This includes the hierarchical clustering with centroid linkage (mistakenly quoted as average linkage) (Esier et al ), K-means, self-organization map, Bayesian clustering, PLAID model ( Statistica Sinica 2002, Lazzeroni, Owen), and Generalized association plot (Chen 2002). Clustering is also referred to unsupervised learning in the engineering literature. The supervised learning, known as classification or discriminant analysis in Statistics, is also loaded with various tools such as CART, support vector machine, nearest neighbor, density estimation, and Fisher's linear discriminant analysis. Both areas are still very active and I would expect more innovative ideas to surface in the future.

**4.4. Profile similarity and co-expression dynamics**..

Profile similarity is perhaps the most fundamental notion behind all microarray elucidation methods. Two genes with similar expression profiles are likely to participate in a common structural complex, metabolic pathway, or biological process. Pearson correlation coefficient has become a popular measure of similarity measure.  Let's look at one example.  This is a scatterplot matrix showing the co-expression patterns of seven genes MCM1-MCM7 from the yeast cell-cycle data. The positive association among the 6 genes, MCM2,..,MCM7, is a sharp contrast to that between each one of them and MCM1. It turns out that MCM2,…,MCM7,  form a hexameric complex that binds chromatin. It is a part of the pre-replicative complex forming at the origins of DNA replication between late M phase and G1/S transition. On the other hand, MCM1 is a transcription factor of the MADS box family. Thus it has a entirely different function from the other MCM genes.

Despite of the successful application of this notion in many ways, a major restriction is clear when one realizes that most genes have multiple cellular roles to play. Those genes engaging in one process may later dissociate and undertake activities of their own as the cellular conditions change.  A current project of mine is to study such co-expression dynamics.  I have applied Stein's lemma to come up with a simple statistics for quantifying the average amount of co-expression change between a pair of genes as the expression level of a third gene varies.  The results are quite encouraging and several interesting examples are obtained. In one example, we show how the entire urea cycle pathway is rationally expressed by yeast cells in order to achieve an optimal management of arginine biosynthesis and degrading.  Currently we are applying this new method to several large scale gene expression data on cell lines and various types of cancers.

**5.A concluding note**. To conclude this talk, let me go back to the title. I  have purposely placed Statistics in front of biology to emphasize the leading role  that statisticians can play in this wide-open area. Reflecting this consensus,  the bioinformatics programs launched by many institutes in US including UCLA have Statistics courses as a major part in the curriculum. In my view, this new discipline amounts to the integration  of three components: data, knowledge, and searching. The greatest challenge is how to formulate a meaningful biological problem that can be solved with available DNA, protein sequence  and  Microrray data, using the mounting  biological knowledge accessible through creative searching via the internet.  The statistical training on how to handle errors and model stochastic phenomena is a big advantage  in this complex field. Finally,  for those interested, the January issue of Statistica Sinica gathered 17 articles which may be useful for jumping into this area.