

## COMPLEXITY

# The practical problems of post-genomic biology

Sui Huang

In his visionary commentary, Glen Evans<sup>1</sup> describes the emergence of new post-genomic disciplines like proteomics, epigenomics, and phenomics. He proposes that integrated “omic” databases in the future will allow the in silico re-enacting of the entire organism once the totality of the information at the various levels, from genome to proteome and functions, is available. This is an attractive idea: just think of virtual organisms that could serve for virtual clinical trials. However, the “omic” strategy is based on the same linear, monocausal, deterministic thinking that has reigned in molecular biology over the past decades<sup>2</sup>; it just extrapolates it to the entire genome. This is brute-force, genocentric reductionism in the guise of *entireness*, rather than a novel integrative approach devoted to *wholeness*, or as *Nature Biotechnology* put it in an editorial: “Complicated is not complex.”<sup>3</sup>

In fact, the present efforts in the drug development industry to simulate organismal function still ignore the very elementary parts, and instead employ the traditional “top-down” modeling approach of physiologists and systems engineers, in which organs and cells are black boxes. Such approaches vastly dominated a recent workshop on computer modeling in biology, despite the promising conference title: “From gene to organ.”<sup>4</sup> With their dream of simulating organisms “from bottom-up,” genomics scientists touch upon the old riddle of genome-to-phenome mapping. They will have to overcome the mentality of data collecting, clustering, and classification, and develop a qualitatively new conceptual understanding of the complexity of living organisms. Help in tackling this daunting task is already on its way: not only are physical scientists increasingly interested in studying complex systems in general, but one of the first institutions devoted to such studies, the New England Complex Systems Institute (NECSI) in Boston, MA, has launched collaborations with molecular biologists to harness the data flood triggered by the success of genomic technology.

To counter, but not stifle Evans’ enthusiasm, let me give an overview of some of the

major obstacles and challenges that we will have to confront in a post-genomic endeavor that will ultimately lead to the design of cyber-guinea pigs based on our genomic databases:

*Ontological questions concerning database structure and representation of “gene function”.* A major goal of making genome databases more user-friendly is the systematic annotation of genes with functional properties, such as the biochemical function, interaction partners, and physiological role. This has worked pretty well for simple organisms, as demonstrated in the *Saccharomyces* Genome Database<sup>4</sup>. However, in more complex organisms, the same task represents an ontological challenge, as most proteins are embedded in a regulatory network and don’t just code for metabolic enzymes. For instance, genes such as *ras*, *myc*, *rho*, and *NF-κB* either stimulate growth and survival, or they induce apoptosis. The “function” of a gene product appears to depend on its “cellular context.”

This raises fundamental questions. For example, what is a “gene function?” What is its “context?” And what is the modular entity to which the concept of context applies: a protein domain, a single protein, a stable complex, or a whole pathway? Recently, the principle of “modularity” in cellular processes has been proposed to facilitate the task of dealing with biological functions<sup>5</sup>. For example, a specific signal transduction pathway could be viewed as a functional module. However, it remains to be seen whether modules are “natural kinds” or just logical constructs of our mind and how their boundaries have to be delineated given the extensive “crosstalk” between the historically defined pathways. These questions must be addressed first before useful higher level “omic” databases can be constructed.

*Limitations of computational power.* In proposing the in silico re-enacting of organismal function, Evans seems to ignore the immensity of computational capacity required. Take the human genome with 100,000 genes and let every gene be simply either “on” (expressed) or “off” (silent). This minimal, idealized, and discrete setting alone would lead to the astronomical number of  $10^{30,000}$  possible gene expression profiles! The computing and testing of all these patterns with the existing serial computers would take more time than the age of

the universe. The cell, in contrast, in which zillions of molecular events occur at a time, computes in a parallel fashion. Moreover, the cell has evolved a wiring architecture for genomic interactions that enables it to spontaneously choose stable, physiological expression profiles<sup>6,7</sup>.

Even with the envisaged knowledge of the genomic wiring architecture and possible novel algorithms that could facilitate our computing approaches, many generic problems in genomic-scale computation, such as analysis of genome-wide interaction maps, are thought to belong to the class of problems known as “hard NP” to computer scientists. On traditional serial computers such problems can have computational running times that grow exponentially with the number of variables. Feeding “omic” databases into computers and running them to simulate life would require parallel supercomputers. However, in contrast to traditional computers, parallel computers must be built to the specifications of the given task, which for simulating phenomes out of genomes still awaits formalization. With the need for ever-increasing simulation capacity, biology will join the physical sciences as a discipline where simulation capacity is a notorious rate-limiting factor. Ironically, in an opposite development, computer scientists facing the limits of their machines have started to explore the possibility of moving from in silico to in vitro computing to exploit the inherent parallelism of mixtures of specific DNA molecules<sup>8</sup>.

*Lack of formal understanding of the design principle of large biochemical networks.* Unexpected results in metabolic engineering have taught us some illuminating lessons. For instance, an *Escherichia coli* strain with a mutated pyruvate kinase gene shows the same central carbon flux ratio as the parental wild-type strain, and knockout mice lacking a critical gene that fail to exhibit the expected phenotype are no sensation anymore. These examples point to an “emergent property” of biochemical networks at a higher level of integration that defies traditional genetic determinism; they illustrate our inability to predict such aggregate behavior from the knowledge of the individual pathways.

Several attractive concepts have been proposed in the past for understanding the integrated behavior of metabolic networks and predicting the outcome of metabolic

Sui Huang is research fellow at the department of surgery, Children’s Hospital and Harvard Medical School, Boston, MA 02115 ([huang\\_su@al.tch.harvard.edu](mailto:huang_su@al.tch.harvard.edu)).

## COMMENTARY

engineering<sup>9-11</sup>. However, we are still far from a unifying, formal understanding of the link between functional macroscopic properties and the architecture of large regulatory networks of intertwined, nonlinear, stochastic interactions exhibiting quasi-stationary states far from thermodynamic equilibrium. Analytical approaches to small biochemical networks using mathematical tools for describing nonlinear dynamic systems has brought some insights into design principles of biochemical circuitries<sup>12</sup>. Importantly, they have addressed generic phenomena such as feedback regulation, hysteresis, multistability, robustness, which were counterintuitive to the biologist used to thinking within the linear framework of the central dogma of molecular biology. However, it remains unclear whether living organisms with larger networks of interacting parts can be reduced to explicit differential equations. Recent studies combining computer modeling with recombinant DNA technology approaches to real, well-documented, small networks, such as the one involved in bacterial chemotaxis, are illustrative examples of the depth of understanding to be targeted in the near future<sup>13</sup>. The extension of such approaches to larger networks of higher organisms will have to incorporate multiple layers of interactions, including those established by physical, humora and neutral intercellular communications. This will be the next big challenge.

*Uncharacterized consequences of a distinct physicochemical reality at the microdimensions of the cell.* Most quantitative modeling approaches rely on our chemical intuition gained from the study of ideal, dilute, homogeneous solutions where the laws of mass action allow the simple formalization of molecular reactions. The fact that conditions in the cell are far from ideal has many important kinetic and thermodynamic consequences. First, the small number of individual molecules, ranging from two (specific promoter elements) to a few thousand (regulatory proteins), precludes the use of concepts like concentrations or equilibrium constants. Such characteristic variables are derived from the statistical mechanics treatment of large numbers of molecules. Second, the local crowding of macromolecules in the cell affects structure, function, and diffusion of biopolymers<sup>14</sup>. Third, cellular biochemistry, such as signal transduction cascades, occurs not in free three-dimensional space, but on the surface of membranes and cytoskeletal structures. This leads to fractal and non-Michaelis-Menten kinetics<sup>15</sup>. Finally, the small number of specific reaction partners gives rise to relatively large, random fluctuations<sup>16</sup>.

It is still debated whether and when this noise is a disturbance (necessitating the design of robust regulatory networks) or is even exploited by the organism. For example, randomness can be used to generate variability in pattern formation, for the noise-driven transport of macromolecules based on a ratchet mechanism or for signal processing based on stochastic resonance<sup>17</sup>. Taken together, these microscale reaction

**Features of the very same system depend on the scale of observation. This precludes the extrapolation of knowledge at one level to higher levels where the "complexity" increases.**

conditions within the cell preclude a modeling approach that relies on traditional chemokinetics using "macrovariables" such as concentrations and rate constants. The development of novel visualization tools and the advent of cellular nanotechnology will pave the way toward an understanding of the real "device physics" of a cell, which in turn will provide access to the relevant "microvariables" for a precise bottom-up modeling.

*Structural and mechanical complexity of cells.* Genomic science is concerned with the abstract level of information storage and processing in cells while neglecting their mechanical structure. Borrowing terms from the world of computers, genomics and its "omic" grandchildren can be said to solely deal with the software, and to neglect the hardware of living systems. Yet the organism is more than an extended biochemical network obeying logical rules and chemical laws. Genomics scientists should be reminded that the organism is also a physical entity with geometric dimensions and thus is subjected to the laws of macroscopic mechanics. Understanding gene function must include knowledge about the hardware aspect. Cells are compartmentalized, and localization of proteins affects their function. Information transfer takes place not only through the specificity of protein binding. The cell responds to topological clues and mechanical forces that play a central role during morphogenesis and yet do not encode genetic information. For example, within the same biochemical milieu, a normal mammalian cell will divide, differentiate, or apoptose depending just on its externally

imposed shape<sup>18</sup>. With the advent of microfabrication in cell biology, bioengineers are designing powerful tools, like the microelectromechanic systems (MEMS), which will become available for biologists who wish to study the microscopic hardware of the living organism.

*Biology and the new science of complex systems.* The fundamental question of how simple parts (self-) assemble into a complex whole with novel properties is attracting an increasing number of physical scientists who are now consolidating the new science of complexity. As physicist and Nobel Prize laureate Phil Anderson acknowledged, psychology is not applied biology is not applied chemistry is not applied physics. At each level, novel laws can be found whose necessary, separate study has defined each discipline. Features of the very same system depend on the scale of observation. This precludes the extrapolation of knowledge at one level to higher levels where the "complexity" increases (and vice versa). Understanding why this is so, and determining how to formalize the problem of emergent features and multiscale description is one of the goals of the science of complex systems. Biology, whose object of study extends within one same entity across many scales, from molecule to animated organism, should thankfully embrace the efforts and watch their progress carefully<sup>19,20</sup>.

\*"From Gene to Organ," February 23-27, 2000, Georgia Institute of Technology, Hilton Head, SC.

1. Evans, G. *Nat. Biotechnol.* **18**, 127 (2000).
2. Strohmman, R.C. *Nat. Biotechnol.* **15**, 194-200 (1997).
3. Complicated is not complex (editorial). *Nat. Biotechnol.* **17**, 511 (1999).
4. Ball, C.A. et al. *Nucleic Acids Res.* **28**, 77-80 (2000).
5. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. *Nature* **402** (Suppl.), C47-C52 (1999).
6. Kauffman, S.A. *The origins of order*. (Oxford University Press, New York, NY; 1993).
7. Huang, S. *J. Mol. Med.* **77**, 469-480 (1999).
8. Chen, J. & Wood, D.H. *Proc. Natl. Acad. Sci. USA* **97**, 1328-1330 (2000).
9. Kacser, H. & Burns, J.A. *Biochem. Soc. Transact.* **7**, 1149-1160 (1973).
10. Schuster, S., Fell, D.A. & Dandekar, T. *Nat. Biotechnol.* **18**, 326-332 (2000).
11. Schilling, C.H. & Palsson, B.O. *Proc. Natl. Acad. Sci. USA* **95**, 4193-4198 (1998).
12. McAdams, H.H. & Arkin, A. *Ann. Rev. Biophys. Biomol. Struct.* **27**, 199-224 (1998).
13. Alon, U., Surette, M.G., Barkai, N. & Leibler, S. *Nature* **397**, 168-171 (1999).
14. Zimmerman, S.B. & Minton, A.P. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 27-65 (1993).
15. Savageau, M.A. *J. Theor. Biol.* **176**, 115-124 (1995).
16. McAdams, H.H. & Arkin, A. *Trends Genet.* **15**, 65-69 (1999).
17. McClintock, P.V. *Nature* **401**, 23-25 (1999).
18. Huang, S. & Ingber, D. *Nat. Cell Biol.* **1**, E131-E138 (1999).
19. Coffey, D.S. *Nat. Med.* **4**, 882-885 (1999).
20. Bar-Yam, Y. *Dynamics of complex systems* (Perseus Books Group, NY; 1997).