BEYOND BIOLOGY: A STATISTICIANS' PERSPECTIVE

Ravindra Khattree

Human genetics and statistics are like twin sisters who have grown together and shared many similar growing pains over the past hundred and twenty-five years. Important ideas in statistics such as regression analysis, analysis of variance and correlation were formalized in statistics mainly out of their necessities in genetics (Ewens, 1999). Since then the statistical ideas have been central to the development of quantitative genetics and conversely, genetics has fed statistics with many important research problems, solutions to which have gone on to make impact in many other disciplines. Later, as these two fields matured, the exchange of ideas between the fields became more intense and contributed to the mutual benefit and enrichment of each of them. It must be noted that the father of modern statistics, R. A. Fisher, was a geneticist by profession, and the noted geneticist, J. B. S. Haldane, chose to work at Indian Statistical Institute doing genetic experiments and applying statistical techniques to analyze them. He did this during much of the later part of his life, because he felt that the statistical techniques were central to the development of modern genetics. The direction seems to have reversed now, when noted statisticians and probabilists, like Sam Karlin, have essentially transformed thinking and approaches in genetics by their fundamental biostatistical research.

One of the latest examples of the love between the two fields is the human genome project. The human genome project has not only revolutionalized the biological and medical sciences, but also strongly impacted other scientific disciplines such as biomedical engineering, mathematics and, of course, statistics. At a time when statistics was beginning to be thought as a well developed field like mathematics and physics, where fundamental principles are well established, it came as a shock to statisticians when they realized that many of its fundamental ideas and tools were not adequate to solve the important problems arising out of human genome project. This essentially provided an impetus to develop many newer, fundamental statistical methods and to revisit as well as reexamine the basic philosophical and methodological foundations of statistical science.

Many standard statistical approaches proved incapable of handling the challenge created by the sheer size and wealth of data. This problem has been the statistician's delight as well as his nightmare. The genome data often has hidden structures, which can cause conventional methods to become infeasible. New tools were required to deal with this challenge. Specifically, the following few research questions few being addressed by statisticians.

1. **How do we assess the validity of our measurements and ensure the accuracy of genome data?** We must learn how to clean such datasets, which potentially have many freak or inconsistent observations and a possible large amount of bias. These data cleaning problems are compounded by the fact that the related bioengineering technology needed for collecting the genome data and for making such measurements has developed at a much faster pace than the statistical knowledge components needed to make full sense of these data. A case in point is the use of micro-array technology in biological experiments to study the ex-

pressions of genes and to identify the genes whose expression may be causing a particular disease.

For example, in many micro-array experiments, where thousands of genes are examined simultaneously, the results can be different when the experiments are repeated under identical conditions. Reproducibility requirements, therefore, can be violated. We need to know whether such irreproducibility stems from bad data items or due to other technology related causes. Thus, data cleaning is often a necessary task and statistical techniques, in combination with biological expertise are needed to do this.

2. **What information if any, can be mined from such a large data sets.** The interdisciplinary field of "data mining" deals with careful and systematic computer based search for the information hidden in the large dataset, using the multivariate statistical tools. For example, one would like to identify a few important genes, which may be responsible for a particular disease, among a collection of a few thousands. Studying one gene at a time is likely to miss much of the valuable information because some of the genetic markers or segments may be linked to each other and interactions may exist. Thus, any approach of statistical data mining must be capable to uncover in the hidden linkages, the complex genetic structures however small they may be within this large amount of data. Fortunately, there are many well established multivariate techniques for this purpose and important discoveries have been made with their help.

3. **We also observe that certain problems in genome research need tools, which are unique to the problem at hand.** This has resulted in important scientific research through the combination of biology, statistics, probability, mathematics and operations research. As an example, a biologist may perform 50 or 60 micro-array experiments, where thousands of genes are

being studied simultaneously. The sample size in this case, is, say 50 but the numbers of variables may be 2000. In conventional statistics, we must have the sample size larger than the number of variables, but in the present context the two are reversed. Researchers are attempting to solve such problems using the new advanced statistical techniques but a satisfactory solution is yet to be obtained.

4. **How should one design the genome experiments.** One needs to make fair and unbiased comparisons, extract the maximum possible amount of information and minimize the number of experiments. This is a very difficult statistical problem.

5. **How do we approach the problem of multiple testing?** Any statistical decision making process is prone to two types of errors, namely false positives (e.g., a person who does not have AIDS is erroneously found to test positive for AIDs) and false negatives (e.g. a person who has AIDS is tested negative). One attempts to fix the first kind of these errors (which we call the type I error) at a desired minimum level (say, 1 percent) and minimize the second kind. However, when thousands of statistical tests are performed on these thousands of genes, the first type of error may be much higher than the desired level. The main emphasis is to come up with analyses of genomic data, which minimize the false-discovery rate that is, the rate of erroneously labeling genes as important when they were really not important in the particular context. Use of statistical techniques proves to be an enormous help to biologists to narrow down the search and reduce the further probing to these few important genes from the large list of few thousands.

In closing, Genomics and genome data have some features that have resulted in unique statistical approaches, calls for more interdisciplinary collaboration and have impacted

greatly the statistical domain. I list some of these here as a summary (also see, Liu, 1997).

a. Some genome data are mixtures of discrete and continuous variables, such as combination of genotypes of genetic markers, which are discrete and values of quantitative traits, which are continuous.

b. Many test statistics for genomic hypotheses do not have nice statistical probability distributions.

c. Genomic data sets are unusually large and computationally intensive.

d. Standard textbook approaches familiar to biologists such as those based on regression and analysis of variance are inadequate.

e. Genomic data usually have a large number of variables (genes) and small number of samples (experiments), which is not a very convenient statistical scenario.

f. Genomic data have hidden probabilistic structures embedded on them which are hard to capture due to sheer volume.

g. Einstein once said, "all that can be counted does not necessarily count, and all that counts, cannot be fully counted." This is especially true for genome data. Since the data sets are so large in size and hence often cannot be used directly, the amount of data must be reduced. This data reduction essentially amounts to identifying key features or variables in data and by ensuring that ignoring the other variables does not result in significant loss of valuable information. However, various interdependences and cause and effect relations make such a data reduction very difficult especially when the nature of interdependence has not been fully understood. This has regenerated interest in many classical multivariate statistical techniques such as principal components analysis, factor analysis and clustering. New methods, with special reference to

genome data are being devised and their appropriateness is being examined.

I personally believe that while these unique features are puzzling and at times frustrating to biologists and statisticians alike, the work is clearly for the betterment of the two professions and will define a theme for the important research in this new century.

REFERENCES

Ewans, W.J. (1999). Statistical methods in human genetics, (In *Statistics in Genetics*, Ed., M.E. Holloran, S. Geisser). 147–162, Springer, New York, N.Y.

Liu, B.H. (1997). *Statistical Genomics: Linkage, Mapping and QTL Analysis*, CRC Press, Boca Raton, FL.