In the Postgenomic Era, Can High-Dimensional Analyses Deliver the Answers We’re Looking for? 1, 2

Stephen Barnes*

Department of Pharmacology and Toxicology, UAB Center for Nutrient–Gene Interaction in Cancer Prevention, and Mass Spectrometry–Proteomics Shared Facility, Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294

Expanded Abstract

Although significant investigation of the health of this nation (the United States) underwent a significant step forward with the establishment of the National Institutes of Health in 1948, most younger investigators (and many politicians) see the beginning of this millennium as the place where substantive research began because it was then when the global and detailed features of the human genome became clear for the first time. This expansive vista was supposed to radically change our ability to treat, cure, and prevent disease. Scientists, entrepreneurs, and clinicians alike looked to it for unique solutions to the pressing problems of disease and aging. Although large by many standards, the total number of genes discovered in the human genome was a disappointment. However, it spawned new (sic, old) technologies, the study of proteins and metabolites, and gave them new names, proteomics and metabolomics, to take part in the search for recognizable signatures of disease. Serum, plasma, urine, and other physiological fluids were collected, separated, and analyzed at an ever-increasing pace and scope. Many promises were made, and discovery became the standard for research excellence. However, this frenetic pace of research largely ignored other scientific fundamentals of good science. It is worth taking stock of where we are and what it will take to deliver the solutions we so urgently strive for.

The study of disease systems is in an analogous position to where physics was at the turn of the 20th century, except that fundamental particles in physics are now in the form of genes, proteins, and metabolites. It is realized that what follows from these discoveries is an integration of biological knowledge, a discipline known as bioinformatics or advanced library networking. The Internet, because of its virtual library capabilities, allows for an enormously diverse group of facts to become immediately available after the initial discovery of a protein. Although this is impressive, it does not necessarily take us much further forward in real understanding. Indeed, the view of biology that the application of all the current methodologies allows is merely classical in nature. And it suffers from being for the most part low dimensional, particularly in the time domain. This view limits the understanding of how biological systems really work. Where the future lies is contained in perturbations of systems in time. Biological stimuli are not the square waveforms used by most investigators in cell culture studies but instead are highly rhythmic and thereby are in resonance with processes in cells. And by further analogy to physics, it is likely that biological systems are quantized and will exhibit weird behavior. An example of a quantized system is a group of cells that have on average 3 transcripts of a rare transcription factor (the classical view). If all cells in this group were analyzed, they would be instead shown to have a distribution of transcripts, centered around 3. The number of transcripts in any 1 cell cannot be predicted, and if one resampled a cell with no transcripts at a time in the past, it would have finite probability of having a transcript now. This is somewhat similar to determining the position and motion of an electron is in an atom. Uncertainty is a feature of biology as well as physics.

To probe how a complex system responds, we currently have DNA microarrays to measure gene expression in cells (1), proteomics methods to identify proteins (2,3), examine protein abundances (4) and protein modifications (5), and metabolomics to determine how metabolite levels change (6) in physiological fluids and tissues. Each of these methods is a technical tour de force within the context of the analytical chemistry that is available in the early 21st century.

DNA microarrays depend on the specific hybridization of message (mRNA) with short oligonucleotides (20–25 bp) that are specific for an individual gene in a genome. Related methods use spotted cDNAs; these are more reliable because they have a much greater hybridization but suffer from cross-reactivity between gene homologs. Most of the known genes in a genome can be spotted on a single glass slide. Although this is at first sight very impressive, the statistics of such systems are poorly understood; there is a competition between optimizing discovery and controlling the false discovery rate. And even if a gene is shown to have significantly changed, does this mean that it is a meaningful event? It is argued that what occurs at the level of
proteins is the more important event. The magnitude of the change may also be misleading: for proteins that regulate major events in a cell, the range of normality is often very small. Also, the locale of the protein is just as important as its amount.

The analysis of the proteome is still rather primitive in nature. Two approaches have been tried; in 1, proteins from either the entire cell lysate or a particular subfraction are hydrolyzed to create a peptidome that is analyzed by 2-dimensional liquid chromatography–mass spectrometry (2D-LC-MS) (3). The advantage of this technique is that it is easily automated. However, it suffers from the loss of information as the protein is fragmented into many pieces. Also, the number of tandem mass spectrometric (MSMS) spectra that were acquired (50,000–60,000/run) is overwhelming and has deterred most from ever questioning the quality of the data. However, 2 pieces of work brought some sanity to the science. The first from Natalie Ahn’s group showed that almost the same number of peptides detected with this procedure were also detected when the peptidome database was reversed in silico before the data were fitted. This has led to a more stringent evaluation of the fits made by this procedure. The second problem came from a comparative study on the analysis of a common reference sample by 6 commercial companies. Only 3% of the >1700 proteins detected in the sample were observed by all 6 companies. This lack of reproducibility would be a death knell for application in biomarker analysis. Many labs, after the initial euphoria of discovery, are taking a more systematic and quantitative approach to sample analysis.

The alternative approach relies on first separating the proteins (2) and then carrying out the mass spectrometry analysis (7). 2D-electrophoresis requires strong technical skills. However, all-liquid 2D-protein separation analytical systems have been introduced (8,9). The advantage of protein separations is that it enables rigorous tests of the hypothesis that we can predict the proteome. Even if the sequence of a peptide does not match any from the gene-based proteome, we can be certain that the protein exists.

Metabolomics is not new; it’s a method that has been used in biochemical investigations ever since A. J. P. Martin and his colleagues, R. L. M. Synge, and A. T. James developed partition and gas-liquid chromatography (10,11). As for many other techniques, today’s methods are just more sensitive and more adapted to computerized analysis. In terms of understanding how a system works, sustains itself, and responds or resists external stimuli, ultimately a lot depends on the levels of the small molecules: without ATP and other key intermediates, the cell won’t function even with its otherwise full set of expressed genes and proteins. Enzymes are very responsive to small molecules; their observed activity at a point in time is a function of their intrinsic activity in combination with the concentration of substrates and inhibitors.

Extracting information from genomics, proteomics, and metabolomics can be either simplistic (and largely erroneous) or difficult (because of the need to optimize experimental design, carry out many repeat experiments, and develop robust statistical methods for the analysis of the data). What is really needed is the development of rigorous mathematical models of the systems (or parts of them) that we are studying. We might call this part of the approach dynamic bioinformatics. It is an engineering phase of understanding biological systems.

At the present time, we are largely only able to take simple time slices of a complex system: we are the observers in the types of conundra that Einstein considered in developing his theories of relativity. We photograph (i.e., gene arrays and protein gels) moments in a system, assuming a priori that what we record is related. To go further in our understanding, we must recover the time and spatial domains. We assume specific concentrations of compounds produce biological effects, but it is more likely that it is the changing, wavelike variations in concentration that are the real signals in a cell. Measuring those changes is the challenge that faces us all. Although this is yet to occur for mammalian systems, some progress has been made by those who study microorganisms (12).

**Literature Cited**