

Bioinformatics Moves into the Mainstream

FEATURE

by Jennifer Ouellette

With the mapping of the human genome completed, bioinformatics is undergoing a sea change. Now that scientists possess maps of the human genome and those of several other animal species, they can look for differences and similarities between all the genes of multiple species, with the ultimate goal of gaining a comprehensive view of biological systems as a whole.

An explosion of data is being tamed with new systems

But genome mappings, those completed and those in progress, have generated a vast amount of biological data, and now more than ever, scientists need sophisticated computational techniques to make sense of it. To meet those ever-increasing needs, bioinformatics is shifting from software designed for a specific project in academic laboratories to the commercial mainstream.

Bioinformatics is an interdisciplinary research area loosely defined as the interface between the biological and computational sciences. In practice, the definition is narrower, according to Michael Zuker, a professor of mathematical sciences at Rensselaer Polytechnic Institute (RPI) in Troy, New York. For Zuker and many others, the term applies to the use of computers to store, retrieve,

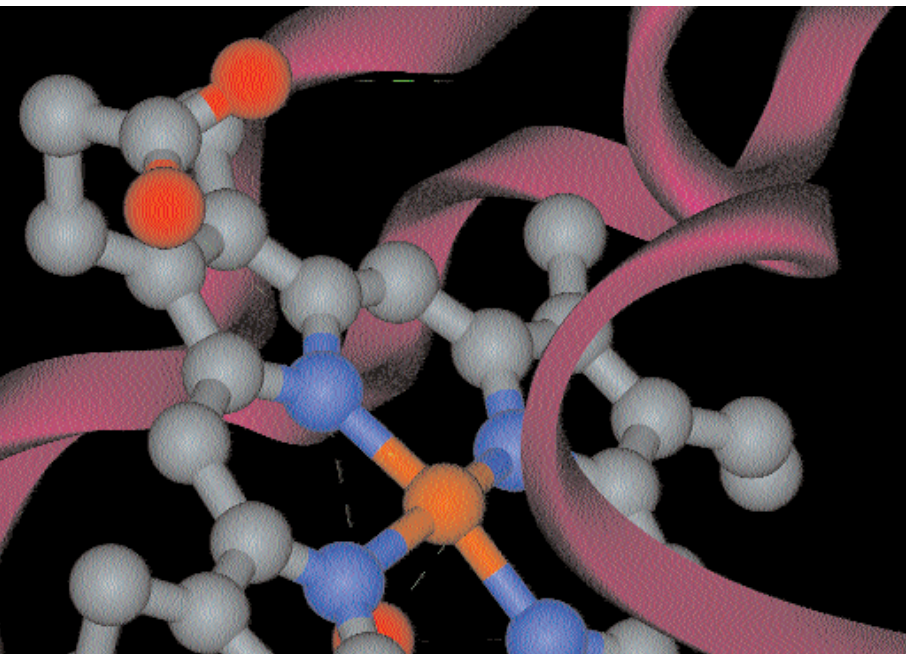
analyze, or predict the composition or structure of biomolecules. These include genetic materials such as nucleic acids, as well as proteins, the end products of genes.

Biology emphasizes three major types of information: one-dimensional structural data from DNA and genes; the three-dimensional structures of proteins; and complete biological systems with their emergent behaviors. Eric Jakobsson, director of the National Institutes of Health's fledgling Center for Bioinformatics and Computational Biology, maintains that biology "has always been an information-driven science. But it has taken time for the culture in the field to evolve to the point where biologists realize that to fully exploit computation, bioinformatics has to grow out of the cottage industry of each lab developing its own software for specific projects."

It has also taken time for computer speed, networking, and software tools to reach the point where they can help biologists. And with the development of high-throughput machines to sequence biomolecules and similar techniques, scientists can now perform multiple biochemical experiments, each of which generates enormous amounts of data. For example, the Human Genome Database contains approximately 3 terabytes of data, the equivalent of 150 million pages of information, and the volume of life sciences data is doubling every six months, according to Caroline Kovac, vice president of IBM's Life Sciences unit.

The need to manage and analyze this data largely drives the current bioinformatics boom. "Biology is awash in data," says Jakobsson. "We cannot exploit the body of data that is currently out there—we cannot mine it—without computers, and now we cannot even handle the data in our own individual labs without sophisticated computation." Doug Bassett, vice president and general manager of Rosetta Biosoftware (Kirkland, WA), agrees. "Researchers need smart software that can understand the biological complexity of the experiment and automate the routine analysis and data mining that need to take place," he says. "Software is no replacement for a biologist, but it can prioritize the information and present the researcher with the key data he or she needs to see."

This situation has stimulated a proliferation of start-up companies seeking to meet those needs and substantial investment by computer giants such as IBM. Front Line Strategic Consulting, Inc. (San Mateo, CA), predicted last year that the bioinformat-



Accelrys

Figure 1. Many start-up companies are seeking to meet a growing need for software programs that model, simulate, and analyze biomolecules, exemplified by this three-dimensional ladder diagram of DNA.

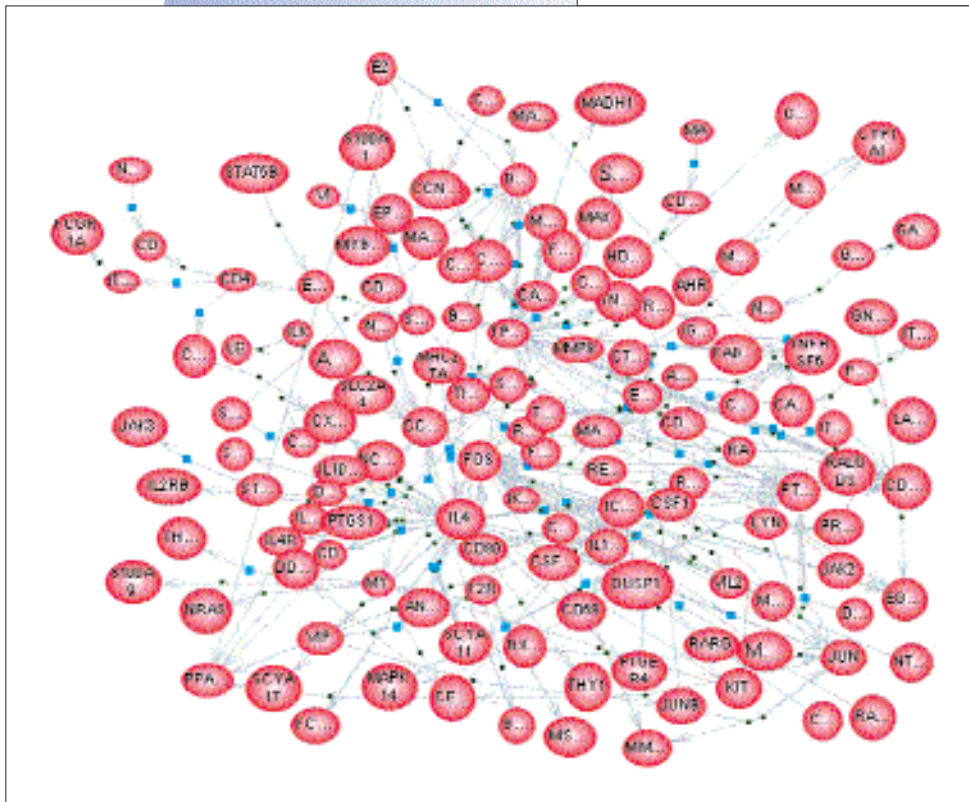
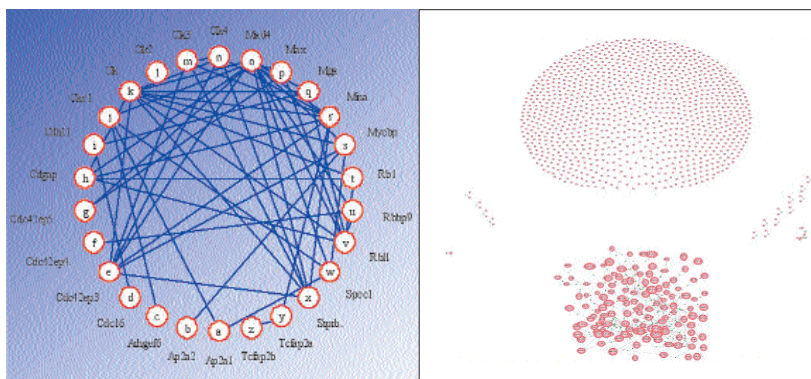
ics business will reach \$1.7 billion by 2006, growing by approximately 20% annually, and shaving 33% off the cost and two-years time off the drug discovery process by then. “Software written in academia is designed for a specific research purpose and is not always as broadly applicable as some users might wish,” says Scott Kahn, chief science officer for Accelrys, Inc. (San Diego, CA), citing the growth of commercial bioinformatics software. Accelrys creates software programs for modeling, simulation, and analysis of biomolecules (Figure 1).

Among its competitors is Rosetta Biosoftware, which markets a range of bioinformatics software for gene and protein expression data analysis, including the Rosetta Resolver system (Figure 3). Silicon Genetics (Redwood City, CA) produces GeneSpring 5.1 for gene expression analysis, and tools for automating the most common analytical projects in genomics labs. Bioinformatics Solutions, Inc. (Waterloo, ON), has developed advanced algorithms and innovative software for drug discovery. It has licensed its PatternHunter software to deCODE Genetics (Reykjavik, Iceland), among others, to identify genes and potential drug and diagnostic targets. And in Europe, LION Bioscience AG (Heidelberg, Germany) markets its Discovery Center, which integrates drug discovery data, applications, and documents on a single desktop computer.

Applications

Although bioinformatics aids a broad range of life-sciences research, Jakobsson divides it into roughly three categories: the application of principles of physics and chemistry to the modeling of biological systems at the atomic and molecular level; dynamical systems modeling, that is, representing how biological systems evolve as differential equations or stochastic processes; and pattern analysis, the process of searching for patterns in sequences of genes or proteins to gain insight into how a biosystem works.

For example, bioinformatics tools enable scientists to make predictions about what is called the secondary structure of proteins. “When people talk about structure, they usually mean the three-dimensional structure of living matter, such as a cell or membrane, or, in the case of DNA or proteins, they are referring to a three-dimensional model at atomic resolution,” explains Zuker, who created a bioinformatics Web site that registers up to 150,000 hits a month. “Secondary structures are reduced versions of these three-dimensional models; they don’t model every single atom, they model globular shapes.” Zuker’s algorithms have been used to find structural patterns in noncoding regions of genes—



those that do not specify the makeup of proteins—for drug design and DNA-folding research, and to predict the folding pattern of the large SARS virus.

One pioneer of pattern recognition research is Isidore Rigoutsos, manager of IBM’s Bioinformatics group. In 1996, he developed the Teiresias algorithm—named after a blind seer in Greek mythology—a combinatorial algorithm for discovering patterns and associations in streams of data. Since then, his group has generated and published other algorithms for tackling these problems. A simple way to explain pattern discovery is to think of an English text with all punctuation and spaces removed so the text runs together in a continuous stream. “If you give a non-English-speaking person this kind of textual input, the person should be able to identify the existing words and phrases as combinations that appear near one another and repeat multiple times, i.e., as patterns,” says Rigoutsos. “Given sufficient text of this type, one could attempt an automated reconstruction of the English vocabulary by recognizing these patterns. We play such

Figure 2. Some of the complex connections being studied in the Alliance for Cell Signaling, a project funded by the National Institutes of Health that focuses on genomes, gene products, functions, and pathways.

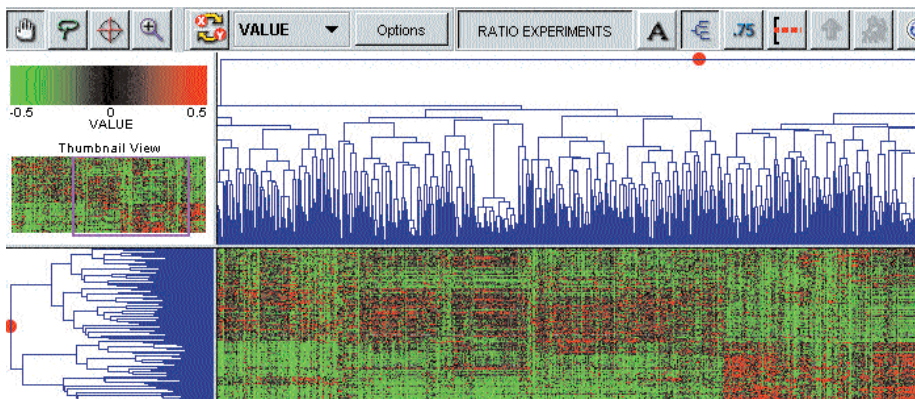


Figure 3. The visual scripting in this software enables a comprehensive analysis of the microarray data (large panel at bottom right) by integrating user-defined analysis plug-ins into a gene-expression profiling study.

games with amino acid sequences, nucleic acid sequences, gene expression data, and so forth.”

Rigoutsos has used Teiresias to process several large public databases of amino acid sequences and compile more than 57 million patterns, or seqlets—amino acid combinations that appear frequently in the data set—into a database dubbed the Bio-Dictionary, which associates the various seqlets with functional, structural, and other information already available in the public domain. One use of these patterns is the annotation of amino acid sequences in an automated manner. A system he developed uses this database to determine which of these patterns are present in a given amino acid sequence, and then the system attaches the patterns’ meaning(s) to the corresponding regions of the amino acid sequence. “It’s the same thing we would do if we had a dictionary and a lot of text in a language we don’t speak,” he says. “We would look up words, find the meaning, then chain the meanings together to make sense of the sentence.”

Proteomics

A major emerging application for bioinformatics is proteomics, the science of proteins and their interactions. The consulting firm Multimedia Research Group, Inc. (Sunnyvale, CA), estimates that the proteomics market will grow from \$565 million in 2001 to \$3.3 billion by 2006. “We

see a lot of potential in the proteomics arena in identifying gene and protein expression biomarkers that can help scientists diagnose a disorder, determine a patient’s prognosis, or whether a patient will respond to a particular drug,” says Bassett. Gene expression technologies measure cellular gene activity under various conditions to elucidate the molecular basis of a given disease and discover new treatments, although analysis of gene expression data is just one application of intensive computation used to infer protein function.

“There are relatively few applications [in biology] in which you need enormous computational power, but genetics and proteomics definitely have many of them,” says Mark Wilkins, vice president of bioinformatics for Proteome Systems (Sydney, Australia), which provides integrated Web-based tools and databases for fundamental proteomics research.

Determining the structure of a protein is necessary to determine its function. Proteins consist of an array of amino acids that fold and bend into complex three-

dimensional shapes that determine the function of each protein. If their shape changes because of some

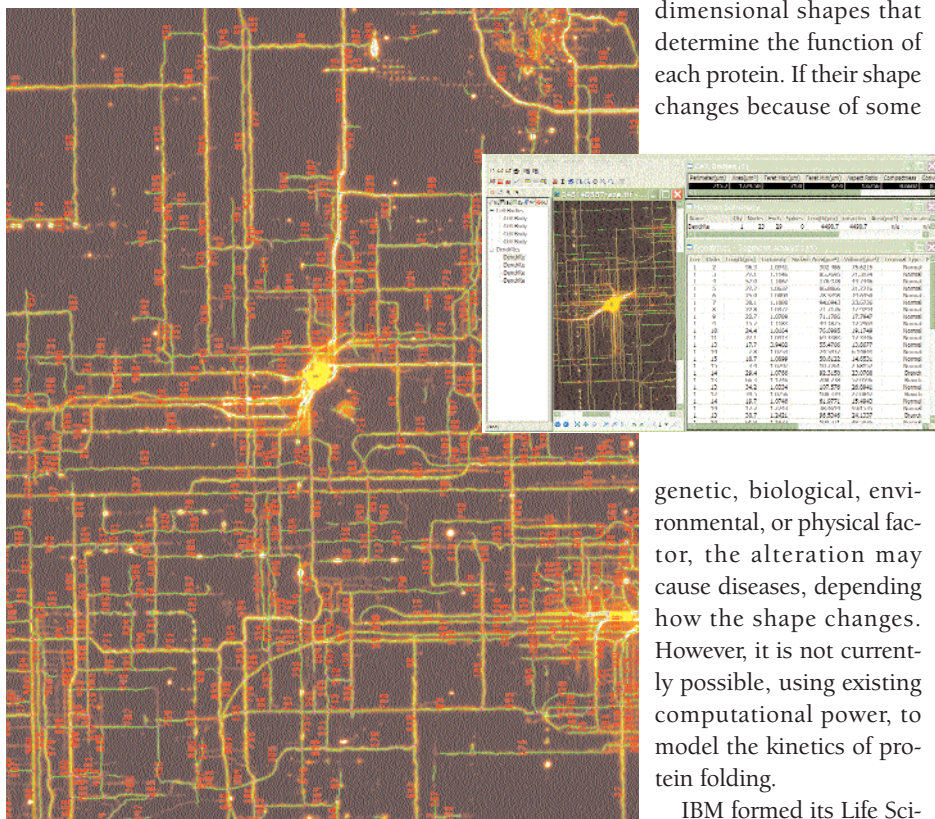


Figure 4. Medical-imaging analysis is moving from tedious manual sorting and examining of three-dimensional images to algorithms that automatically crunch raw image data into Excel spreadsheets.

genetic, biological, environmental, or physical factor, the alteration may cause diseases, depending how the shape changes. However, it is not currently possible, using existing computational power, to model the kinetics of protein folding.

IBM raised its Life Sciences business unit in August 2000 to bring its supercomputing expertise to bear on proteomics in

Badripath Royyam, Rensselaer Polytechnic Institute

particular, because proteins control all cellular processes in the body. Scientists have compiled vast databases of proteins, including IBM's own Bio-Dictionary. However, "all this data is worthless without the information technology that can help scientists manage and analyze it," says Kovac.

The linchpin of IBM's development effort is Blue Gene, a supercomputer that will be 100 times faster than any available today and is designed to advance our understanding of the mechanisms behind protein folding through large-scale biomolecular simulation. Blue Gene will feature IBM's next-generation cellular architecture design, in which chips will contain cells—processors that contain memory and communication circuits. IBM believes that cellular architecture will help scale computer performance from teraflops (10^{12} calculations/s) to petaflops (10^{15} calculations/s). In 2001, IBM announced plans to build Blue Gene/L, an intermediate step to Blue Gene, in collaboration with Lawrence Livermore National Laboratory. Slated to debut in 2005, Blue Gene/L is expected to operate at about 200 teraflops.

Medical-imaging analysis is another emerging bioinformatics application area. A typical biology wet laboratory generates vast amounts of data in the form of three-dimen-

sional images, which contain critical structural and functional information about tissue cells, according to Badri-nath Roysam, a professor of electrical, computer, and systems engineering at RPI. Today, trained technicians sort and examine these images manually—a tedious and time-consuming activity and one highly subjective and vulnerable to human error (Figure 4). Roysam has developed algorithms that enable researchers to crunch the raw image data down into Excel spreadsheets, which statisticians can analyze to determine significant differences between normal tissue and test samples. "The technique brings objectivity and consistency to image analysis," says Roysam. "The computer is relentlessly consistent. If you give it the same image on four different days, it will give you the exact same answer." Automation also makes it easier to scale up processes to higher throughput rates.

Roysam has also combined this real-time image analysis technique with eye-tracking instruments used in laser retinal surgery to provide a safer procedure. The software will be accessible online so that physicians worldwide will have access to it. Automated image analysis will also help speed the interpretation of Pap smears, a common test for cervical cancer that can yield false negatives as much as 20% of the time.

Systems biology

According to Bassett, a big step for bioinformatics will be developing “biologically aware and intelligent” analytical software that enables a researcher to tailor it to a particular experimental design. But bioinformatics’ real future lies in systems biology—“basically, linking the various pillars of bioinformatics data so they can be used in synergy for drug discovery and life science research,” says Bassett. Accelrys’s Kahn agrees. “The biggest challenge right now in bioinformatics is the integration of disparate data,” he says. “People are trying to bring an increasing amount of information around a specific question and bring to bear other data that otherwise had not been connected.”

Systems biology involves the analysis of all components of a biological system, which includes an in-depth analysis of how genes are expressed and their complex interactions within a cell, tissue, organ, or whole organism. The Institute of Systems Biology (ISB), a nonprofit research center in Seattle, Washington, likens this approach to trying to understand the modern medical health care system, which consists of many individual groups that must interact with one another: patients, physicians, nurses, hospitals, insurance providers, and so forth. In systems biology, the various types of biological information—DNA, RNA, protein, protein interactions, cells, tissues, and organs—all have individual elements. A comprehensive model of the entire biological system requires determining and integrating the relationship among all of them.

In June, for example, IBM announced a collaboration with Lynx Therapeutics, Inc. (Hayward, CA), and the ISB to study how cells of the human immune system respond to infectious diseases, with the goal of uncovering correlations between activated genes and the cellular response of macrophages to microbial infections. (Macrophages are critical players in the body’s response to infection, and when activated, they act as part of the body’s defense against infectious diseases.) Handling the data will help identify basic information technology requirements for future systems-biology research.

Because of the massive amounts of data involved, bioinformatics is a critical component to realizing true systems biology. This effort will require collaborations among biologists, computer scientists, chemists, engineers, mathematicians, and physicists to develop new global technologies—with global standards—and integrate them with the data acquisition, storage, and analysis tools of bioinformatics. “Ultimately, bioinformatics is really just an enabling technology, in the same way that any other piece of wet-lab technology is enabling,” says Proteome Systems’ Wilkins. “Systems biology is one very demanding user of that technology.”

However, “there is something qualitatively different between other technical advances that permit us to gather more and more precise and accurate data, and bioinformatics, which permits the transformation of those data into knowledge,” says Jakobsson. “The other technologies are powerful extensions of our senses; bioinformatics is a powerful extension of our brain.” 