The challenges of in silico biology

Moving from a reductionist paradigm to one that views cells as systems will necessitate changes in both the culture and the practice of research.

Bernhard Palsson

The advent of high-throughput technologies, such as genomics and proteomics, is enabling biologists to study cells as systems. This is not only creating a whole new set of logistical problems, but also forcing a conceptual reevaluation of the concept of cells as a collection of individual cellular components. What do we do with this developing list of cellular components and their properties? As informative as they are, these lists basically give us the molecules that make up cells and their individual chemical properties. How do we now arrive at the biological properties that arise from these detailed lists of chemical components?

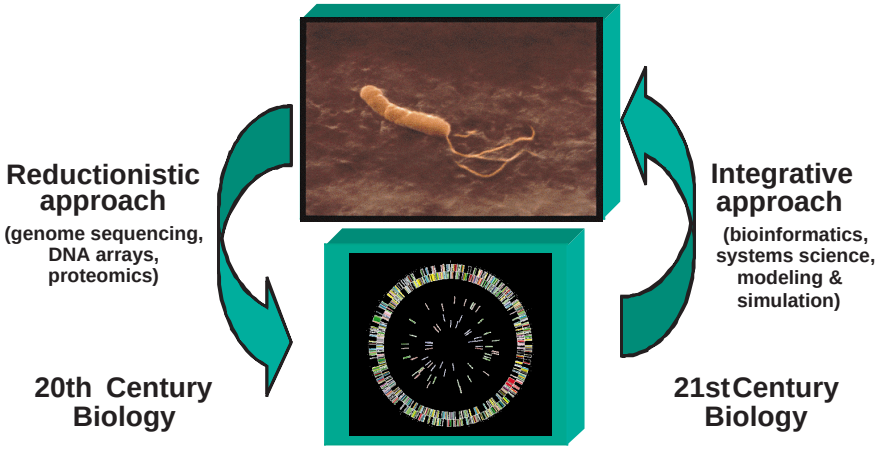

Figure 1. The shift in emphasis of biological research. Biology has traditionally followed a reductionist approach in which individual components of a living system are studied separately. It is becoming clear that we need to reverse the process and to study how these components interact to form complex systems using an integrative approach.

From reductionism to globalism

During the latter half of the 20th century, biology was dominated by reductionist approaches that successfully generated information about individual cellular components and their functions. Over the past decade, this process has been greatly accelerated by the emergence of genomics. We now have entire DNA sequences for a growing number of organisms and are continually defining their gene portfolios. Although functional assignment to these genes is presently incomplete, we can soon expect the assignment and verification of function for the majority of genes on selected genomes. Extrapolation between genomes will then most likely accelerate the definition of what amounts to a "parts catalog" of cellular components in a large number of organisms[1]. Expression array and proteomic technologies give us the capability to determine when a cell uses particular genes and when it does not. The reductionistic process is schematically depicted on the left in Figure 1.

However, it has become generally accepted that the integrative analysis of the function of multiple gene products has become a critical issue for the future development of biology (e.g., see refs 2–8). Such integrative analysis will rely on bioinformatics and methods for systems analysis (right side of Fig. 1). It is thus likely that over the coming years and decades, the biological sciences will be increasingly focused on the systems properties of cellular and tissue functions. These are the properties that arise from the whole, and represent "real" biological properties. These properties are sometimes referred to as "emergent" properties because they emerge from the whole and are not properties of the individual parts. This new challenge comes with several fundamental scientific issues and implications for the biological community. Only a few will be addressed here.

In silico biology

As in other fields before, biology will experience an increased use of systems mathematics and computer simulations. We have already begun to experience this trend, and it is likely to continue. Many other fields of science and engineering have developed systems science and complicated mathematical simulations to a high level of sophistication. These capabilities influence our everyday life. When placing a telephone call, one enters a complex and optimized network. The chemicals that we all use originate from refineries and other highly integrated chemical processes with complex control structures that rival those of living cells. Aircraft pilots are trained in simulators, and the aerospace industry now simulates aircraft designs so accurately in a computer that prototypes are no longer built. This would have been unthinkable only a few years ago. Many fields of science and engineering have thus gone through productive periods of rapid data generation, data analysis, mathematical model building, and computer simulation.

What about biology? Will it ever reach a level of sophistication in mathematical modeling and simulation similar to other fields? Opinions are mixed. The complexity of living systems and their continual change through evolution makes many skeptical about the success of such endeavors. Of course, only time will tell how successful they will be.

With the Human Genome Project's completion at hand, and with increasing amounts of expression data becoming available, growing attention is being paid to in silico biology. Broadly speaking, the term in silico biology refers to the use of computers to perform biological studies. Computations of the structures of complex biomolecules are currently routinely performed. Now, the mathematical description and computer simulation of the simultaneous action of multiple gene products is growing in importance, and in the view of many, will take center stage in biology in the coming decades.

How will we proceed?

Mathematical model building in biology is likely to differ, at least initially, from that practiced in the physicochemical sciences. In these fields, one starts with basic thermodynamic concepts such as chemical potential, fundamental rate equations such as the diffusion equation, and the basics of electrochemistry such as the Nernst equations. These

Bernhard Palsson is professor at the Department of Bioengineering, University of California–San Diego, La Jolla, CA 92093-0412 (palsson@ucsd.edu).

equations are based on fundamental physical theories and concepts, and contain a large number of parameters, most of which can be individually measured. Computer models of complex processes have information both on the properties of each component in the system as well as on their interconnectivity.

In spite of impressive bioinformatic databases, we cannot obtain all of the information needed to build a computer model of a whole cell at this detailed level of description. One day, this goal might be achieved, but at present a different approach is needed if we want working and useful computer models of whole cells. At present, we can obtain the network structure of multigenic processes (e.g., through knowledge of stoichiometry and the use of yeast two-hybrid systems), but obtaining information about the physciochemical properties of gene products, such as binding constants and turnover rates, is much more difficult.

In the absence of detailed information, an alternative approach can be formulated that is based on the fact that cells are subject to certain constraints that limit their possible behaviors. Imposing these constraints, one can then determine what is possible to a cell and what is not. Imposing a successive series of constraints, one can limit likely cellular behavior, but never predict it precisely. This approach is illustrated on the left in Figure 2. It leads to the formulation of solution spaces rather than the computation of a single solution. Behaviors within this space can be attained, each basically representing a different phenotype based on the component list, the biochemical properties of the components, and the imposed constraints. If all of the constraints are known, the solution space shrinks to a single point, as shown on the right in Figure 2. The question then becomes, will we ever reach this state of knowledge of cellular processes? Most likely not, at least in the foreseeable future, except in rare cases such as for the human red blood cell[9] or for simple viruses[10]. However, this approach does lead to models that are helpful in analyzing, interpreting, and even predicting the genotype–phenotype relationship.

Table 1. Physicochemical factors constraining metabolic function

Factor	Type of constraint
Capacity	
Maximum fluxes	Nonadjustable maximum based on maximum association rates
Connectivity	
Systemic stoichiometry	Hard nonadjustable constraints
Rates	
Mass action, enzyme kinetics, regulation	Highly adjustable by an evolutionary åprocess
Others	
Osmotic pressure, electroneutrality, solvent capacity, molecular diffusion	Hard nonadjustable constraints

Types of constraints and their imposition

Cells are subject to a variety of constraints (see Table 1); there are both invariant (i.e., nonadjustable) and adjustable constraints. The former can be used to bracket the range of possible behavior. The latter can be used to further limit allowable behavior, but these constraints can adjust through an evolutionary process. In addition, the adjustable constraints, such as kinetic constants, will vary from one individual to another. A set of successive constraints can be applied to the analysis of metabolic fluxes to narrow attainable flux distributions achievable from a defined metabolic genotype (see Fig. 3).

The first part of Figure 3 shows a space where the axes represent fluxes through all individual reactions in the metabolic network. Not all the points in this space are attainable because of the interrelatedness of the fluxes. The stoichiometric matrix limits the steady-state fluxes to a subspace, and metabolic transients are rapid so any deviations from this subspace are short-lived. If the reactions are defined so that all the fluxes are positive, this plane is converted to a cone through the use of convex analysis. The edges of this cone become a set of unique, systemically defined metabolic pathways (see review in ref. 11), and all the points on the interior of the cone can be represented as positive combinations of these fundamental pathways. Because of the capacity constraints on the individual steps in the pathways, the length of each edge is limited. These capacity constraints close the cone (step 3 in Fig. 3) and form a closed solution space in which all allowable metabolic flux maps lie. This space can be searched for optimal phenotypes using linear optimization[12,13]. Recent experimental studies in my laboratory have shown that growth of *Escherichia coli* lies along an edge that represents optimal growth in minimal media. With this information in hand, one can seek the kinetic constraints that force the solution to the edge of the closed cone.

The application of successive constraints to metabolism is probably just the first such example. It is an approach that marries the use of unambiguous physicochemical constraints to the evolutionary change inherent in biological processes, by allowing for time-varying or adjustable constraints. This approach offers an attractive alternative to mathematical modeling of systemic functions in biology. The more classical physicochemical approach to studying biological dynamics has been developed and described in several text books[14,15], and specialized theories have been developed, such as metabolic control analysis[16], for biological systems analysis.

The iterative model-building process

The process of building mathematical models of complex biological processes and their computer simulation will be an iterative one. We will begin to construct "in silico organisms" that are computer representations of their in vivo counterparts. Initial versions will be synthesized using genomic, biochemical, and physiological data. These models will have some interpretive and predictive capabilities. However, because of incomplete knowledge of constraints and erroneous annotation, these initial models will be able to represent only some functions of the organism correctly.

In carrying out this iterative model building process, we must learn to embrace failure. The main difference between the in silico and in vivo organism is that the in silico version is missing some features. Therefore, we must set out to formulate experimentally testable hypotheses based on the in silico analysis, perform the experiments, and update the models (see Fig. 4). Interestingly, this iterative process

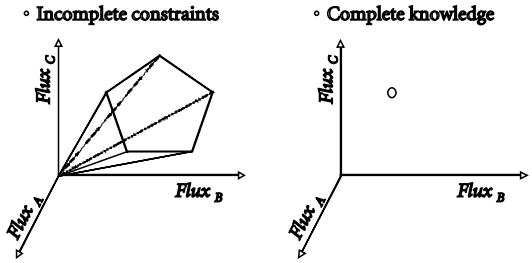

Figure 2. Constraining possible behaviors. Because biological information is incomplete, it is necessary to take into account the fact that cells are subject to certain constraints that limit their possible behaviors. By imposing these constraints in a model, one can then determine what is possible and what is not, and determine how a cell is likely to behave, but never predict its behavior precisely.

for building in silico organisms is likely to have two feedback loops. One is the classical experimental loop (the one on the right in Fig. 4), and the other is in silico (on the left in Fig. 4). Many corrections and adjustments for these models are likely to originate from analyzing and searching the ever-growing availability of bioinformatic databases.

What will we do with these in silico models? They are likely to have some basic scientific use, for purposes such as comparative genomics and evolutionary studies. The initial metabolic models will likely have practical uses associated with study of human pathogens and design and operation of industrial bioprocesses. We will move from talking about genetic engineering of single genes, to what may become known as "genome engineering," where the whole organism is the context of the design. Some early studies along these lines are appearing[6].

One additional issue is worth comment in this iterative model-building process. The high-throughput technologies are generating a "need-to-know everything" mentality. However, as experience has shown in other fields, one can construct powerful and useful computer models without "knowing everything." If we insisted on having computer models that account for every detail of a process being studied, we would not build airplanes or refineries. In fact, one of the arts of model building is to determine what is needed in order to synthesize an insightful and useful computer model. It is likely that the lessons learned from other fields will benefit model building in biology.

Simplicity from complexity

It is clear that even though the molecular composition of living cells is complex (i.e. their genotype) the number of distinct behaviors (i.e. their phenotypes) that they display is much fewer. This important principle of simplicity from complexity is emerging from singular value decomposition of gene expression data that clearly shows that many expressed gene products behave in a highly coordinated fashion[17,18]. For instance, these studies show that two principal underlying modes of motion govern the genome-wide expression pattern in yeast during its cell cycle. Studies of mathematical models of complex biochemical reaction networks exhibit similar features. Temporal decomposition of complex metabolic and growth models show that there are only a few governing dynamic determinants[19] and robustness analysis of models of complex biological processes, such as those for bacterial chemotaxis[20] and pattern formation in the drosophila embryo[21], show that their overall behavior is relatively insensitive to the exact numerical values of the kinetic constants used.

The elucidation of the underlying simplicity will rely on well established methods of system identification and model reduction that have been practiced in a number of fields of science and engineering. The approach of successive application of constraints described above similarly leads to few allowable behaviors based on a large number of interacting components. These analysis methods applied to large volumes of biological data being currently generated are likely to lead to the elucidation of the principal 'genetic circuits'[1] that underlie cell function.

What is rate limiting?

High-throughput experimental technologies are generating biological data at unprecedented rates, and the pace will only accelerate. The bioinformatic infrastructure that tabulates, curates, and makes these data retrievable is developing (e.g., WIT, EcoCyc, Mips, Kegg, Biology WorkBench, EMP, Swiss-Prot). Many initial visualization tools and statistical analysis methods, such as clustering, are becoming available for data analysis. With few exceptions, mathematical models are generally not available. Models like the human red blood cell and *Mycoplasma genitilium*[22], however, are beginning to become available in transportable format. The talent that goes into formulating these models, performing the numerical analysis, and interpreting the results is presently in short supply. So far, the available computational power for solving these models has not been a limitation.

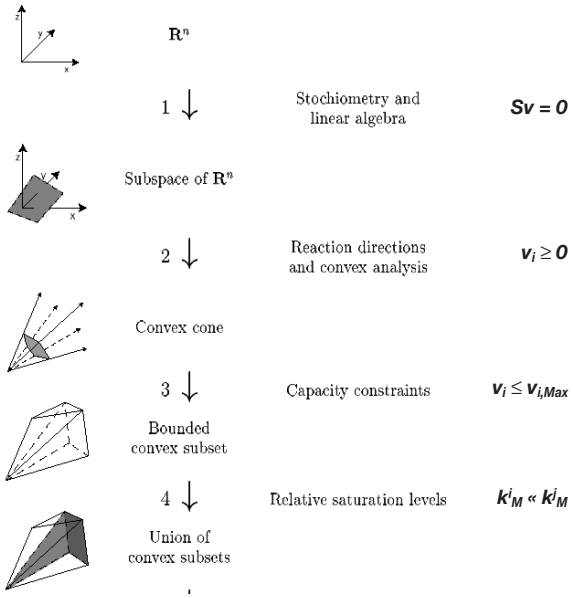

Figure 3. Narrowing down the alternatives. The application of successive constraints to a set of reactions in a pathway allows one to narrow down the attainable outcomes ("flux distributions") from a defined metabolic genotype (see text for further details).

The implications of the process depicted in Figure 1 are not just that of a major shift in scientific emphasis and outlook. The educational infrastructure in the biological sciences must respond. The biological scientist of the future will have to become more computer literate and will have to possess a higher level of mathematical and informatics training. It is likely that the major changes needed will be difficult to achieve within the structure of existing biology departments. Change in faculty orientation, research, and teaching skills that are required may not be possible with the peer review system that is currently in place.

It seems likely that new educational pro-

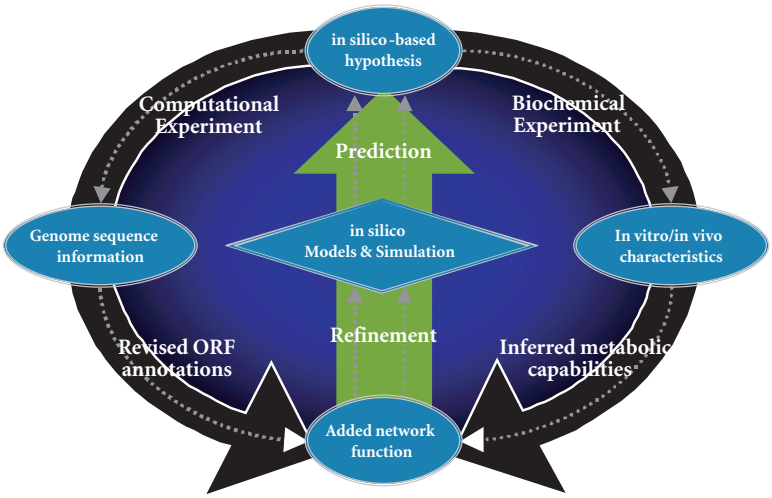

Figure 4. If at first you don't succeed. . .Iterative in silico model building in biology involves the formulation of experimentally testable hypotheses based on the in silico analysis, collection of experimental data, and subsequent refinement of the models based on these data.

grams and departments will arise. The new curricula that need to be synthesized will comprise not only computer science and biology, but also mathematical modeling, numerical analysis, and systems science. New biologically based engineering programs are likely to emerge, just as chemical engineering emerged from chemistry and mechanical engineering early in the last century.

Conclusions

High-throughput experimental technologies are not only forcing researchers to accommodate the systems point of view in cellular and molecular biology, but also enabling us to study cells as systems. Given the complexity of even the simplest cellular function, this capability is demanding the development of mathematical models and computer simulations to study the simultaneous function of multiple gene products. Such models are likely to be developed for well-studied biological model systems and organisms (e.g., *E. coli*, yeast, *Drosophila*), and will then be used to analyze, interpret, and predict the genotype–phenotype relationship. This study of phenotypes with knowledge of the genotypes can be called "phenomics[22]," which is analogous to genomics. Phenomics will have an important theoretical component through mathematical model building and computer simulation. The complexity and specific properties of biological systems, such as robustness, redundancy, and time-varying constants (evolution), are likely to make model building different from other fields of science and engineering.

Acknowledgement

I thank Markus Covert, Jeremy Edwards, David Letscher and Christophe Schilling for preparing the figures.

1. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823-826 (2000).
2. Palsson, B.O. What lies beyond bioinformatics? *Nat. Biotechnol.* **15**, 3–4 (1997).
3. Strothman, R.C. The coming Kuhnian revolution in biology. *Nat. Biotechnol.* **15**, 194–199 (1997).
4. Hartwell, L.H., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).
5. Evans, G.A. Designer science and the "omic" revolution. *Nat. Biotechnol.* **18**, 127 (2000).
6. Bailey, J.E. Lessons from metabolic engineering for functional genomics and drug discovery. *Nat. Biotechnol.* **17**, 616–618 (1999).
7. Aebersold, R, Hood, L.E., & Watts, J.D. Equipping scientists for the new biology. *Nat. Biotechnol.* **18**, 359 (2000).
8. McAdams, H.H. & Arkin, A. Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 199–224 (1998).
9. Lee, I.-D. & Palsson, B.O. A comprehensive model of human erythrocyte metabolism: extensions to include pH effects. *Biomed. Biochim. Acta* **49**, 771–789 (1991).
10. McAdams, H.H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269,** 651–656 (1995).
11. Schilling, C.H. et al. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* **15**, 296–303 (1999).
12. Varma, A. & Palsson, B.O. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* **12**, 994–998 (1994).
13. Bonarius, H.P.J., Schmid, G. & Tramper, J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.* **15**, 308–314 (1997).
14. Reich, J.G. & Sel'kov, E.E. *Energy metabolism of the cell* Edn. 2. (Academic Press, New York, NY; 1981).
15. Heinrich, R. & Schuster, S. *The regulation of cellular systems*. (Chapman & Hall, New York, 1996), p. 372.
16. Fell, D. *Understanding the control of metabolism*. (Portland Press, London, UK; 1996).
17. Alter, O., Brown, P.O, & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* **97**, 10101–10106 (2000).
18. Holter, N.S., Mitra, M., Martian, A., Cieplak, M., Banavar, J.R. & Fedoroff, N.V. Fundamental patterns underlying gene expression profiles. *PNAS* **97**, 8409–8414 (2000).
19. Palsson, B.O. Joshi, A., & Ozturk, S. Reducing complexity in metabolic networks. *Fed. Proc.* **46**, 2485-2489 (1987).
20. Alon, U., Surette, M.G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).
21. von Dassow, G., Meir, E., Munro, E.M., & Odell, G.M. The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
22. Tomita, M. et al. E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84 (1999).
23. Schilling, C.H., Edwards, J.S. & Palsson, B.O. Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.* **15**, 288–295 (1999).