The launching of this new section of EHP *marks a critical stage in the evolution of the field of toxicogenomics.*

Toxicogenomics: An *EHP* Section

Kenneth S. Ramos

Thomas J. Goehl

Toxicogenomics is one of the newest fields of study that will play a major role in future research breakthroughs in environmental health. In January, when we expanded *Environmental Health Perspectives*'s (*EHP*) coverage of this new field by initiating a new quarterly section of the journal, the change was welcomed enthusiastically by the *EHP* readership. However, to clarify that articles in the Toxicogenomics Section are part of *EHP* and not a separate publication, we have instituted several changes: All articles will now carry an *EHP* citation, and each year all sections will have the same volume number with pages numbered consecutively. Articles will be abstracted immediately by abstracting services and will enjoy the same high impact factor as all other *EHP* articles. So that all articles in the Toxicogenomics Section have the same citation, we are re-publishing articles from the premier issue. Note that the Digital Object Identifier (DOI) code portion of the original citation remains unchanged.

We will consider articles for publication in the Toxicogenomics Section of *EHP* from the related disciplines of pharmacogenomics, proteomics, metabonomics, bioinformatics, molecular epidemiology, translational aspects of genomic research, and molecular medicine. The section has full color capabilities and features online publication of extensive data sets and supplementary materials. As with other *EHP* articles, accepted research articles will be published within 24 hours. These articles are completely citable using the DOI code that is managed by CrossRef, a licenseee of the International DOI Foundation. The *EHP* Toxicogenomics-in-Press articles can be found on our website (http://ehp.niehs.nih.gov/txg/).

Please join us as we explore the interactions between genes and the environment and the complexity of the biological circuitry involved in the cellular response to stressful environments. We will nurture the field by maintaining the highest standards of excellence. The launching of this new section of *EHP* marks a critical stage in the evolution of the field of toxicogenomics.

Kenneth S. Ramos
Toxicogenomics Editor, *EHP*
University of Louisville Health Sciences Center
Louisville, Kentucky
E-mail: ksramo01@gwise.louisville.edu

Thomas J. Goehl
Editor-in-Chief, *EHP*
Research Triangle Park, North Carolina
E-mail: goehl@niehs.nih.gov

Pluralitas non est ponenda sine necessitate.
William of Ockham (ca. 1280–1349)

Model Selection in Genomics

Ilya Shmulevich

With the discovery of DNA, the completion of genome sequencing of a number of organisms, and the advent of powerful high-throughput measurement technologies such as microarrays, it is now commonly said that biology has gone through a revolution. But I also have heard it said that biology is only about to go through a scientific revolution, much as physics did in the 17th century. In messianic hopes, people foretell the coming of the Newton of biology, but it is up to us, the scientific community, to set the stage for that to happen.

Both views are valid, each in their own sense. The discovery of DNA and the more recent development of powerful new technologies have certainly revolutionized our understanding of the inner workings of life and allowed us to probe deep into the machinery of living organisms, much as the Copernican system and Galileo's telescope helped revolutionize astronomy. It was Sir Isaac Newton, however, who placed science on a solid footing by formalizing existing knowledge in terms of mathematical models and universal laws. In some sense, this was the real scientific revolution because it permitted prediction of physical phenomena in a general setting, as opposed to simply describing individual observations. The difference is profound. Whereas a mathematical equation can adequately describe a given set of observations, it may be missing the needed universality for making predictions. Kepler's equations pertained to planets in our solar system. Newton's laws could be used to predict what would happen to two arbitrary bodies anywhere in the universe. The universality of a scientific theory coupled with mathematical modeling allows us to make testable predictions. This ability will have a profound effect on the field of biology.

The hallmarks of a great scientific theory are universality and simplicity. Newton's law of gravity is a case in point. The fact that the force of attraction between any two bodies is proportional to the product of their masses and inversely proportional to the square of the distance between them is both universal and simple. These issues are especially important today in the rapidly evolving field of genomics, where formal mathematical and computational methods are becoming indispensable. So what should be our guiding principles, our beacons of scientific inquiry? One such fundamental principle underpinning all scientific investigation is Ockham's razor, also called the "law of parsimony."

Consider the following, seemingly straightforward problem. We are presented with a set of data, represented as pairs of numbers (x,y). In each pair, the first number (x) is an independent variable and the second number (y) is a dependent variable. The problem is to choose whether to fit a line (of the form $y = a + bx$) or a parabolic function (of the form $y = a + bx + cx^2$). The knee-jerk response might be as follows: Let's fit the parabolic function, since the linear function is clearly a special case of it, just by letting $c = 0$; thus, the parabola will always provide a better fit to our data set. After all, if it so happens that our data points are arranged on a line, the estimation of parameters (*a*, *b*, and *c*) will simply reveal that *c* is indeed equal to zero and the parabolic function will reduce to a linear one. Thus, it would seem, three "adjustable" parameters are better than two. Of course, such reasoning could be taken *ad absurdum* if we had freedom to choose as many parameters as we like. Thus, there must be a tradeoff. Although three parameters surely provide a better fit to the data, the model becomes more complex and so, we sacrifice simplicity. But why is that bad?

To give a general answer, by making a model overly complex, we forfeit predictive accuracy. A complex model may be able to describe the observed data very well, but will it accurately predict future instances? For example, if the data contain random fluctuations or noise, an excessively complex model will "overfit" the data along with the noise and will obviously provide a poor fit to future (unseen) data. The chief goal of model selection is to find the right balance between simplicity and goodness-of-fit.

Consider gene expression–based cancer classification. The basic idea is simple: Take a number of tumor samples of a known type, measure expressions of thousands of genes for each one, and on the basis of these observations, construct a classifier (model) that will predict the tumor type when presented with an unknown sample. A fundamental question is "What type of classifier should we choose?" This is a crucial step in model selection (in machine learning, the model is called the "hypothesis space"). The next step—actually selecting a particular classifier from the model class (i.e., selecting a particular hypothesis)—is fairly well understood, as it involves the estimation of parameters.

As discussed, it would be unwise to devise an overly complex classifier, consisting of hundreds or thousands of parameters, especially in light of rather small sample sizes (number of tumors) available, which is typically below 100. Such a classifier may have extremely small or even no error on the seen data but may exhibit very high error on unseen data. Hence, its predictive accuracy would be very poor.

So, suitable criteria or methods are needed that would help us strike the right balance between simplicity and goodness-of-fit, such that predictive accuracy can be maximized. Fortunately, recent statistical literature is replete with various approaches, such as the Bayesian information criterion, Akaike's information criterion, minimal description length principle, and cross-validation methods.

In the field of toxicogenomics, issues related to prediction and model selection are of vital importance. For example, toxicogenomic biomarkers should reliably predict toxic effects to help us develop safer drugs and chemicals and understand molecular mechanisms of pathogenesis. Models of genetic networks and gene expression–based classifiers are expected to predict consistently a cell's response to a stressful challenge and to classify unknown compounds. A keen awareness of Ockham's razor will help guide us on our quest to understand the nature of living systems and their behavior under various environmental conditions.

Ilya Shmulevich
Cancer Genomics Laboratory
The University of Texas M.D. Anderson Cancer Center
Houston, Texas, USA
E-mail: is@ieee.org

Ilya Shmulevich is an assistant professor at the Cancer Genomics Laboratory at The University of Texas M. D. Anderson Cancer Center. He is an associate editor of the Toxicogenomics Section of Environmental Health Perspectives. *His research interests include computational genomics, systems biology, nonlinear signal and image processing, and computational learning theory.*

. . . just as genetic toxicology co-evolved with the fields of genetics and molecular biology, so will toxicogenomics co-evolve with the fields of genomics and systems biology . . .

On the 50th Anniversary of Solving the Structure of DNA

Leona D. Samson

As biochemistry students at Aberdeen University in Scotland, our class studied and strategized together to prepare for our final honors degree exams, and in the British tradition, the results of those final exams would, alone, determine our final grade after four years of undergraduate study. During that final academic year (1973–1974), the 20th anniversary of the famous Watson and Crick publication (Watson and Crick 1953) was being loudly celebrated in the scientific literature. Our class predicted that questions about DNA structure and function would be heavily represented, if not overrepresented, in the final exams. We were right. Thirty years later it is an unexpected pleasure to be invited to join the chorus, indeed the symphony, celebrating the golden anniversary of the DNA double helix and the sequencing of a complete human genome and to reflect upon how deciphering the structure of DNA was fundamental to the fields of mutagenesis and genetic toxicology and more recently to the emerging field of toxicogenomics.

I have studied various aspects of mutagenesis and genetic toxicology for nearly three decades, and upon looking back at the history of genetics and molecular biology (wherein Watson and Crick obviously played a pivotal role), it becomes immediately apparent that with each insight into the structure and function of DNA came an accompanying insight into how DNA structure and function can go awry. While Watson and Crick's discovery of the complementary nature of the bases inside the DNA double helix immediately suggested a mechanism by which DNA could replicate, it did not suggest how this molecule ultimately dictates the nature of all proteins present in the cell (Watson and Crick 1953). Indeed, even with an immediate insight into how DNA might replicate, it was 5 years (1958) until the beautiful Meselson and Stahl experiment (Meselson and Stahl 1958) demonstrated semiconservative DNA replication, as predicted by Watson and Crick. It was to take 13 years (1966) before the genetic code was finally cracked, and during those 13 years there emerged a reasonably complete picture of how DNA, mRNA, tRNA, and ribosomes collaborate to produce proteins of genetically predetermined sequence.

After the Watson and Crick paper in 1953, along with every experiment that produced an ever more detailed molecular picture of how DNA replicates and of how DNA makes RNA makes proteins, there came immediate insights into how each of these processes can go wrong. For example, until we understood the workings of triplet codons and the genetic code, we could not understand (at the molecular level) how changes in the DNA sequence might ultimately produce missense, nonsense, frameshift, and other mutations. A detailed understanding of DNA chemistry also led to an exploration of how chemical and physical agents could alter that chemistry. From this followed the concept that damage to DNA might lead to permanent sequence changes and thus to different kinds of mutation. This is not to say that damage to cells had not already been shown to cause mutations. Indeed, Muller demonstrated in 1927 that X-rays could induce heritable mutations in *Drosophila melanogaster*, and for this he won the 1946 Nobel Prize in Physiology or Medicine (Muller 1927). But this discovery was 25 years before Hershey and Chase (1952) finally convinced the scientific world that genes reside in DNA, and 26 years before the structure of DNA was solved (Watson and Crick 1953). Thus, although the fields of mutagenesis and genetic toxicology have a history long before the structure of DNA was discovered, it was only since 1953 that a molecular picture co uld be drawn of how toxic agents might interact with DNA to produce the biological end points of mutation and cytotoxicity. Moreover, the 1953 publication of Watson and Crick launched exquisitely detailed characterization of how DNA is faithfully replicated, and from this came an understanding of the role that DNA polymerases and such processes as recombination must play in the generation of DNA sequence changes. Parallel to these fundamental revelations were the observations that all organisms are equipped with a battery of genes that produce proteins whose primary roles are to prevent or repair chemical and physical damage to DNA; such activities protect against mutation and cell death induced by DNA-damaging agents, and studies of these activities eventually evolved into the field of genetic toxicology.

Genetic toxicology has been approached in two ways: *a*) with questions specifically aimed at understanding the molecular processes that influence the induction of DNA damage, and the toxic effects of such DNA damage; and *b*) with more general questions about the genes that influence the susceptibility of cells to toxic agents. The difference between these two approaches lies in the fact that the first is concerned only with toxicity resulting from genetic damage, and the second is concerned with genes that influence the toxicity of an agent, whether or not that toxicity emanates from damaged DNA. Both of these approaches to genetic toxicology are now evolving into the field of toxicogenomics.

With the dawning of the new millennium came one of the finest achievements in the history of biological research, namely, the sequencing of a complete human genome. Surely this was one of the most profound achievements to flow from the 1953 discovery of the structure of DNA. The working draft of this roughly 3.2 billion base pair sequence, the technological advances that were developed because of it, and the rapid electronic publication of the sequence as it was generated changed forever the ways in which biological and health-related research is being conducted. It is now possible, in principle, to address questions about *all* human genes in a massively parallel way, that is, questions related to the entire human genome, hence the term "genomics." The National Institute of Environmental Health Sciences (NIEHS) was very quick to realize the awesome potential of being able to interrogate the role of each and every gene in protecting humans against the detrimental health effects of exposure to environmental agents. The prescience of the NIEHS led to the launch of two major extramural research initiatives that have fostered the application of genomics to the environmental health sciences, namely, the Environmental Genome Project and the Toxicogenomics Research Consortium.

Several years ago the NIEHS established the Environmental Genome Project (http://www.niehs.nih.gov/envgenom/home.htm) to identify all common DNA variants, mainly single nucleotide polymorphisms (SNPs) for more than 500 human genes known (or likely) to influence cellular responses to toxic environmental agents. In the long term we will have an inventory of common SNPs for every gene in the human genome, but in the short term the Environmental Genome Project will provide us with focused information for genes already known to influence the biological consequences of exposure to toxic environmental agents. It is not difficult to imagine that it will soon be possible to screen individuals to determine their constellation of SNPs in these 500 or so genes deemed relevant to environmentally induced disease. This foray into genomic scale analysis will provide an important first step toward our being able to predict the response of an individual upon exposure to toxic environmental agents. However, it is quite clear that being able to identify the gene variants present in an organism is simply not enough. Genomic analyses must stretch far beyond the DNA to include RNA and protein; after all, DNA makes RNA makes protein. It is clear that we need to know the temporal aspects of how the environmentally relevant genes are expressed (in each cell type), as well as how their expressed products (RNA and protein) are modified and localized in the cell. We also must be able to predict how such expression, modification, and localization will change over time when individuals are exposed to environmental agents. Finally, armed with all this knowledge we must learn how to integrate the information into a systems biology view that not only is descriptive but also is predictive of the phenotype of cells, tissues, and ultimately people. We have not yet grasped how to do this, but we will have achieved one of the most exciting and powerful insights into biology when we find the ways.

The field of toxicogenomics has thus emerged to address these genomic-scale questions; moreover, the National Center for Toxicogenomics at the NIEHS recently established the Toxicogenomics Research Consortium (http://www.niehs.nih.gov/nct/trc.htm) to help launch and foster the development of the field. At the very least, transcriptional profiling using DNA microarrays and proteomic analysis using mass spectrometry represent the current major thrusts in toxicogenomics. In addition, the development of genomic approaches to systematically assess how each gene influences the phenotypic response of cells to environmental agents is well under way for model organisms such as *Saccharomyces cerevisiae*, and such "genomic phenotyping" is now being initiated for mammalian cells. It seems likely that within the next few years, libraries of small inhibitory RNAi constructs will be available for the systematic knock down of expression for each and every human gene in each of many different human cell types. It is inevitable that the fields of genomics and systems biology will mature as more efficient and sophisticated technologies emerge for quantitatively measuring global gene expression, global RNA and protein modification, and the dynamic trafficking and localization of cellular molecules. And just as genetic toxicology co-evolved with the fields of genetics and molecular biology, so will toxicogenomics co-evolve with the fields of genomics and systems biology.

The future test of toxicogenomics will be in our ability to predict accurately human susceptibility to the adverse effects of environmental agents. Perhaps, long before the golden anniversary of sequencing the human genome, it will be possible to determine individualized risk to environmental agents as part of a routine annual checkup. But before a time-line for this can even be envisioned, we must first learn to apply quantitative molecular assessments, engineering principles, and the informatics tools necessary to conduct successful predictive toxicology in model cellular systems.

Leona D. Samson
Biological Engineering Division and Center for Environmental Health Sciences
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
E-mail: lsamson@mit.edu

Leona Samson is professor of biological engineering and toxicology at the Massachusetts Institute of Technology (MIT), director of the MIT Center for Environmental Health Sciences, and a member of the Executive Steering Committee for a new Initiative at MIT in Computational and Systems Biology (CSBi). She is also an associate editor for the Toxicogenomics Section of Environmental Health Perspectives.

References

Hershey AD, Chase M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen physiol 36:39–56.

Meselson M, Stahl FW. 1958. The replication of DNA in *Escherichia coli*. Proc Natl Acad Sci USA 44:671–682.

Muller HJ. 1927. Artificial transmutation of the gene. Science 46:84–87.

Watson JD, Crick FHC. 1953. The structure for deoxyribose nucleic acid. Nature 171:737–738.