Systems Biology Is Taking Off

Måns Ehrenberg,[1,5] Johan Elf,[1] Erik Aurell,[2] Rickard Sandberg,[3] and Jesper Tegnér[4]

[1]*Department of Cell and Molecular Biology, Uppsala University, BMC, 751-24 Uppsala, Sweden;* [2]*SICS, SE-14-29 Kista, Sweden;* [3]*Microbiology and Tumor Biology Center, Karolinska Institute, S-171-77 Stockholm, Sweden; and* [4]*Division of Computational Biology, Department of Physics, Linköping University, S-581-83 Linkoping, Sweden*

There is a revolution occurring in the biological sciences. It took off just a couple of years ago and is now clearly visible in the literature. Some scientists in the field like to refer to the development as the birth of systems biology, whereas others prefer not to put a label on what is happening.

Modern molecular biology was born with the discovery that genetics is based on nucleic acid chemistry (Watson and Crick 1953), and one way to define it is to say that molecular biology is a large box of tools to do genetics by manipulating DNA. This definition may sound disheartening, but its positive side is that the tools can be applied to all aspects of biology to solve essentially all scientific problems that may arise.

One result of molecular biology is large-scale sequencing of genomes from a rapidly growing number of organisms. Genome sequencing is not possible without the use of computers with large memory and tools to handle the enormous amounts of data that are generated in the massive sequencing efforts. The need for data handling led to another box of tools, called bioinformatics, which is now an established part of molecular biology. However, when all this sequence data got into computers, it became obvious that the genetic blueprints by themselves tell us very little about the functional behavior of cells and multicellular organisms; that is, about what we really want to know about biological systems. In this way, the human genome project, which is perhaps the most spectacular success of molecular biology, also meant that a vast space of future research of a radically different kind became visible. To understand the causal connections between genotype and phenotype will require a very significant expansion of the traditional toolbox used by molecular biologists. It must include concepts and techniques from many other scientific disciplines such as physics, mathematics, numerical analysis, stochastic processes, and control theory. Many novel tools that do not exist today must be forged to understand how dynamic, adapting, and developing systems can emerge from the information buried in the genomes.

The development of such an extended toolbox for quantitative reasoning about the dynamics of living systems, and the application of its contents to solve a variety of scientific problems, is one way to define systems biology, analogous to our definition of molecular biology above. It is our belief that systems biology will enrich the biological sciences and transform our thinking about biological problems, in analogy with what has been happening in molecular biology during the 50 years that have passed since the discovery of the double helix. Systems biology will always bring the functional aspects into focus, sometimes close to genomics and sometimes far out in areas not visited before. Below will follow some examples of what we consider significant developments of systems biology, which is still in its infancy but has great future promise. The selection of topics is limited by the format of this mini-review, and many important contributions could therefore not be covered.

[5]Corresponding author.
E-MAIL ehrenberg@xray.bmc.uu.se; FAX 46184714262
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1763203.

Models of Growing Bacteria

One question that has been with us for a long time is whether bacteria have evolved to maximize their growth rate (Ehrenberg and Kurland 1984), and in its simplest form, it refers to logarithmically growing cells in media containing different types of carbon sources. It is now possible to use genomic information to reconstruct the metabolic pathways in whole cells (Edwards et al. 2001; Forster et al. 2003). The freedom of choices of pathways can be reduced by in silico models of metabolic flows, in which physical and other constraints are introduced to decrease the number of degrees of freedom. Interestingly, the growth rate of bacterial cells is accounted for in such in silico models and can be maximized by simple linear optimization techniques (Ibarra et al. 2002) and then compared with the growth rate of real bacteria. It turns out the *Escherichia coli* bacteria grow with maximal rate according to the prediction of its in silico model in media containing a number of common carbon sources. In contrast, when the carbon source is glycerol, the growth is much slower than that in the optimized in silico model. When, however, the cells are allowed to grow in the presence of glycerol for a large number of generations, they reproducibly evolve toward a maximal growth rate, as predicted by the optimized in silico model. In this case, the step from the annotated genome sequence of *E. coli* to predictions about how fast the organism can grow is amazingly short. Because rapid growth correlates positively with the population genetic fitness parameter, it is now possible to make in silico simulations of the evolution of bacteria that lead to predictions that are experimentally testable.

Another case, in which an analysis of metabolic flows in growing bacteria leads to a number of interesting testable predictions, concerns amino acid limitation. In all organisms, many amino acids are encoded by several synonymous code words (Crick et al. 1961), and these are often read by several tRNA isoacceptors (Björk 1996). When such an amino acid becomes rate-limiting for protein synthesis in the cell, the charged levels of the members of a family of isaoaccepting tRNAs will react very differently to the deficiency. One tRNA, or a few tRNAs, will totally lose its charging with amino acid, whereas others can remain almost fully charged even when the rate of supply of the amino acid becomes negligible (Elf et al. 2003). This result, following from a very simple model, has been used to rationalize the choice of synonymous codons in control systems for transcriptional regulation (attenuation of transcription; Yanofsky 1981), as well as in genes encoding enzymes that synthesize amino acids. These results also indicate that the codon adaption index (CAI) for highly expressed genes (Sharp and Li 1987) should be complemented with a starvation CAI (sCAI) for gene expression during amino acid limitation.

Functional Motifs in Transcriptional Networks

Genome sequences can also be used to identify common motifs in networks for transcriptional regulation. Searches for such motifs in *E. coli* (Shen-Orr et al. 2002) and *Saccharomyces cerevisiae* (Lee et al. 2002; Maslov and Sneppen 2002) have revealed a number of recurrent motifs, and for some, their functional role can be

inferred. One example is a feed-forward coherent loop, and kinetic analysis reveals that this motif produces a significant output signal when the input signal lasts a sufficiently long time but not when the input is a short transient. Identification of such motifs that occur with frequencies much higher than in random networks, and understanding their functions, will become an important tool for identifying ubiquitous control units in cells.

Unifying Principles in Biological and Engineered Systems

Another interesting development in systems biology is the recognition that the way organisms are designed may be very similar to how man-made machines are constructed, as the result of many generations of trials and errors (Csete and Doyle 2002). The insight that robustness, that is, the stability of system performance to perturbations in machine parameters or to variation in external conditions, is important not only for engineered objects but also for organisms has proved particularly fruitful.

One example regards bacterial chemotaxis, in which Barkai and Leibler (1997) suggested a simple and yet apparently realistic feedback system that allows the cells to retain full sensitivity to relative gradients in repellents or attractants, although the absolute concentrations of these compounds may vary by orders of magnitude. They also showed that this property of their modelled chemotactic control system was robust to any changes in enzyme concentrations and other intracellular parameters. They suggested that robustness is expected to be a universal design principle among many intracellular control systems.

This insight was applied to attempts at modeling the embryonic pattern formation in *Drosophila melanogaster* (Eldar et al. 2002). A huge set of theoretical models, all compatible with existing experimental data on morphogen gradient patterns in the *Drosophila* dorsal region, were characterized with respect to robustness. The analysis revealed that only a small subset of models were robust, and that all these robust models shared the same molecular mechanism. The existence of this particular mechanism in *Drosophila* was subsequently verified by experiments.

This result corroborates the suggestion that robustness is a universal organizing principle in the design of organisms, and shows how a criterion of robustness can be used to effectively eliminate false models. Even the discovery of just a few such universal principles could be the starting point for a systems biology that reaches far beyond the modeling of particular instances.

Engineering of Genetic Networks

The engineering aspect of systems biology is further emphasized by a number of recent attempts at building gene circuits with desired properties, just as one builds electronic circuits for various purposes.

These engineered gene circuits have been used to test mathematically formulated hypotheses about dynamics and regulation of small genetic networks. The engineered systems are plasmid-born or chromosomally integrated by using homologous recombination (Court et al. 2002). One rationale for studying engineered, rather than natural, systems is to prevent the complexity of an unknown cell physiology to obscure the outcome of experiments. Understanding properties of the components (promoters, repressors, reporters, etc.) and their interaction in the engineered systems will give clues to design principles used also in natural systems. Engineered gene circuits have been designed to toggle between stationary states (Gardner et al. 2000), to oscillate (Elowitz and Leibler 2000), or to behave as logical gates (Guet et al. 2002). Starting from the principle of regulated recruitment of RNA polymerase to different promoters (Ptashne and Gann 2001), Buchler et al. (2003) have developed quantitative models, which can be used to guide the design of small *cis*-regulatory networks.

In the study by Guet et al. (2002), combinatorial design using three well-characterized transcriptional regulators and five of their promoters made it possible to construct a large variety of three-gene networks. In response to different combinations of the inducers of the regulators, the network implemented different logical gates. An additional interesting but subversive observation in this study was that networks with the same connectivity, that is, expected to behave similarly, could display different response patterns. This means that even in very simple logical circuits that are carefully designed, the phenotype can often not be predicted, emphasizing the formidable problems that lie ahead to understand naturally occurring global intracellular networks. Kim and Tidor (2003) have used mathematical modeling to explain how networks with apparently identical architecture, as used by Guet et al. (2002), can implement different logical functions.

The Stochastic Nature of Intracellular Networks

The stochastic nature of all chemical reactions necessarily leads to random fluctuations in intracellular molecule copy numbers. Careful characterization of fluctuations in biological systems is often required to understand their modes of operation. This is intuitively obvious when the molecule copy numbers are low, as for messenger RNAs in gene expression (Berg 1978). Artificial systems for gene expression have been engineered with the purpose of studying how chemical noise affects their performance (Elowitz et al. 2002; Ozbudak et al. 2002; Blake et al. 2003) and how noise can be attenuated by feedback inhibition (Becskei and Serrano 2000). It is often assumed that the expected size of fluctuations (i.e., the standard deviation from the mean) is equal to the square root of the number of molecules of a particular kind in the system (Schrödinger 1944). However, this intuition is based on equilibrium thermodynamics (van Kampen 1992), and because chemical reactions in the living cell often occur far from equilibrium and frequently have strongly nonlinear kinetics, fluctuations can both be much larger and much smaller than the square root of the average number of molecules. A rapid method to estimate the size of fluctuations in intracellular networks of any kind has been presented by Elf and Ehrenberg (2003).

The power of mathematical modeling that takes stochastic aspects into account was demonstrated by the analysis of cell cycle control in fission yeast (Sveiczer et al. 2001) and studies of the control of plasmid copy number in bacteria (Paulsson and Ehrenberg 2001). In the first of these studies, the molecular basis of cell size control was identified, and it was argued that the cell-cycle time is regulated by the concentration, rather than by the amount, of the activator (Cdc2/cyclin) in the cytoplasm. Noise in gene expression affects the stability and robustness of states of stable gene expression, and quantitative descriptions of this phenomenon have been made for the paradigmatic example of λ phage lysogeny (Aurell and Sneppen 2002).

Massively Parallel Experiments Reveal the Organization of Genetic Networks

It is now possible to follow changes in the expression from a vast number of genes in an organism at the transcriptional level with microarray techniques (Lander 1999), and at the protein level with two-dimensional gels or mass spectrometry with isotope labeling (Gygi et al. 1999). Changes in metabolite pools can also be followed in a massively parallel fashion (Raamsdonk et al. 2001), and with the increasing knowledge of protein–protein interactions (Uetz et al. 2000), global analysis of genetic intracellular networks is becoming feasible.

One example of the power of these new experimental techniques is work on *Caulobacter crescentus* (Laub et al. 2000), in which microarray analysis identified differential transcriptional regulation as a crucial mechanism for cell-cycle control and revealed a hierarchical pattern of control in which a few proteins steer the transcriptional pattern of a large number of genes downstream in the cell cycle. In this case, the microarray observations fell into the fertile soil of extensive genetic work over many years and an in-depth knowledge of the physiology of this organism.

Another example comes from work by Ideker et al. (2001). They used microarrays and mass spectrometry to analyze the responses of messenger RNA and protein levels, respectively, to systematic perturbations of expression from the genes in the galactose utilization pathway in the yeast *S. cerevisiae*.

Pitfalls of Reverse Engineering of Genetic Networks

Although massively parallel experimental detection of transcripts, proteins, and metabolites in combination with knowledge of protein–protein interactions allow effective discrimination against erroneous models of biological networks, these methods do not generally allow unique reconstructions of existing networks through reverse engineering. Statistical techniques, such as cluster analysis, have been used to demonstrate that currently available experimental data on gene expression levels are generally not enough to reconstruct the network structure of pathways and regulatory networks. Major challenges for reverse engineering are therefore to develop optimal strategies for the design of perturbation experiments (Tegner et al. 2003; Zak et al. 2003), to devise statistical methods to estimate the confidence of model choices and parameter estimates (Moles et al. 2003), and to put models obtained by reverse engineering to critical tests with data from living systems.

An alternative approach is to integrate different types of data, for example, expression and sequence data. Binding sites of transcription factor can be recognized in sequence signatures (see Djordjevic et al. 2003), and their differential number counts, upstream of proximate genes, give more information on which to base reconstruction efforts. Work in this direction has been done in yeast (Bussemaker et al. 2001). Very likely, we will see more large-scale data fusion of this kind. Successful integration of data from multiple sources (e.g., protein–protein, protein–DNA, sequence, and metabolite data) requires the development of new optimized software for data management and fusion. One example of such a development has been given by Shannon et al. (2003); another approach to the same problem, specifically aimed at T-cell maturation, by Elfroni et al. (2003).

Conclusions

Systems biology is here to stay. It has left its lag phase behind, and we are now witnessing how a new scientific discipline allows an ever-increasing number of biological problems to be approached and solved with new techniques and theoretical concepts that unify hitherto separate areas of biology. The methods and concepts of systems biology will not only expand into all areas of the biological sciences; its results are bound to have repercussions in and inspire other sciences such as physics, engineering, mathematics, and social sciences.

REFERENCES

Aurell, E. and Sneppen, K. 2002. Epigenetics as a first exit problem. *Phys. Rev. Lett.* **88:** 048101.

Barkai, N. and Leibler, S. 1997. Robustness in simple biochemical networks. *Nature* **387:** 913–917.

Becskei, A. and Serrano, L. 2000. Engineering stability in gene networks by autoregulation. *Nature* **405:** 590–593.

Berg, O.G. 1978. A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **71:** 587–603.

Björk, G. 1996. *Stabile RNA modification in* Escherichia coli *and* Salmonella typhimurium: *Cellular and molecular biology* (eds. F. Neidhardt et al.), pp. 861–886. ASM Press, Washington, DC.

Blake, W.J., Kaern, M., Cantor, C.R., and Collins, J.J. 2003. Noise in eukaryotic gene expression. *Nature* **422:** 633–637.

Buchler, N.E., Gerland, U., Uwa, T. 2003. On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci.* **100:** 5136–5141.

Bussemaker, H.J., Li, H., and Siggia, E.D. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* **27:** 167–171.

Court, D.L., Sawitzke, J.A., and Thomason, L.C. 2002. Genetic engineering using homologous recombination. *Annu. Rev. Genet.* **36:** 361–388.

Crick, F., Barnett, L., Brenner, S., and Watts-Tobin, R. 1961. General nature of the genetic code for proteins. *Nature* **192:** 1227.

Csete, M.E. and Doyle, J.C. 2002. Reverse engineering of biological complexity. *Science* **295:** 1664–1669.

Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res.* (this issue).

Edwards, J.S., Ibarra, R.U., and Palsson, B.O. 2001. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19:** 125–130.

Ehrenberg, M. and Kurland, C.G. 1984. Costs of accuracy determined by a maximal growth rate constraint. *Q. Rev. Biophys.* **17:** 45–82.

Eldar, A., Dorfman, R., Weiss, D., Ashe, H., Shilo, B.Z., and Barkai, N. 2002. Robustness of the BMP morphogen gradient in *Drosophila* embryonic patterning. *Nature* **419:** 304–308.

Elf, J. and Ehrenberg, M. 2003. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.* (this issue).

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300:** 1718–1722.

Elfroni, S., Harel, D., and Cohen, I.R. 2003. Toward rigorous comprehension of biological complexity: Modeling, execution and visualization of thymic T-cell maturation. *Genome Res.* (this issue).

Elowitz, M.B. and Leibler, S. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature* **403:** 335–338.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. 2002. Stochastic gene expression in a single cell. *Science* **297:** 1183–1186.

Forster, J., Famili, I., Fu, P., Palsson, B.O., and Nielsen, J. 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13:** 244–253.

Gardner, T.S., Cantor, C.R., and Collins, J.J. 2000. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403:** 339–342.

Guet, C.C., Elowitz, M.B., Hsing, W., and Leibler, S. 2002. Combinatorial synthesis of genetic networks. *Science* **296:** 1466–1470.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17:** 994–999.

Ibarra, R.U., Edwards, J.S., and Palsson, B.O. 2002. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420:** 186–189.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292:** 929–934.

Kim, P.M. and Tidor, B. 2003. The topology and behavior of genetic networks: A mathematical treatment. *Genome Res.* (this issue).

Lander, E.S. 1999. Array of hope. *Nat. Genet.* **21:** 3–4.

Laub, M.T., McAdams, H.H., Feldblyum, T., Fraser, C.M., and Shapiro, L. 2000. Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **290:** 2144–2148.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.

Maslov, S. and Sneppen, K. 2002. Specificity and stability in topology of protein networks. *Science* **296:** 910–913.

Moles, C.G., Mendes, P., and Banga, J.R. 2003. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Res.* (this issue).

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. 2002. Regulation of noise in the expression of a single gene. *Nat. Genet.* **31:** 69–73.

Paulsson, J. and Ehrenberg, M. 2001. Noise in a minimal regulatory network: Plasmid copy number control. *Q. Rev. Biophys.* **34:** 1–59.

Ptashne, M. and Gann, A. 2001. *Genes and signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., et al. 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19:** 45–50.

Schrödinger, E. 1944. *What is life?* Cambridge University Press, Cambridge, UK.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* (this issue).

Sharp, P.M. and Li, W.H. 1987. The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acid Res.* **15:** 1281–1295.

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31:** 64–68.

Sveiczer, A., Tyson, J.J., and Novak, B. 2001. A stochastic, molecular model of the fission yeast cell cycle: Role of the nucleocytoplasmic ratio in cycle time regulation. *Biophys. Chem.* **92:** 1–15.

Tegner, J., Yeung, M.K., Hasty, J., and Collins, J.J. 2003. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci.* **100:** 5944–5949.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.

van Kampen, N.G. 1992. *Stochastic processes in physics and chemistry*. Elsevier Science, Amsterdam, The Netherlands.

Watson, J.D. and Crick, F.H. 1953. A structure for deoxyribose nucleic acid. *Nature* **171:** 737–738.

Yanofsky, C. 1981. Attenuation in the control of expression of bacterial operons. *Nature* **289:** 751–758.

Zak, D., Gonye, G., Schwaber, J., and Doyle, F.J. 2003. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network. *Genome Res.* (this issue).