

GENOMICS

Designer science and the “omic” revolution

Glen A. Evans

Following the success of the human genome project effort, several other “omic” disciplines have emerged, with the goal of analyzing the components of a living organism in its entirety. Proteomics (the complete set of proteins produced in a cell), phenomics (the complete set of mutational phenotypes), epigenomics (the complete set of methylation alterations in the genome), ligandomics (the complete set of organic small molecules), and so forth, have each focused on the accumulation of biological information of a molecular type. As the “omic” strategy progresses, additional new fields will become part of the biological lexicon, with increasing volumes of molecular data added to ballooning databases. So widespread has this approach become that the term “omics” has been taken to be a general term of reference to studies of entities in aggregate¹.

Array technology

Much of the recent accumulation of biological data has been accomplished through the use of microarray techniques. This technology is one of the most important developments of post-sequencing genomics as it allows the large-scale parallel assessment of gene expression profiles in rapid and reproducible benchtop experiments. Thousands of DNA hybridization probes are attached to microscale arrays on glass or silicon, hybridized with fluorescently labeled whole-cell cDNA, and the quantitative signals on each array element are measured in parallel. In spite of the simplicity of the approach, powerful insights can be obtained into biological function and clustering of coregulated genes².

In spite of the ability to generate large quantities of data quickly, microarray expression analysis depends largely on the assumption that mRNA levels reflect protein levels, that coregulated genes are functionally associated, and that hybridization signals reflect intracellular biological processes. Current microarray technology provides information on the levels of mRNA, but most biologists, if given the choice, would rather know the levels of particular proteins, the activities and reaction rates of enzymes, and transport processes through cellular membranes at given times. We might expect

the model of microarrays to be expanded to arrays of proteins, antibodies, enzymes, and organic compounds, analyzed in parallel, to fill out the rich biological data set necessary to define a living organism.

Simulation space

“Omic” databases of the future can be conceived with gene function and interactions in mind, such that the database would comprise a simulation space allowing mechanism and organismal function to be re-enacted. Rather than compiling individual “omic” databases in isolation, a stratagem or organizing paradigm is needed to allow the wide variety of data to be eventually integrated into a single model of biological function. One approach is to use the “central dogma” of molecular biology in its simplest form, accumulating data contributing to the virtual information flow from gene to metabolic or structural function.

In this simple strategy, the “omic” parameters we would like to know and could seek to accumulate would include the number and replication rates of genes, transcriptional rates, mRNA degradation rates, protein synthesis rates, protein metabolic function (if an enzyme), gene targets (if a transcription factor), protein partners (if part of a structure or multicomponent protein), binding constants for ligands and other small organic compounds, concentrations of substrates and products, and other related parameters. Such data, accumulated from sets of “array-style” strategies and grouped in a database, would comprise a “simulation space.” An “omic” simulation would contain for each gene an estimate of each critical parameter, as well as rules for interactions at each level to provide networks of interactions and allow estimations of biological functions. When an “omic” experiment is completed, the data could be entered in the simulation and the model “run” in a context where predictions could be made and tested. The resulting database would comprise an *in silico* biological laboratory.

On the basis of these emerging principles and data sets, some impressive starts have been made on biological data framework models in the form of simulation tools like the *e*-cell of Tomita³ and the metabolic simulation space of Palsson⁴. What is now needed is a simulation space data structure that can be applied to emerging “omic” data sets and where results of array analysis of each step in molecular analysis can be placed in a context in relationship to other data types. Such a simulation space, when completed with all of

the 140,000 or so human genes (including RNA expression levels, proteomics, metabolic, signal transduction, and protein–protein interactions data), will provide a first pass at the simulation of processes within a living cell. Ultimately, based on a more profound understanding of gene networks and systems rules than currently available, the simulation space would allow construction of an *in silico* mammal.

Beyond “omics”

The “omic” approach to large-scale biology has now revealed the complete genetic blueprint of almost 30 organisms from archeons to *Drosophila* and *Caenorhabditis elegans*, all of which are available and accessible from GenBank. Saturation transposon-mediated mutagenesis of the simplest organisms has been used to refine the estimated number of minimally required genes to support independent life⁵. For instance, among the 480 protein-encoding genes of *Mycoplasma genitalium*, the simplest known free-living life form with a genome size of only 580 kb, only 265–350 genes are required to support the basic life processes of metabolism, replication, and homeostasis. The end result of these simulation spaces may be the genomic bioengineering *in silico* of novel microorganisms based on the knowledge of interacting systems and networks of genes and gene products.

The ability to engineer genomes on the wet laboratory bench and develop DNA-based bioengineering of novel life forms would follow directly from *in silico* simulation. It is now routine to synthesize large collections of specific oligonucleotides in microarrays and to use these to systematically assemble synthetic genes. Following a successful simulation *in silico*, modifications of existing robotics devices could be used to facilitate the total synthesis of DNA molecules for the assembly of a synthetic genome. Designer genomics based upon known genes, protein products, and interacting networks extracted from databases of many different organisms could be combined into novel “alien” life forms with diverse properties. Synthetic genomics may become a mainstay of future biotechnology much as recombinant vectors are today.

Glen A. Evans is director of the Genome Science and Technology Center, The University of Texas, Southwestern Medical Center, Dallas (gevans@utsw.swmed.edu).

1. Weinstein, J.N. *Science* **262**, 628 (1998).
2. Marshall, E. *Science* **286**, 445 (1999).
3. Tomita, M. et al. *Bioinformatics* **15**, 72 (1999).
4. Schilling, C. & Palsson, B. *Proc. Natl. Acad. Sci. USA* **95**, 4193 (1998).
5. Hutchinson, C.A. et al. *Science* **286**, 2165 (1999).