

# Challenges for Intelligent Systems in Biology

Russ B. Altman, *Stanford University*

**B**iological processes have produced the ultimate intelligent system (us), and now we are trying to understand biology (and ourselves) by building intelligent systems.

Intelligent systems research in biology strives to understand how living systems perform difficult tasks routinely (ranging from molecular phenomena such as protein-folding to organism-level phenomena such as cognition).

The definition of intelligent systems in biology can lead to hours of debate. Some—the lumpers—say that all high-performance systems that do something difficult with (or to) biological data should be considered intelligent systems. Others—the splitters—insist that the term “intelligent system” should be reserved for systems using the methods typically associated with modern AI. For this article, I will be a lumper. However, some systems are clearly more intelligent than others.

## An emphasis on molecular biology

Biology is the study of living systems and how they work. Using intelligent systems to understand biology can be applied across many scales, from the atomic details of biological molecules to the interactions of species in an ecosystem. The areas that have received the most attention, however, are those where the data glut is most evident, and methods are needed immediately to manage this information. DNA sequencing technologies were the first to produce large amounts of data, and they provided the founding impetus for bioinformatics. Figure 1 in the Guest Editor’s Introduction on page 9 shows the number of DNA bases in Genbank, the major DNA database, over the last 20 years. The human genome contains approximately 3 billion DNA bases, and a rough draft of this sequence is now available.<sup>1,2</sup>

More recently, biologists have developed other high-throughput experimental methods that produce large amounts of data. These include methods for measuring the expression of all genes within a population of cells simultaneously and quantitatively (using DNA microarrays), rapidly assessing the ability of biological molecules to

interact with one another (using yeast-two hybrid), quickly identifying the compounds present in a mixture of biological molecules (using mass spectroscopy), and determining the detailed 3D structure of biological molecules (using x-ray crystallography and nuclear magnetic resonance [NMR] spectroscopy). If you collect a lot of data, the intelligent systems will come.

Certain key features of biological data make intelligent systems critical for their analysis.

- Biological data is normally collected with a relatively low signal-to-noise ratio. This creates a need for robust analysis methods.
- Biology’s theoretical basis is still in its infancy, so few “first principle” approaches have any chance of working yet. This creates a need for statistical and probabilistic models.
- Despite the wealth of biological data, biology is still relatively knowledge rich and data poor. We know more about biology in a qualitative sense than a quantitative one. This creates a need for complex knowledge representations.
- Biology (and its associated data sources) operate at multiple scales that are tightly linked. This creates a need for cross-scale data integration methods.
- Biological research efforts are distributed, and the associated databases focus on particular types of data. This creates a need for data integration methods.
- Biologists think graphically about their work. This creates a need for user interfaces and graphical metaphors for communicating information.

## DNA sequence, the master plan

DNA sequence analysis is a primary magnet for attracting computer professionals into biology, because DNA sequencing is digital and compatible with decades of work on the string algorithms. However, DNA sequencing is a false magnet, because it is perhaps the only digital information in biology, and some computer scientists feel betrayed when they realize how fuzzy the rest of biology is. Nonetheless, chal-

lenges in the analysis of DNA sequencing abound; it is still amazing that a linear sequence of four characters (ATGC) is sufficient to specify (in the context of some egg and sperm “initial conditions”) a living process.

How can we parse a genome to find the segments of DNA sequence with various biological roles: encoding proteins and RNA, and controlling when and where those molecules are expressed? The best methods rely on work from speech processing and various forms of hidden, semi-, and high-order Markov models.<sup>3</sup> How can we align the sequences in DNA sequences to examine what is the same and what is different across the samples? We can ask these questions at a fine level of detail (aligning individual fragments of the genomes) or broadly across entire genomes. Alignment methods are critically important in finding difficult-to-recognize signals in the sequences that only emerge in multiple examples. They are used to compare genomes (and thus, for example, to understand why things work differently or the same in mice and humans) or to compare the control regions of individual genes within a genome (to understand why sets of genes are turned on and off in a concerted fashion). These techniques are usually based on statistical-sampling theory or substrating analysis.<sup>4</sup>

### RNA, the message

RNA has a number of biological functions, but its primary function is to be the working copy of the gene (made directly from the DNA) that is then used to synthesize proteins. The emergence of techniques for measuring the level of RNA expression for each gene in a population of cells promises a high-resolution understanding of the coordinated expression of genes over time, as an organism develops and responds to its environment.

Biologists are excited about the potential uses of RNA expression data. Current challenges include the development of methods to cluster genes on the basis of common patterns of expression in the cell, to classify genes on the basis of supervised machine learning techniques (when a subset of genes has a known class), and to reconstruct genetic interactions by trying to identify coregulated genes that are controlled by common “master” genes.<sup>5</sup>

The challenge for interpreting this data is to find reliable gold standards against which to measure new methods. Clustering can produce groupings that are reminiscent

of known associations, but this new data is so comprehensive that it is not clear how to further validate clusters. So, there is a fair amount of work in developing internal and external measures of consistency for the clusterings.

Similarly, the task of recreating the control pathways for turning genes on and off suffers from the lack of gold standards, as well as from a relative paucity of data. In principle, not only can each gene affect every other gene, but groups of genes can combine in nonlinear ways to affect other genes. As investigators attempt to model the relationships between genes using Boolean networks, linear models, and nonlinear models, many of these models have simply too many parameters (compared to the number of data points) to adequately constrain the problem. Thus, the relative data abundance can be misleading, and requires the use of other sources of data and knowledge to let us distinguish competing gene control models.

The final challenge in messenger RNA (mRNA) expression analysis is combining these data with other data sources, including the published literature, the sequence and structure databases, and so on. Single data sources are most useful in the context of other, relatively orthogonal sources of data, where the noise in one dataset offsets the signal in another. The most exciting work in this field, therefore, is often combining expression data with other data sources to draw new inferences.

### Protein, the effector

The expressed mRNA is brought to the ribosome, where the genetic code is used to read off the sequence of amino acids that create a protein.<sup>6</sup> The linear string of amino acids then folds (reliably and reproducibly) into the 3D protein structure that then can manifest many biological functions. Protein structures are responsible for enzymatic catalysis, structural support, motion, signal transduction of physical signals (light), cell-to-cell communication, and many other functions in the organism. An understanding of a protein’s 3D structure often yields valuable insight about the mechanism and details of its function. However, the availability of 3D structures has trailed behind the availability of DNA sequences (and RNA expression) because of the great expense (and chance) involved in experimentally determining 3D structure.

So, a second major magnet for computational biology in the last 30 years has been the Holy Grail of predicting 3D structure from 1D amino acid sequences. Although this challenge remains with significant progress recently—using knowledge-based approaches combining physical principles with information from the set of known 3D structures—it is possible that it will be mostly of academic interest. The success of efforts to experimentally determine a large number of sample 3D structures can be used as templates for building models of the rest. Of course, major technical challenges remain, ranging from the robotics of large-scale experimental design to the search for proper experimental conditions, to techniques for automatically generating the samples. Structural genomics has great promise to finally increase the number of 3D structures available for analysis by at least an order of magnitude. There are many computational challenges for protein structure in intelligent systems, including

- The analysis of the database of known structures and the sequence databases (with mostly unknown structures) to identify potential high-impact targets for 3D structure determination.
- The prediction of (full or partial) elements of 3D structure from 1D sequence information (for example, predicting the class of protein structure, the location of secondary structural elements, or the overall topology), a supervised learning problem.
- Understanding how biological molecules interact physically to transfer signals, including protein–protein and protein–DNA interactions as well as protein–drug (or other small-molecule) interactions. Approaches typically combine physical principles with empirical models.
- Analyzing 3D structures to find common structural motifs that can explain function. These are routinely formulated as pattern recognition problems.
- Understanding how the 3D structure of proteins evolved over time and how related 3D structures in different organisms have adapted to the special needs of those organisms.
- Designing new proteins to have new or modified functions for medical or industrial purposes.
- Supporting the analysis of experimental data, often requiring solutions to combi-

natorial hypothesis generation problems.

The future of studying intelligent systems in protein sequence and structure is bright because of the anticipated increase in available data. The physical and spatial quality of these molecules make them good targets for the application of methods from robotics, planning, computer-assisted design, image understanding, and general machine learning.

### Pathways and networks

The past few years have witnessed a notable shift of attention within computational biology. Efforts previously focused on the analysis of DNA and protein sequence (with some 3D structural analysis). Interest is now increasing in how these molecules interact to form pathways for metabolic conversions from one substance to another, and how genes form networks to regulate the timing and location of events within the cell.

The natural relationship between biological network data structures and general graphs suggests an opportunity to apply the results from graph theory to biology. The primary challenge is using experimental data to deduce the hierarchy of control that exists between genes and gene products (the proteins). Promising initial work includes the use of Bayesian belief networks to infer genetic regulatory networks.<sup>7</sup>

The definition of a generally applicable representation for biological processes is an important open problem for intelligent systems. Whereas biological structure is well defined (usually Cartesian coordinates for atoms), the concept of function is much more difficult to capture. First, function exists at many levels of description (“metabolism” is a high-level concept, “glycolysis” is a medium-level concept, and “add phosphate to glucose” is relatively low-level), and function representations need to be able to capture these hierarchies. Second, function has temporal and spatial connections that can also be difficult to manage. So, functional descriptions tend to be cognitive structures useful for biologists when constructing and evaluating models, and as such they do not have the sharp edges and crisp definitions of physical models. The study of biological pathways and networks expose the inadequacy of current approaches and represents an important driving application for the development of improved methods. Success in this area has implications for functional

representations across all scales of biology, to the organism and ecosystem level.

### One important organism: homo sapiens

The publication of the draft human genome is clearly one of the magnificent contributions to science in this century (all of one year old). The draft contained approximately 30,000 genes and an even larger amount of nongene coding, control, and structural elements. Much of human biology for the next 20 years will focus on making sense of this genome and using it to understand biology and improve human health. Currently, the challenges relate to finishing the genome’s “final” draft and understanding the sources and consequences of variation in the genome. It is quite sobering to look at three billion bases and realize that a small percentage variation in these bases explains all human variation—less than one percent.

For informatics, the challenge is to analyze the DNA to locate the regions that code for proteins are and that control the expression of proteins. There is a sense that the signals that control human development and biological function are created with a combinatorial approach—multiple proteins that can bind the DNA control regions of genes exist. Nature mixes and matches these to create control strategies for overlapping (for example, by sharing a subset of protein control regions). So, we need methods that can look for weak signals in the DNA and combine them with external experimental data sources to define and recognize these control regions.

The human genome draft provides the challenge of associating genomic variation in individuals (their genotype) with the functional manifestations of this variation (their phenotype). The area of genotype–phenotype correlation is active and difficult. There is much more genetic variation among humans than is functionally important—some variation is just noise. Thus, there are many false statistical associations that can be found between variations in the genome and variations in function. We need robust methods (most likely, the combination of statistical and empirical methods) that can sort through the possible associations to find the ones that are both statistically and biologically sensible. At this stage, the main emphasis is identifying the regions that vary (one important type of variation is the single nucleotide polymorphism, or

SNP, which is a single position in the genome that differs across individuals, usually by having one of two DNA bases). Databases have been established to characterize the major types of variation, and the search for the variations of functional importance is being undertaken in earnest, because it might have significant financial implications to those who discover the important associations (for defining drug targets, predicting disease risk, or providing other prognostic information).

Another important area for genotype–phenotype studies is pharmacogenomics, which focuses on how genetic variation contributes to variation in the response to drugs. An understanding of how genetic variations affect the efficacy of drugs, the levels achieved in the blood, and the occurrence of side effects might lead to an ability to prescribe medications more precisely and with a much better understanding of a prescription’s likely outcomes. The PharmGKB resource ([www.pharmgkb.org](http://www.pharmgkb.org)) is an NIH-funded resource that gathers information about pharmacogenomics and makes tools available for analyzing this data.

### A community of independent, interacting organisms

Biology has at its roots the study of the natural living world. Darwin studied birds and Mendel observed garden peas. Understanding the origin and development of species is an important goal of modern biology. With the increasing availability of completely sequenced genomes, we have an unprecedented opportunity to compare and contrast organisms, and to gather a high-resolution understanding of where they share elements and where they have diverged or acquired unrelated genetic elements. The rough draft of the human genome, for example, suggests a number of genes with viral and bacterial origin that seem to have become a stable part of the human germ line. Similarly, every organism’s genome will have a story to tell about that organism’s history—and the history of individual sets of genes (how they were acquired, how they have evolved, and what they do today). A reconstruction of these histories will provide a picture of what has happened on earth during the last four billion years.

In addition to studying the diversity of species, the availability of multiple genomes is extremely useful in studying individual species. By examining the similarities and

differences between organisms that are “close” to humans (that is, they diverged evolutionarily relatively recently) such as mice, pigs, and monkeys, we can get useful information about which elements of our genome are generic to all mammals and which are particular to humans. Comparative genomics relies heavily on computational comparison of genomes and on reasoning about the resulting differences to understand the way that particular systems function.

### **Ontology design and maintenance**

The creation of methods for defining and maintaining shared domain models within biology is critical, particularly representation of biological function. These techniques will be important for creating an infrastructure that is compatible with computational approaches. Most biological knowledge is currently stored in natural language text, representing problems for computational approaches that require more structured access to data. The first step in structured access is a conceptual space and set of shared vocabularies that allow at least a subset of biological discourse to be written down formally. The Gene Ontology effort to create a controlled vocabulary for function is a first step,<sup>8</sup> and efforts to create hierarchical data models for biological knowledge bases and to use description logics for describing the domain are important.

### **Natural language processing**

As a corollary to the previous discussion of ontologies, it is also critical to nurture natural language processing (NLP) techniques in biology. The Medline database contains abstracts of more than 10,000,000 published articles in biomedical research ([www.ncbi.nlm.nih.gov/PubMed](http://www.ncbi.nlm.nih.gov/PubMed)). In addition, almost all biomedical research journals are making full text available on the Web. The creation of controlled vocabularies (such as the Unified Medical Language System)<sup>9</sup> will facilitate the extraction of information from these data resources, thereby facilitating the growth of intelligent systems within biology. Currently, most NLP work in biology focuses on extraction of relationships from literature (such as gene–gene interactions, protein–gene interactions, and gene–drug interactions for the purpose of cataloging). There has been intriguing work in the use of the literature to infer the overall structure of genetic networks.<sup>10</sup>

### **Data mining and intelligent integration of information**

Biology is fortunate to have a plethora of useful (distributed) databases. The task of mining these databases is particularly challenging because it can require the integration of generic data mining approaches with relatively deep biological models. In addition, there is a problem of poorly shared semantics across multiple (often Web-accessible) databases. Failure to accurately integrate biological knowledge into machine learning algorithms can lead to both trivial and nonsense inferences that cause biologists to lose confidence in the methods. On the other hand, there is clearly a need for general-purpose methods to generate exploratory hypotheses, as well as methods for trying to prove these hypotheses. The marriage of information integration and the development of ontological technologies for structured representation of knowledge will fuel the continued use of data mining techniques.

### **Tools for intelligent visualization and interaction**

Biologists think about their work graphically. The contribution of many papers can be summarized as a change to a paradigmatic graphic that summarizes the understanding of a system.<sup>11</sup> A major challenge for intelligent computational systems for biology is the creation of graphical interfaces that allow biologists to operate in familiar territory while giving computational access to these models. The use of electronic publishing has also created an opportunity for graphical display of biological data that is more dynamic than previously possible in printed journals and textbooks.

### **Emerging trends**

Some trends are affecting how intelligent systems in biology might evolve in the next decade. Biology has its roots in reductionism. For over a century, biologists have been trained to take complex phenomena and break them down into reduced systems that can be manipulated, controlled, and studied. This is the basis for biology’s success during the last 50 years. However, biologists are realizing that it is necessary to move away from reductionist approaches and toward systematic approaches to biology. Some have even suggested that certain reductionist paradigms (for instance, about how enzymes work) have led to results that might be

mostly irrelevant to the actual biological reality of these systems.<sup>12</sup>

A shift away from reductionism would have major consequences for biology. First, virtually all biologists are trained to work under the reductionist paradigm, so we would need a new model for biological work that is currently not emphasized. Second, the use of qualitative and quantitative simulation at a systems level would become increasingly important. Our understanding of a system would have to be expressed as our ability to predict, understand, and manipulate the integrative phenomena that emerge from such simulations. Finally, we would have an opportunity to unify biology from molecular to ecosystem levels, because a relatively uniform systems view of biology would let us consider interacting systems at different granularities, perhaps with more facility.

Intelligent systems research in biology has amazing potential. In fact, the annual meeting for the International Society for Computational Biology ([www.iscb.org](http://www.iscb.org)) is the “International Conference on Intelligent Systems for Molecular Biology” (<http://ismb02.org>). An important goal is to harness the enthusiasm in the field to pursue society aims. In particular, one aim is to nurture the emerging field of computational biology and bioinformatics within academic structures, governmental funding agencies, and foundations.

**T**he intellectual challenges to information and knowledge processing in biology are exciting and promise to provide problems that will continue to drive the development of improved methods for intelligent systems. As our understanding of biology increases, the role of intelligent systems will be not only to assist in creating new knowledge but also to provide methods for storing this knowledge. Biology’s complexity is too great for the human mind to track, and written publications only provide a single projection of knowledge onto paper. The long-term resting place for our full understanding of biology will be the intelligent systems we build.

### **References**

1. E.S. Lander et al., “Initial Sequencing and Analysis of the Human Genome.” *Nature*, vol.

- 409, no. 6,822, 15 Feb. 2001, pp. 860–921.
2. J.C. Venter et al., “The Sequence of the Human Genome,” *Science*, vol. 291, no. 5,507, 16 Feb. 2001, pp. 1304–1351.
  3. C.B. Burge and S. Karlin, “Finding the Genes in Genomic DNA,” *Current Opinion in Structural Biology*, vol. 8, no. 3, June 1998, pp. 346–354.
  4. R.A. Durbin et al., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge Univ. Press, Cambridge, UK, 1998.
  5. R.B. Altman and S. Raychaudhuri, “Whole-Genome Expression Analysis: Challenges beyond Clustering,” *Current Opinion in Structural Biology*, vol. 11, no. 3, June 2001, pp. 340–347.
  6. M.M. Yusupov et al., “Crystal Structure of the Ribosome at 5.5 Å Resolution,” *Science*, vol. 292, no. 5,518, 4 May 2001, pp. 883–896.
  7. N.M. Friedman et al., “Using Bayesian Networks to Analyze Expression Data,” *J. Computational Biology*, vol. 7, nos. 3–4, Aug. 2000, pp. 601–620.
  8. M. Ashburner et al., “Gene Ontology: Tool for the Unification of Biology,” *Nature Genetics*, vol. 25, no. 1, May 2000, pp. 25–29.
  9. D.A. Lindberg et al., “The Unified Medical Language System,” *Methods of Information in Medicine*, vol. 32, no. 4, Aug. 1993, pp. 281–291.
  10. D. Proux et al., “A Pragmatic Information Extraction Strategy for Gathering Data on Genetics,” *Proc. 8th Int’l Conf. Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, Calif., 2000, pp. 279–285.
  11. R.M. Felciano et al., “RNA Secondary Structure as a Reusable Interface to Biological Information,” *Gene*, vol. 190, no. 2, Feb. 1997, pp. 59–70.
  12. M.A. Savageau, “Reconstructionist Molecular Biology,” *The New Biologists*, vol. 3, no. 2, Feb. 1991, pp. 190–197.

## Acknowledgments

Russ B. Altman is supported by the Burroughs Wellcome Fund, NIH grants GM-61374, LM-06422, GM-07365, and National Science Foundation DBI-9600637.

**Russ B. Altman** is an associate professor of genetics, medicine and (by courtesy) computer science at Stanford University. His research interest is in the application of advanced computational techniques to problems in molecular biology. He received an AB in biochemistry and molecular biology from Harvard University and an MD from Stanford University, where he also received a PhD in medical information sciences. He is a member of the ACM, ACP, AAAI, and IEEE, and is the president of the International Society for Computational Biology. Contact him at Stanford Medical Informatics, Stanford Univ., 251 Campus Dr., MSOB X-215, Stanford, CA 94305-5479; russ.altman@stanford.edu; <http://smi-web.stanford.edu/people/altman>.

## The Impact of European Bioinformatics

Alfonso Valencia, *Protein Design Group*

Europe is undergoing a political integration process, the goals of which are still under discussion. Some countries, such as Germany, France, Italy, Spain, Holland, and Belgium, have coordinated aspects of their economy, including a common currency, under the administration of the European Commission. Key countries such as the UK and Denmark are still undecided about their degree of participation, and others, including most of the former Eastern-block countries, are actively seeking integration. The EC has limited power in equilibrium with the European parliament and national and regional governments.

Direct funding of the sciences was not included in the EC’s mandate. It was considered a strategic responsibility of the national governments. This political concept evolved into new modes of collaboration between countries and the EC. Bioinformatics could be a key area for these developments, because, if properly coordinated, it has a great potential for generating added value to the new genomics and proteomic technology.

Currently, some bioinformatics-related research is funded as networks of groups from separate countries, and others are funded as basic research services, such as databases. For example, the E-Biosci project, coordinated by the European Molecular Biology Laboratory as a large research institution, includes the participation of groups from different countries and one or two medium-sized companies. Other computational science and telecommunications-related programs are rarely accessible to the bioinformatics community.

### Common scientific projects

Bioinformatics and computational biology were first developed around structural biology groups, particularly among crystallographers in London and Cambridge. Part of this tradition is still alive

in Europe, where computational biology is strong in structure classification, comparison, and databases.

During the early ’90s, large multinational groups in the EMBL, London, and Cambridge generated considerable scientific activity, including the support of databases such as SwissProt and EMBL and the development of the first modeling and threading software. During these years, the yeast sequencing projects, in which Europe took the lead, combined the efforts of many small laboratories with the bioinformatics analysis coordinated by the Munich Information Center for Protein Sequences. Unfortunately, the project’s success, in part based on centralized management, did not lead to the creation of a network of bioinformatic resources. This situation, however, began to change with the coordination of various bioinformatic groups working on separate genomic projects.

During the late ’90s, bioinformatics branched out to many smaller centers, thanks to the new technologies available—that is, faster connections, Web access, distributed databases, and computer power. Incorporating new groups of computer scientists created a new focus of activity. Unfortunately, during this period some key senior scientists left Europe for research positions in the US. The European efforts shifted significantly toward sequencing smaller genomes and systematically analyzing organism function. The Sanger center, and sequencing centers in Germany (the Institute of Molecular Biotechnology in Jena) and France (the Genoscope and Centre National de la Recherche Scientifique), were the only European contributors to the sequencing of the human genome. The associated bioinformatics effort has concentrated exclusively on the collaboration between the Sanger center and the European Bioinformatics Institute (for example, the Ensembl project).

A clear example of the difficulties of adaptation to the intrinsic European diversity is the EMBnet organization. Conceived by the EMBL in the ’80s as a network of repositories of databases and basic analysis software, it has grown into a worldwide structure,

although it has failed to integrate into the new European framework by collaborating with the many emerging scientific groups.

The EBI is the key for the development of European bioinformatics. The institute, founded in 1992 as an outstation of the EMBL (the largest Europe common research facility in molecular biology), is maintained with funds provided largely by the member states of the EMBL and substantial support from the EC. The institute inherited the EMBL sequence database, the collaboration in the construction of Swiss-prot, and the intention to contribute to the protein structures. The optimistic view at that time was that it could be the European counterpart of the American National Center for Biological Informatics. The lack of a continuous scientific direction, accumulated problems in the recruitment of scientists, difficulties in structuring a network of collaborations both locally (the Sanger Center or the Cambridge scientific community) and internationally, the lack of a well-structured European training program, and budgetary problems of the EMBL headquarters have considerably delayed the EBI's evolution. Hopefully, the activity of the new director, whose appointment is pending, the involvement of the EC in the future of the institute, and the allocation of a separate budget inside the EMBL will be solved favorably for the only European common institute in bioinformatics.

## Organization

Germany, the UK, France, Holland, and Spain have organized networks and associations on computational biology and bioinformatics, with annual meetings in their own national languages (linguistic diversity is an important part of European culture). The main organizations in biology, such as the European Federation of Biochemical Societies or the European Molecular Biology Organization, have paid little attention to this field. The involvement of European scientists in the activities of the International Society of Computational Biology, which has been held in England, Greece, Germany, and this year Denmark, is positive. The recent decision of the European science foundation to launch a five-year program for functional genomics, with the aim of mobilizing the research and funding bodies in associated countries, particularly in areas related to bioinformatics, is also positive.

## Education

A few universities, including British and Swiss, offer master's degrees at the European level, and many universities have started their own bioinformatics programs. Still, efforts to coordinate corresponding programs have not been successful. Some of the most relevant efforts to create a common teaching structure are the ones carried out by the Universities of Bielefeld and Stockholm—both participate in the 5-Star consortium—and the bioinformatics project for the definition of standard teaching topics led by the University of Manchester.

## Companies

Small bioinformatics companies are spreading all over Europe, even in traditionally less active countries such as Ireland and Spain. The particular fragmented structure of the European economy has made this process very heterogeneous. Its success depends on the strength of the local academic environment, the legal facilities for knowledge transfer, the possibilities for scientists to participate in companies, the availability of private funds, and the existence of regional plans. A successful example of a positive combination of these factors is Lion Bioscience, a EMBL spin-off, under the German plan for regional development of biotechnology. Lion Bioscience started in 1997 with an initial capital of \$63,000; it now has a value of more than \$1.34 million.

## Strength in diversity

There's a lot to learn from the history of European bioinformatics in the past decade, particularly its participation in genomics and proteomics projects. While the massive cost and equipment involved in these new projects can be managed in larger centers, European strength resides in its diversity, which is better served by collaboration between specialized groups in different institutions and countries. To face this challenge, the VI Framework Programme of the European Commission (2002–2006) will focus not only on infrastructures but also on fostering new developments. This will be tackled by bigger integrated projects closely associated with networks of excellence, including collaboration between programs of different countries under the EC umbrella. The final aim would be the creation of what is called the European Research Area. This ambitious new concept will face serious questions

about bureaucratic cost, real scientific value, and consequences in countries that do not possess large technology centers.

At the time of this writing, the first positive evidence of the reactivation of European bioinformatics community has appeared. The EC has announced direct support (20M Euros) for a large network of laboratories around the EBI, including the development of science and technology on biological sequence and structure databases, DNA array repositories, and protein interactions.

## Acknowledgments

I am indebted to Victor de Lorenzo (CNB-CSIC), Luis Serrano (EMBL-Heidelberg), and Carlos Martinez-Riera (ECC), for their interesting opinions.

**Alfonso Valencia** is a group leader with the Protein Design Group at the Campus de la Universidad Autónoma de Madrid. His research interests include comparative genomics, protein structure and function prediction, collaboration in different biological systems, and text mining in scientific text. He is vice president and founding officer of the International Society for Computational Biology, coordinator of the Spanish National Bioinformatics Network, and is a member of the editorial board of *Bioinformatics*. Contact him at [Diseno de Proteinas, Centro de Nacional de Biotecnologia at the Campus de la Universidad Autónoma, Cantoblanco, M-28049, Madrid, Spain; valencia@cnb.uam.es](mailto:Diseno de Proteinas, Centro de Nacional de Biotecnologia at the Campus de la Universidad Autónoma, Cantoblanco, M-28049, Madrid, Spain; valencia@cnb.uam.es).

## The Asia-Pacific Regional Perspective on Bioinformatics

Satoru Miyano, *University of Tokyo*  
Shoba Ranganathan, *National University of Singapore*

The Asia-Pacific region spans the Asian and Australasian continents as well as the Pacific-rim countries. As such, the seeds of bioinformatics in this region have been sown as early as 1989 in India, followed by Japan and Australia in 1991. While bioinformatics research, service, and education have reached laudable heights in these countries as well as in Singapore, Taiwan, Korea, Malaysia, New Zealand and Russia, several other countries (Thailand, Indonesia, and the

Philippines to name three) are making considerable progress. Following the success story of Japan, the status of bioinformatics in the Asia-Pacific region is presented here.

### Bioinformatics in Japan

The Japanese bioinformatics project started in April 1991 as the Genome Informatics Project as the informatics part of the Japanese Human Genome Project. Minoru Kanehisa, at Kyoto University, established a new research area bridging biological sciences and computer science, and developed new computational techniques for genome research. As a result, he successfully organized the GenomeNet Service ([www.genome.ad.jp](http://www.genome.ad.jp)) operated by the Supercomputer Laboratory of Kyoto University in collaboration with the Human Genome Center at the University of Tokyo.

The second stage of the Genome Informatics Project began in April 1996. Data collection and knowledge organization were emphasized, but informatics technology development also continued. Along with the intensive sequence productions of various organisms (starting in 1995 with *Haemophilus influenzae* and *M. genitalium*), the systematic compilation of information about genes and gene products of these organisms became a central issue. KEGG<sup>1</sup> (Kyoto Encyclopedia of Genes and Genomes) is a unique knowledge database that attempts to computerize knowledge of molecular and cellular biology in terms of wiring diagrams of genes and gene products. This project also emphasized developing techniques for knowledge discovery from genomic data<sup>2</sup> and sequence interpretation.<sup>3</sup> New computational methods have been developed for handling and analyzing systematic data generated by functional genomics experiments, such as for predicting networks of interacting genes from microarray gene expression profiles.<sup>4</sup>

A forum for researchers, practitioners, and users working on various aspects of bioinformatics and genome informatics was organized when the Genome Informatics Project began. Its aims are to present recent research results (theory and practice), demonstrate systems, and explore directions for future research and new applications. This project started as a small workshop called the Genome Informatics Workshop (<http://giw.ims.u-tokyo.ac.jp/giw>). GIW 2000 attracted more than 500 participants, and the peer-reviewed papers presented there were pub-

lished in *Genome Informatics*. In 2001, members changed the workshop's name to the International Conference on Genome Informatics (still keeping GIW as its acronym).

The Japanese Society for Bioinformatics, a bioinformatics professional society founded in December 1999, has over 500 members. The papers and posters presented at GIW are electronically available from the JSBi Web site ([www.jsbi.org](http://www.jsbi.org)).

### Recent Japanese government initiatives

In 2001, the Japanese government announced the Millennium Project. The project deals with both the human genome and the rice genome; the results should contribute to the health and welfare of Japanese people. Full-length DNA analysis, standard Single Nucleotide Polymorphisms (<http://snp.ims.u-tokyo.ac.jp>), genes related to diseases and drug responsiveness, and bioinformatics technology are all part of this project. The government has started to introduce bioinformatics education programs. Programs are being planned at some universities in Japan, including the University of Tokyo. As is similar with most countries, however, it is hard to find enough researchers to fulfill this aim of bioinformatics education. Many other genome related government projects are also sprouting up like bamboo shoots after the rain.

### The future for Japanese bioinformatics

The Japanese infrastructure for bioinformatics is getting better. The DNA Database of Japan (DDBJ) at the National Institute of Genetics has provided a service for international DNA sequence databases since 1986. There are three DNA sequence submission sites: the DDBJ, EMBL, and GenBank. One-quarter of the DNA collected in 2000 was collected at DDBJ.

Currently, Japan has a problem with human resources in bioinformatics—not enough researchers. But the Japanese government is devoting considerable resources that are geared toward training scientists in this developing discipline. After bioinformatics education programs are installed, these universities must train graduate students to be better prepared for the growing needs in a post-genomic era. We hope that this will actually work. Moreover, judging from the potential of JSBi and the enthusiasm during GIW meetings, the Japanese bioinformatics com-

munity might grow soundly with a partnership between academia, industry, and government, with the infrastructure constructed by governmental policy and direction.

### Asia-Pacific Bioinformatics Scene

The Asia-Pacific Bioinformatics Network ([www.apbionet.org](http://www.apbionet.org)), formed in January 1998, started as a nonprofit organization at the Pacific Symposium of Biocomputing. It fosters bioinformatics network infrastructure development, data and information exchange, training program, workshop, and symposia development, and collaborations in the bioinformatics field. Founding member countries include Australia, Canada, China, Japan, Korea, Malaysia, Singapore, and the US. APBioNet addresses the essentials of bioinformatics as set out by Walter Gilbert. “We must hook our individual computers into the worldwide network that gives us access to daily changes in the databases and also makes immediate our communications with each other,” Gilbert said. “The programs that display and analyze the material for us must be improved—and we must learn to use them more effectively.”<sup>5</sup>

APBioNet works with network issues (its collaboration with APAN will link DDBJ, GenomeNet, HGC, NCC, Angis, and Maffin), hardware, and biological databases (the Biomirrors project, a joint APAN-APBioNet initiative providing core biological information), software, training, and outreach. BioGRID, a new transregional distributed-computing project has been launched for compute-intensive applications.

Since 1998, the membership of APBioNet has grown and now includes 256 individual members and 92 organizational members. Individual and academic institutional membership continues to remain free and open to all those interested in bioinformatics. The 2001 Annual General Meeting also resolved to accept corporate memberships, which will provide the organization with much-needed funding for travel fellowships.

Several member countries have taken a leading role in starting formal bioinformatics education to meet the growing manpower requirements both regionally as well as worldwide. The S\* Life Science Informatics Alliance ([www.s-star.org](http://www.s-star.org)) for global distance education in bioinformatics includes Singapore and Sydney, Australia from the Asia-

*continued on page 61*

Continued from page 20

Pacific region. The National University of Singapore started its graduate program in bioinformatics in July 2000. In Japan, a national committee is actively looking into initiating formal bioinformatics training programs. In Australia, several universities have started undergraduate programs and courses in bioinformatics and computational biology, while in New Zealand, the University of Auckland has recently started a module in bioinformatics.

APBioNet is working toward active involvement of the regional bioinformatics pioneers such as Japan, Australia, and Singapore to help initiate, sustain, and develop bioinformatics both within this region as well as to encourage collaborations with researchers in Europe, via its partner organization, the EMBnet as well as in the US and Canada.

Computational biologists in Africa have adopted the APBioNet model, and will soon be initiating the African Bioinformatics Network. ■

## References

1. M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, Jan. 2000, pp. 27–30.
2. H. Bannai et al., "Views: Fundamental Building Blocks in the Process of Knowledge Discovery," *Proc. 14th Int'l FLAIRS Conf.*, AAAI Press, Menlo Park, Calif., 2001, pp. 233–238.
3. K. Nakai and P. Horton, "PSORT: A Program for Detecting Sorting Signals in Proteins and Predicting Their Subcellular Localization," *Trends in Biochemical Sciences*, vol. 24, no. 2, Jan. 1999, pp. 34–36.
4. T. Akutsu, S. Miyano, and S. Kuhara, "Inferring Qualitative Relations in Genetic Networks and Metabolic Pathways," *Bioinformatics*, vol. 16, no. 8, Aug. 2000, pp. 727–734.
5. W. Gilbert, "Towards a Paradigm Shift in Biology," *Nature*, vol. 349, 1991, p. 99.

**Satoru Miyano** is a professor at the Human Genome Center and the Institute of Medical Science at the University of Tokyo and is a vice

director of the Institute of Medical Science. His technical interests include bioinformatics, computational knowledge discovery, computational complexity, and learning. He received his BA, MA, and PhD in mathematics from Kyushu University. He is a member of the editorial board of *Theoretical Computer Science* and the chief editor of *Genome Informatics*. Contact him at the Human Genome Center, Inst. of Medical Science, Univ. of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan; miyano@ims.u-tokyo.ac.jp.

**Shoba Ranganathan** is an associate professor of bioinformatics at the National University of Singapore and the vice president of APBioNet. Her research interests include the bioinformatics of protein structure, prediction and modelling, mining bioinformatics databases, and ligand docking and design. She is a member of the editorial board of *Applied Bioinformatics* and a referee for several bioinformatics journals. Contact her at APBioNet Secretariat, Bioinformatics Centre, Nat'l Univ. of Singapore, 10 Kent Ridge Crescent, Singapore 119260; shoba@bic.nus.edu.sg.

## Help Build the Community of Leading Software Practitioners!

### New Offer Just for IEI Members

Subscribe to  
**IEEE Software** magazine at the  
reduced rate of £60 (€76)!

*IEEE Software* delivers reliable, leading-edge, and useful software development information to engineers and managers. Each bimonthly issue—published by the IEEE Computer Society and peer-reviewed by experts—covers crucial advances in software development and professional issues, including these hot topics and important trends:

- Leading-edge programming practices
- Internet development
- Technical project management
- Object-oriented techniques
- Living with rapid technology change

Contact the Institute of Engineers of Ireland today

Tel: +01.6684341 • Fax: +01.6685508 • Email: library@iei.ie