advance from current multibillion transistor chips to the multitrillion transistor range of terascale integration.

References

1. J. D. Meindl, *Proc. IEEE* **83** (no. 4), 619 (1995).
2. R. W. Keyes, *Proc. IEEE* **89** (no. 3), 227 (2001).
3. J. D. Meindl, J. A. Davis, *IEEE J. Solid State Circuits* **35** (no. 10), 1515 (2000).
4. J. von Neumann, *Theory of Self-Reproducing Automata* (Univ. of Illinois Press, Urbana, IL, 1966), p. 66.
5. R. W. Keyes, *Proc. IBM J. Res. Dev.* **5**, 183 (1961).
6. R. M. Swanson, J. D. Meindl, *IEEE J. Solid State Circuits* **SC-7**, 1146 (1972).
7. C. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).
8. J. D. Meindl, *Proc. IEEE* **89** (no. 3), 223 (2001).
9. H. Haken, H. C. Wolf, *Atomic and Quantum Physics* (Springer-Verlag, Berlin, 1984), chap. 7.
10. J. D. Plummer, *Proc. IEEE* **89** (no. 3), 240 (2001).
11. J. B. Roldan, F. Gamiz, J. A. Lopez-Vallanueva, J. E. Carceller, P. Cartujo, *Semicond. Sci. Technol.* **12**, 321 (1997).
12. D. J. Frank *et al.*, *Proc. IEEE* **89** (no. 3), 259 (2001).
13. Q. Chen, private communication.
14. X. Tang, V. K. De, L. Wang, J. D. Meindl, *Proc. IEEE Int. SOI Conf.* **1999**, 42 (1999).
15. J. Davis, J. D. Meindl, *IEEE Trans. Electron Devices* **47**, 2068 (2000).
16. _____, *IEEE Trans. Electron Devices* **47**, 2078 (2000).
17. K. A. Bowman, X. Tang, J. C. Eble, J. D. Meindl, *IEEE Trans. Electron Devices* **45** (no. 3), 580 (1998).
18. J. A. Davis, V. K. De, J. D. Meindl, *IEEE Trans. Electron Devices* **45** (no. 3), 590 (1998).
19. A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, J. D. Meindl, *IEEE Trans. Very Large Scale Integration (VLSI) Syst.* **8**, 235 (2000).
20. *The International Technology Roadmap for Semiconductors (ITRS)* (Semiconductor Industry Association, San Jose, CA, 1999).

VIEWPOINT

Blazing Pathways Through Genetic Mountains

David K. Gifford

It is now widely accepted that high-throughput data sources will shed essential understanding on the inner workings of cellular and organism function. One key challenge is to distill the results of such experiments into an interpretable computational form that will be the basis of a predictive model. A predictive model represents the gold standard in understanding a biological system and will permit us to investigate the underlying cause of diseases and help us to develop therapeutics. Here I explore how discoveries can be based on high-throughput data sources and discuss how independent discoveries can be assembled into a comprehensive picture of cellular function.

To date, most discoveries that have been based on expression data have relied on data visualization. For example, in this issue, Kim *et al.* describe the first large compendium of *Caenorhabditis elegans* expression data (*1*). The 533 microarray experiments discussed characterize the transcriptome of *C. elegans* cells in a wide variety of growth conditions, developmental stages, and genetic backgrounds. The coexpression of genes in these experiments gives important information about potential gene coregulation and the functions of previously uncharacterized genes in *C. elegans*. Thus, these data will be an important basis for further research in the *C. elegans* community.

Kim *et al.* visualize the *C. elegans* expression data in three dimensions for analysis. Groups of related genes in this three-dimensional approach appear as mountains, and the entire transcriptome appears as a mountain range. Distances in this synthetic geography are related to gene similarity, and mountain heights are related to the density of observed genes in a similar location. A three-dimensional approach is a departure from the common practice of analyzing expression data in a single dimension. Single-dimension analysis places genes in a total ordering, limiting our ability to see important relationships.

Department of Computer Science, Massachusetts Institute of Technology, 200 Technology Square, Cambridge, MA 02139, USA. E-mail: gifford@mit.edu

Visualization-based approaches are an important first step toward understanding cellular function. Expression visualization allows us to hypothesize potential gene-gene relationships that can be experimentally tested. For example, when a visualization tool shows that genes are coexpressed, it is natural to search for transcriptional activators that are shared between the genes. The results of such searches are typically expressed in schematic form, with the schematics depicting how genes influence one another's expression and activity. Often posttranslational modifications of proteins play a large role in their activities, and these modifications must also be captured in a schematic diagram to accurately predict the behavior of a system.

The individual elements of understanding that grow out of visualization and subsequent experiments can be naturally organized into a model-based approach to discovery. Model-based approaches codify our understanding of the underlying causes of data variation that is observed in data visualization, and the integration of results into a system model is necessary for broad understanding and insight. In a model-based approach, competing models that describe a function are constructed, and the models are scored against experimental data. The score of a model describes the likelihood of observing the experimental data given the model under consideration. Thus, models provide a principled way of judging the relative likelihood of competing hypotheses. When many models have roughly the same score, it is possible to determine the features that they share in common. The shared features of high-scoring models represent biological relationships that are likely to be important.

Despite the extraordinary discriminatory benefits of models, many biologists retreat from this approach with concerns about complex differential equations, unintelligible computer commands, and a feeling of unease that researchers will not be able to grasp the subtleties of what the models are saying. Furthermore, many model-based approaches require the values of reaction parameters that we do not yet know and that are difficult to approximate from contemporary high-throughput data sources. New approaches to modeling that are intuitive, can capture high-level structure, and are parameter-free would overcome these problems and motivate more biologists to capture and analyze in computational form what they suspect to be true.

Structured computational models, and in particular graphical models, have recently been proposed as a parameter-free approach for modeling biological network structure (*2*, *3*). Just like the schematic diagrams familiar to biologists, a graphical model captures the qualitative relationships between variables. Vertices in a graphical model represent variables such as mRNA expression levels, protein levels, environmental conditions, genotype, and phenotype. Edges in a graphical model describe relationships between variables and can be annotated with typical biological semantics, such as enhances or represses.

Once constructed, a graphical model represents both a conceptual understanding of a biological system and a computational means for predicting the effects of pertubations to the system. For example, Fig. 1 illustrates how a graphical model can explain data in a form that is simpler and more easily interpretable compared with conventional clustering di-

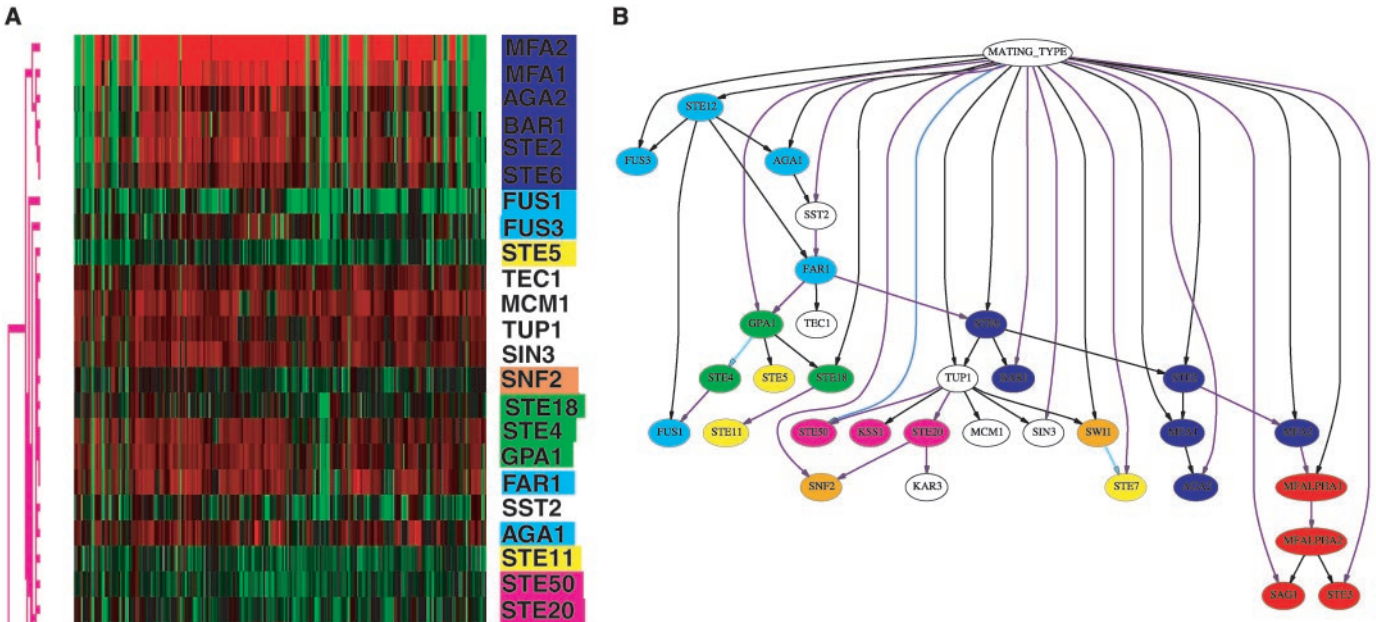

Fig. 1. Expression data can be visualized directly and as genetic regulatory networks. (**A**) shows the hierarchical clustering of 32 genes in 320 *S. cerevisiae* expression experiments (240 shown) and (**B**) shows how the data can be used to automatically reconstruct a tentative genetic regulatory network with graphical models. Genes expressed only in MATa cells are colored dark blue (MFA1, MFA2, STE2, STE6, AGA2, and BAR1); genes expressed only in MATα cells are colored red (MFALPHA1, MFALPHA2, STE3, and SAG1); genes whose promoters are bound by Ste12 are colored cyan (STE12, FAR1, AGA1, FUS1, and FUS3); genes coding for components of the heterotrimeric G-protein complex are colored bright green (GPA1, STE4, and STE18); genes coding for core components of the primary signaling cascade complex are colored yellow (STE5, STE7, and STE11); genes coding for auxiliary or alternate components of the signaling cascade are colored magenta (STE50, STE20, and KSS1); and genes whose protein products form part of the SWI-SNF complex are colored orange (SWI1 and SNF2). Full details on the interpretation of the graphical model can be found in (*5*).

agrams. Computers can be used to compare graphical models from disparate systems to find preserved structure, and to query large databases for models that contain specified components of interest to an investigator. Therefore, graphical models are a natural computational respository for our high-level knowledge about biological systems.

Because graphical models are inherently probabilistic, they can represent areas of imperfect knowledge and complete understanding within the same model. This ability is crucial for our ability to evolve models as our understanding of a system improves. Imperfect knowledge can be represented as simply as "I think these two genes interact somehow," whereas complete understanding can be represented as a precise specification of the transfer function between the genes.

Graphical models are simple, yet they can be scored against noisy high-throughput experimental data. Such scores allow one to precisely judge alternative high-level hypotheses about the structure of a given genetic regulatory network. One nice aspect of graphical models is that they are general enough to permit a wide variety of high-throughput data sources to be fused to provide a focused picture of cellular function. For example, data from location, expression, and proteomic analysis can be combined in a multiple expert approach that overcomes some of the systemic difficulties of relying on a single type of data. Furthermore, the value of additional data can be judged in the context of graphical models, thus allowing experimental work to provide highly informative data given what is already known.

When multiple data sources are fused, it is possible to automate the search for high-scoring models. Millions of alternative model structures can be systematically generated and scored, and the features that are preserved among the best scoring models can be presented. Reverse-engineering genetic regulatory networks in this fashion require highly informative data about the system being discovered in order to avoid model overfitting. Overfitting occurs when a model is sufficiently complex to explain limited data by chance. Experimental design can also be aided by a modeling system because in many instances a model can be used to calculate the marginal value of specific data.

The trade for the simplicity of graphical models is their inability to model fine-grained dynamic behavior. For the faithful detailed replication of cellular behavior, other techniques, such as dynamic stochastic simulation (*4*), will be needed. Dynamic stochastic simulation permits additional knowledge to be incorporated and tested, such as binding constants and reaction rates.

Elucidating complete genetic regulatory networks will entail an immense amount of biological discovery that we expect will dwarf the human genome project in magnitude. Genetic regulatory networks are responsible for cellular control and the development of multicellular organisms and are the foot soldiers in the complex processes involved in neurobiology. Thus, a sizable fraction of the secrets of biology will be uncovered once we have built robust descriptions of the genetic regulatory networks that underlie cellular behavior.

Thus, a challenge that lies before us in postgenome biology is organizing our efforts to discover the genetic regulatory networks of key model systems. The amount of data required will be so vast that it would be unimaginable for a single investigator to produce it all. Unfortunately, unlike genome sequencing efforts, the ability to share information in the modeling context is in an embryonic stage. In the world of genome sequencing, one can physically separate the

Table 1. Parallels between genome sequencing and genetic network discovery.

Genome sequencing	Genome semantics
Physical maps	Graphical model
Contigs	Low-level functional models
Contig reassembly	Module assembly
Finished genome sequence	Comprehensive model

DNA to be sequences into distinct pieces, parcel out the detailed work of sequencing, and then reassemble these independent efforts at the end. It is not quite so simple in the world of genome semantics.

Despite the differences between genome sequencing and genetic network discovery, there are clear parallels that are illustrated in Table 1. In genome sequencing, a physical map is useful to provide scaffolding for assembling the finished sequence. In the case of a genetic regulatory network, a graphical model can play the same role. A graphical model can represent a high-level view of interconnectivity and help isolate modules that can be studied independently. Like contigs in a genomic sequencing project, low-level functional models can explore the detailed behavior of a module of genes in a manner that is consistent with the higher level graphical model of the system. With standardized nomenclature and compatible modeling techniques, independent functional models can be assembled into a complete model of the cell under study.

To enable this process, there will need to be standardized forms for model representation. At present, there are many different modeling technologies in use, and although models can be easily placed into a database, they are not useful out of the context of their specific modeling package. The need for a standardized way of communicating computational descriptions of biological systems extends to the literature. Entire conferences have been established to explore ways of mining the biology literature to extract semantic information in computational form.

Going forward, as a community we need to come to consensus on how to represent what we know about biology in computational form as well as in words. The key to postgenomic biology will be the computational assembly of our collective knowledge into a cohesive picture of cellular and organism function. With such a comprehensive model, we will be able to explore new types of conservation between organisms and make great strides toward new therapeutics that function on well-characterized pathways.

References

1. S. K. Kim *et al.*, *Science* **293**, 2087 (2001).
2. A. Hartemink *et al.*, paper presented at the Pacific Symposium on Biocomputing 2000, Oahu, Hawaii, 4 to 9 January 2000.
3. D. Pe'er *et al.*, paper presented at the 9th Conference on Intelligent Systems in Molecular Biology (ISMB), Copenhagen, Denmark, 21 to 25 July 2001.
4. H. McAdams, A. Arkin, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 814 (1997).
5. A. J. Hartemink, thesis, Massachusetts Institute of Technology, Cambridge (2001).

VIEWPOINT

Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

Machine learning (ML) (*1*) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business, and science. A generalized version of the standard Hidden Markov Models of ML practice have been used for ab initio prediction of gene structures in genomic DNA (*2*). The predictions correlate surprisingly well with subsequent gene expression analysis (*3*). Postgenomic biology prominently features large-scale gene expression data analyzed by clustering methods (*4*), a standard topic in unsupervised learning. Many other examples can be given of learning and pattern recognition applications in science. Where will this trend lead? We believe it will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, ML has the potential to amplify every aspect of a working scientist's progress to understanding. It will also, for better or worse, endow intelligent computer systems with some of the general analytic power of scientific thinking.

Machine Learning Systems Group, Jet Propulsion Laboratory/California Institute of Technology, Pasadena, CA, 91109, USA.

*To whom correspondence should be addressed. E-mail: mjolsness@jpl.nasa.gov

Machine Learning at Every Stage of the Scientific Process

Each scientific field has its own version of the scientific process. But the cycle of observing, creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semiautomatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern recognition. A recent example is event analysis for Cherenkov detectors (*8*) used in neutrino oscillation experiments. Microscope imagery in cell biology, pathology, petrology, and other fields has led to image-processing specialties. So has remote sensing from Earth-observing satellites, such as the newly operational Terra spacecraft with its ASTER (a multispectral thermal radiometer), MISR (multiangle imaging spectral radiometer), MODIS (imaging