What Mathematics Is Required to Make Use of Genomic Data?

Since genetic information has become available for many organisms across the biological spectrum, scientists are now seeking to understand how that information is manifested in the behaviors of cells, organs, organisms, and even communities of organisms.

Below we highlight some mathematical techniques, including statistics, algorithm development and refinement, and modeling of spatial and time-dependent phenomena, required to discover useful information in genomic data and to use that information to enhance biological understanding. The article stems from a report released this spring, *Mathematics and 21st Century Biology*, by a committee of the National Academies Board on Mathematical Sciences and Their Applications. That report, sponsored by the U.S. Department of Energy, gives an overview of the many past and current important interactions between the mathematical sciences and investigations at all biological scales, from genomes up to ecosystems. It makes a number of recommendations for increasing the rate of advance in computational biology.

Genomic data have embedded uncertainties stemming from the laboratory technologies used to obtain them. These variations in the data may or may not be of biological significance. Different sets of data obtained from monitoring a biological process need not be identical, even though they represent the same function. This kind of robustness provides a level of fault tolerance that living organisms need to survive. Given these characteristics, it is not feasible to talk clearly about genomics without the language of statistics.

In addition to providing the concepts and tools for modeling and manipulating information with uncertainties, mathematical techniques have proved invaluable for making genomic research more efficient. For example, it is often desirable to find a short DNA pattern within a longer DNA string. Algorithms have been created to find approximate locations of patterns in texts, to find the best relationship between two or more sequences, and to find the best overlap between two sequences. To try to find patterns in large sets of gene or protein expression data, scientists and mathematicians use machine learning algorithms—both supervised, where some initial structure is imposed on the data, and unsupervised, where no a priori structure is imposed.

Of course, investigations of genomic patterns are meant to inform and to be guided by the observable features of an organism. The traditional method is to estimate whether or not a given genomic region or gene has a causal influence on the appearance of a trait of interest, using a likelihood ratio statistic that expresses the odds of the observed data under the competing hypotheses. Nested models of this fashion can accommodate the evaluation of one genomic site while controlling for the effect of another site or an environmental factor to gain additional power. However, as a result of correlation among nearby variable sites in the human genome, test statistics of this nature are frequently conducted across numerous genes or large regions without a clear picture of what defines a statistically significant finding. As a result, permutation testing has become a critical and highly recommended component of evaluating the significance of these individual single-factor analyses.

More generally, mathematical models serve many key roles in our study of complex biological systems: they capture complex correlations among, and serve as a means to integrate the information in, diverse types of data. They encode substantive biological knowledge and represent our mechanistic and quantitative understanding of systems. In addition, they provide the analytical framework for estimation and inference of unknown parameters and for quantitative prediction.

Consider the modeling of a cell as a system of time-varying variables interacting with each other. A typical goal of the modeler is to reveal a functional structure, representative of what is known or reasonable about the evolution of these variables. Traditionally, these are assumed to evolve according to a set of ordinary differential equations. In the simplest treatment, this dependency is linear and the linear coefficients essentially capture how the various species of molecules affect each other. The analysis quickly becomes challenging when nonlinearity is introduced. The analysis of such coupled nonlinear ordinary differential equations to obtain qualitative understanding of their solution and the ability to make quantitative predictions will likely push the frontier of the theory of differential equations or will certainly require researchers with solid grounding in that theory. A more realistic model of cell processes would take into account the discrete nature of the system, thus involving Markov processes instead of continuous equations. Moreover, the structure of such models should also somehow capture the surprising robustness of many biological systems.

While the previous example illustrates the use of differential equations to construct models of time-dependent behaviors, modeling across the dimensions of both time and space is an important tool that is helpful in many scenarios. However, as spatial structure is taken into account, spatial complexity demands more complicated descriptions that include partial differential and integro-differential equations to account for chemical diffusion, active transport, and other means of communication between neighboring regions. While a large array of analytical tools has been developed to understand the behavior of spatially continuous systems, the understanding of these systems from a mathematical perspective is far from complete.

—Jennifer Slimowitz
—Scott Weidman
Board on Mathematical Sciences and Their Applications
National Research Council

Copies of the report Mathematics for 21st Century Biology *are available from the National Academies Press at 888-624-8373 or online at* `http://www.nap.edu`.

Letters to the Editor

Gender Studies

I read Roitman and Wood's letter "Gender and Mathematics—Again" in the May 2005 *Notices* with interest.

They refer to a "basic study replicated often" which sends the same vita or the same academic paper to some people under a male name and to other people under a female name. The "woman" is ranked lower whether men or women do the ranking.

If this "study" can indeed be replicated over time, its validity becomes very questionable because the respondents seem to represent a bad sample tested by questionable methods.

Where do you find people who consider themselves able to judge an academic paper and who are yet so naive as to not suspect some sort of scam when they recieve a "paper" from an unfamiliar source? Who, after two years in college, does not know of the existence of "bogus" set-ups for psychological studies?

Naive or not, they should have been protected by ethical constraints, in particular a requirement that prospective subjects in a study be made aware of factors that could influence their decision to participate. It is hard to believe that in study after study a reasonable sample would go knowingly through a charade of judging in order that they themselves could be judged.

I hope no AMS members were involved in this "basic study". Refereeing both papers and credentials is an onerous professional responsibility. Refereeing papers at the academic level, especially, is time consuming and is undertaken as a service to the community. Referees are not lab rats.

It seems to me that further attempts to replicate this study would require acceptance of a yet more credulous body of responders or yet more unpleasant schemes of deception and so should be abandoned. Furthermore, recent replications should be reexamined with a critical eye.

—I. David Berg
University of Illinois (retired)

(Received May 16, 2005)

Origins of Grothendieck's Pursuing Stacks

Allyn Jackson's excellent articles on Alexander Grothendieck (October and November 2004) seem to me slightly misleading on the 600-page 1983 manuscript "Pursuing stacks", which has become influential over the years, so I would like to point out that it was written in English in response to a correspondence in English, namely with myself and Tim Porter from Bangor. He sent copies to me and Larry Breen, and with his permission, I sent copies to a few people. So it began its circulation.

—Ronald Brown
University of Wales, Bangor
ronnie@ll319dg.fsnet.co.uk
r.brown@bangor.ac.uk

(Received May 27, 2005)

A Textbook Editions Policy

We have become increasingly concerned about a fairly common practice in the textbook publishing business. New editions are published on a regular cycle, with a period as short as four years, with little if any consideration of whether these new editions are justified on academic grounds. We understand that new book sales decrease in years following publication of new editions, as used copies become more available. Given the very high prices of new mathematics texts, it is not surprising that students often choose to buy used copies. To fight the reduction of income to publishers and authors, new editions appear for no apparent reason. This practice costs students money and forces often trivial but irritating changes in course outlines. This letter is prompted by the announcement that a particular text we have been using for a number of years is now going into the seventh edition. Successive editions of this text have appeared every four years for some time. In our view, they have gotten worse rather than better, because of the inclusion of more examples and verbiage that apparently are only intended to justify each new edition.

We have no problem with books that go into second and third editions because of the desire to correct errors, improve presentation, or change topics covered, based on the experience of users of the original version. But if the author can't get it right by the third edition, he/she should give up. The decision to publish a new edition should be based on pedagogy, not money.

To put some pressure on publishers to adopt what we consider a more responsible approach to this issue, the undergraduate studies committee of the UCLA mathematics department, at its meeting of June 7, 2005, adopted the following resolution: "Whenever one of our textbooks appears in a new edition beyond the third, if there is no evidence that it represents a significant pedagogical improvement over the previous edition, the mathematics department will immediately start a search for a replacement text."

At the same meeting, we decided to change the text mentioned in the first paragraph above. This text has also been used by the UCLA statistics department. Robert Gould, who is my counterpart in that department, has endorsed this letter.

—Thomas M. Liggett
Undergraduate Vice Chair
of Mathematics
UCLA
tml@math.ucla.edu

(Received June 10, 2005)

Correction

A photo caption on page 777 of the August 2005 issue incorrectly identified Congressman Vernon Ehlers (R-MI) as a senator, when he is, in fact, a member of the House of Representatives.