AND FINALLY . . . THE DATA FLOOD

One of the major themes brought out by the workshop was the interplay between theory and data, but the discussions above have not mentioned how much data must be dealt with. The magnitude of the data sets actually creates their own statistical problems (just due to their size).

As an example, Dan Roden of Vanderbilt University reported on research the original goal of which was to use genetics to predict individual responses to drugs. However, the research quickly evolved into a challenge of navigating through a massive data set. Pharmacologists are very interested in understanding why individuals have different responses to the same drugs, and how to predict those variations. The variability in drug response can correlate with a variety of factors, such as gender, age, disease types, concomitant drug therapies, and ethnicity.

Variability in drug response among different individuals may also be due to genetic factors. Each person has two strands of DNA in their genome, shown as two panels in Figure 19. At particular genome locations, the DNA sequences might differ between any two people. Such a difference, called a DNA polymorphism, might be associated with the occurrence of side effects in a given individual.

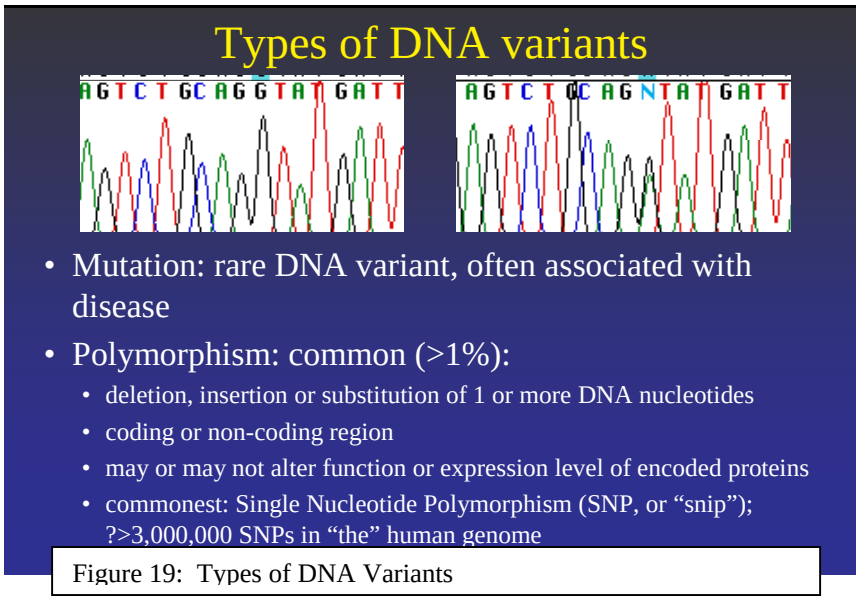


Figure 19: Types of DNA Variants

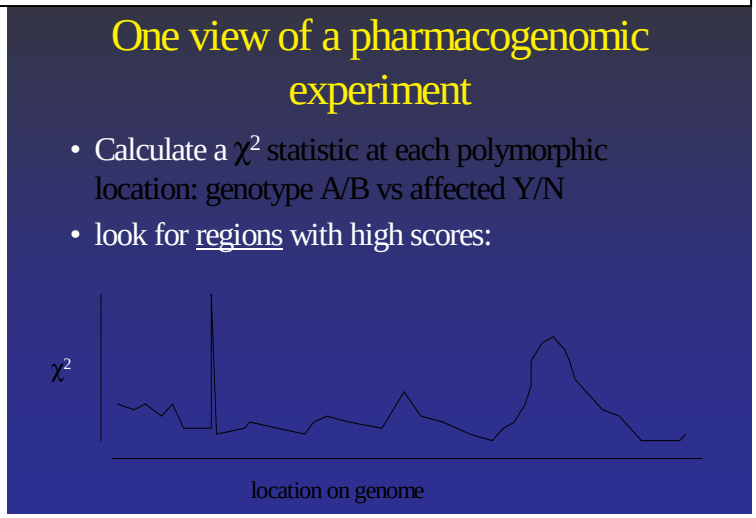
Mutation is one of the factors causing DNA polymorphisms, and which therefore contributes to disease onset. DNA polymorphisms may be due to the deletion, insertion, or substitution of a nucleotide, may occur at coding or non-coding regions of the DNA, and may or may not alter gene function. The occurrence of DNA polymorphism makes it possible to associate a person's response to drugs with particular DNA regions, for example, by correlating the occurrence of the polymorphism with the response. This is the basis of current phamacogenetics, which is the study of the impact of individual genetic variants on drug response.

Roden's research sought to evaluate the role of genetics in determining drug response in the case of a single nucleotide polymorphism (SNP) that is known to predispose

individuals to drug-induced arrhythmias. He approached the problem with the following strategy:

- Define the drug response (phenotype) of interest
- Accumulate patients/DNA/families
- Identify candidate genes that might explain significant response variations
- Identify polymorphisms in candidate genes
- Relate the identified polymorphism to the phenotype

Figure 20: Data from a Hypothetical Pharmacogenomic Experiment



Such an analysis would produce a graph like that in Figure 20, where the χ^2 statistic would be calculated at each SNP. However, such an analysis is actually infeasible for both statistical and economic reasons, because of the "data flood." . Suppose the research has considered 100,000 SNPs in 1,000 patients (500 affected, 500 not affected). The statistical problem is that the data will result in 100,000 χ^2 statistics. With such a multiplicity of tests, there will be many false positives. How then does one set a sensible cutoff point for statistical significance?

Even if the statistical problem can be solved, basic economics makes this straightforward experiment infeasible because of the tremendous cost of recording 100,000 genotypes in each of a thousand people. (If the cost of determining a genotype were only 50 cents, the entire experiment would still require $50 million.)

Therefore, there is a pressing need to solve the problem of handling the bioinformatics data flood.

The data flood pointed out by Roden is only one example of the data handling challenges to be overcome. With the development of microarray experiments, the amount of data available today is enormous. At the April 2001 workshop, Terry Speed of the University of California at Berkeley gave an overview of microarray experiments, which provide a means of measuring expression levels of many genes in parallel.

In the so-called Stanford protocol, genetic material from cells is apportioned into two samples, each of which is exposed to a different treatment. (One of the treatments might be the null treatment, in which case we are comparing a treated sample to a control.) The goal is to determine how the two samples differ in the way their genes are "expressed"—that is, how the genes cause proteins to be created in accordance with their embedded genetic information. One sample is labeled with a red dye and the other with a green dye. The two samples are distributed over a microarray slide (a "gene chip"), which typically has 5000-6000 different segments of complementary DNA (cDNA) arrayed on it. The two samples of red- and green-dye-tagged genetic material adhere to the slide in different patterns according to their chemical bonding to the cDNA. When the dyed genetic material is allowed to express proteins, the level of activity at each coordinate of the gene chip can be measured through the fluorescence of the dyes. From these measurements, one can develop understanding of how the genetic material was affected by the treatment to which it was exposed. More complete background on this process, and a number of valuable other links, may be found athttp://www.stat.Berkeley.EDU/users/terry/zarray/Html/index.html.

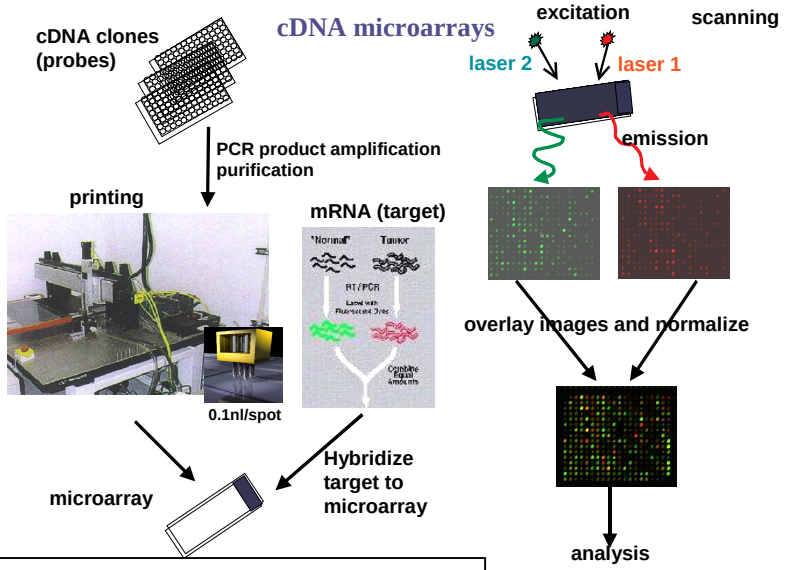


Figure 21: Microarray construction

Many statistical issues arise in the analysis of microarray data, including issues of experimental design, data pre-processing, and arriving at ultimate conclusions. For example, the typical range of expression (on a $\log_2$ scale) is about +- 5, and the amount of background noise in the data could be substantial. Thus, at present, it is usually possible to identify (with some certainty) only those genes that express at a very high or very low level.

Although there are problems with expression levels, and also with bias, a plot of M vs. A, where

$$M = log_2 \text{ (red expression)} - log_2 \text{ (green expression)}$$
$$A = log_2 \text{ (red expression)} + log_2 \text{ (green expression)}$$

can be extremely useful, as in the following experiment described by Speed that identified genes with altered expression between two physiological zones ("zone 1" and

"zone 4") of the olfactory epithelium in mice. Figure 22 shows the log ratios plotted against the average of the logs (which gives a measure of absolute expression). It illustrates the noise level in much of the data. It also shows that a number of genes have very high expression levels, and that these genes show differential expression.

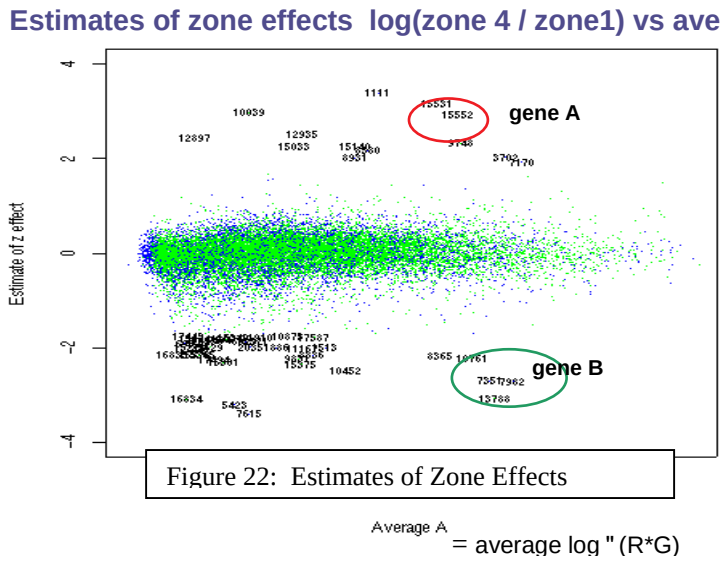


Figure 22: Estimates of Zone Effects

can be extremely useful, as in the following experiment described by Speed that identified genes with altered expression between two physiological zones ("zone 1" and "zone 4") of the olfactory epithelium in mice. Figure 19 shows the $\log_2$ ratios plotted against the average of the logs (which gives a measure of absolute expression). It illustrates the noise level in much of the data. It also shows that a number of genes have very high expression levels, and that these genes show differential expression.

Summarizing, Speed outlined some challenges to current research, such as:

- How to deal with the observed bias associated with whether a sample is treated with red or green dye (which suggest the need to run the complementary experiment of interchanging the red and green labels);
- How to create better designs for microarray experiments, ones that extend beyond merely comparing treatment versus control;
- How to carry out the experiments' pre-processing so as to reduce the noise in the data; and
- How to deal with the fact that, because of the large number of genes tested in microarray experiments, the large number of statistical tests carried out in parallel greatly increases the chance of finding a false positive. (See, for example Tusher, *et al.* (2001), which uses the false discovery rate to set cutoff points for these errors.)

SUMMARY

Throughout the workshop there was much lively discussion from the participants, and there were also two prepared discussants who greatly helped in identifying major themes and challenges for the future. Before summarizing the thoughts of discussants Jim Keener and Keith Worsley, and other comments, we first note that there has been a big cultural change in mathematics and statistics in the past few years. In the past, development of theory would precede the collection of data. Now, in many areas, data drives the development of theory. This is especially true with mathematical sciences research related to the biomedical sciences.

The mathematical sciences have benefited areas of biomedical research by

- Suggesting insights that could not be observed directly (such as "viewing" the interior of the beating heart via a simulation);
- Classifying and describing generic features and processes of biomedical systems); and
- Suggesting how some biomedical systems work and their limitations (through tools such as dynamical analysis of mathematical models that emulate cell signaling networks).

The major challenge to be overcome before the interface between the mathematical and biomedical sciences reaches its potential is to ensure that more mathematical scientists are exposed in depth to research in the biomedical sciences and given the means to contribute. An important step in that direction is that mathematical formulations of important biomedical research be more widely available.

References

Asthagiri, A.R., C.A. Reinhart, A.F. Horwitz, and D.A. Lauffenburger. (2000). The role of transient ERK2 signals in fibronectin- and insulin-mediated DNA synthesis, *J. Cell Sci*. 113: 4499-4510 (2000).

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge MA : MIT Press.

Donahue, J. K., Heldman, A. W., Fraser, H., McDonald, A. D., Miller, J. M., Rade, J. J., Eschenhagen, T. and Marbán, E. (2000). Focal modification of electrical conduction in the heart by viral gene transfer. *Nature Medicine* **6** 1395 - 1398.

Dowling, H. F. (1977). *Fighting Infection*. Harvard Press

Duggan, D.J., M. Bittner, Y. Chen, P. Meltzer, and J.M. Trent. (1999). Expression profiling using cDNA microarrays, *Nature Genetics* **21**:10-14.

Garfinkel, A., O. Voroshilovsky, Y-H. Kim, M. Yashima, T-J. Wu, R. Doshi, H. S. Karagueuzian, J.N. Weiss, and P-S. Chen. (2000). Prevention of ventricular fibrillation by flattening cardiac restitution, *Proc. Nat. Acad. Sci.* 97:6061-6066, 2000

Grakoui, A., Bromley,S. K., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M. and Dustin, M. L. (1999). The Immunological Synapse: A Molecular Machine Controlling T-Cell Activation *Science* **285** 221-227.

Hanahan, D. and Weinberg, R.A. (2000). The Hallmarks of Cancer *Cell* **100**: 57.

Holt, GR, Softky, GW, Koch, C & Douglas, RJ (1996). *Journal of Neurophysiology* **75**:1806-1814

Johnson, D. H., C.M. Gruner, and R.M. Glantz. (2000). Quantifying information transfer in spike generation, *Neurocomputing*, **33** 1047-1054

Kohn, K. (1999). Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems. *Mol. Biol. Cell* **10** 2703-2734.

Lipsitch, M., Bergstrom, C. T. and Levin, B. R. (2000). The epidemiology of antibiotic resistance in hospitals: Paradoxes and prescriptions. *Proc. Nat. Acad. Sci*. **97**: 1938-1943.

McKeithan (1995). Kinetic Proofreading in T-cell Receptor Signal Transduction. *Proc. Nat. Acad. Sci*. **92** 5042-5046.

Nagy, L. (1998). Changing patterns of gene regulation in the evolution of arthropod morphology. *American Zoologist* **38** 818-828.

Sager, B. and Kaiser, D. (1993). Two cell-density domains within the Myxococcus xanthus fruiting body. *Proc. Nat. Acad. Sci*. **90** 3690-3694.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci*. **98** 5116-5121.

von Dassow , G, Meir, E. Munro, E. M. and Odell, G. (2000). The segment polarity network is a robust developmental module. *Nature* **406** 188 - 192

Appendix

Workshop Program

Dynamical Modeling of Complex Biomedical Systems

Holiday Inn-Georgetown
2101 Wisconsin Ave., Washington, D.C.
April 26-28, 2001

Thursday, April 26 (p.m.)
(Kaleidoscope Room)
12:00 Lunch available at workshop site

Overview Session---Part 1

1:00-1:30 Eduardo Marban (John Hopkins Univ.), “Gene Transfer/Gene Therapy”

1:30-2:00 Terry Speed (Berkeley), “Statistics and Microarray Data”

2:00-3:00 Open discussion
De Witt Sumners (Florida State), moderator

3:00-3:30 Break

Mathematical Sciences and Disease States---Part 1

3:30-4:00 Charles Peskin (NYU) "A Virtual Heart for Pathophysiology and Prosthesis Design"

4:00-4:30 James Weiss (UCLA), "Biological Pattern Formation: From Arrhythmias to Embryos"

4:30-5:30 Open discussion
De Witt Sumners (Florida State), moderator
James Keener (Univ. of Utah), summarizer

5:30-6:30 Reception (*Mirage II*)

Friday, April 27
(Kaleidoscope Room)
8:00 am Continental Breakfast

Overview Session---Part 2

8:30-9:00 Michael Phelps (UCLA), "Genetic Engineering, Molecular Imaging, and Molecular Drug Design"

9:00-9:30 Douglas Lauffenburger (MIT), "Cell Engineering: Quantitative Modeling and Experimental Studies of How Cell Functions Depend on Molecular Properties"

9:30-10:30 Open discussion
Jim Weiss (UCLA), moderator

10:30-11:00 Break

Mathematical Sciences and Disease States---Part 2

11:00-11:30 Dan Roden (Vanderbilt), "Using Genetics to Predict Individual Responses to Drugs—Hope or Hype?"

11:30-12:00 Bruce Levin (Emory University), "Mathematical Models of the Population Dynamics of Antibiotic Therapy"

12:00-1:00 Open discussion
Jim Weiss (UCLA), moderator
James Keener (Univ. of Utah), summarizer

1:00-2:00 Lunch

Dynamical Models of Cellular Processes

2:00-2:30 John Tyson (Virginia Tech), " CyberYeast: A Computational Model of Cell Cycle Regulation in Budding Yeast "

2:30-3:00 Byron Goldstein (Los Alamos), “Modeling Immunoreceptor Signaling: From the Generic to the Detailed”

3:00-3:30 Garrett Odell (Univ. Washington), title to be determined

3:30-4:00 George Oster (Berkeley), “The Mysterious Meanderings of Myxobacteria”

4:00-5:30 Open discussion
Iain Johnstone (Stanford), moderator
Leon Glass (McGill Univ.), summarizer

5:30-6:30 Reception (*Kaleidoscope Room*)

Saturday, April 28
(Kaleidoscope Room)
8:30 am Continental Breakfast

Neuroscience

9:00-9:30 John Rinzel (NYU), “Modeling the Thalamus in Sleep and Awake States”

9:30-10:00 Don Johnson (Rice Univ.), "Information Processing: Data Analysis and Theory"

10:00-10:30 Larry Abbott (Brandeis University), "The Effects of Noise on Neural Response Dynamics and Gain"

10:30-11:00 Emery Brown (Harvard Medical School), "Dynamics of Spatial Information Encoding in the Rat Hippocampus"

11:00-12:30 Open discussion
Peter Bickel (Berkeley), moderator
Keith Worsley (McGill Univ.), summarizer

12:30 Adjourn

Program committee: Peter Bickel, University of California at Berkeley
David Galas, Keck Graduate Institute
David Hoel, Medical University of South Carolina
Iain Johnstone, Stanford University
Alan Perelson, Los Alamos National Laboratory
De Witt Sumners, Florida State University
James Weiss, University of California at Los Angeles

This workshop is made possible by grants from the Burroughs Wellcome Fund, Department of Energy, Microsoft Corporation, and the Sloan Foundation.