023
024
025
026
00034
00044
gaccccgacacgacaatttcatcatactac
gactagactgatcg

Data Explosion

Bringing Order to Chaos with Bioinformatics

Digital Vision, PhotoDisc, Matt Ray/EHP

Scientists say a clearer understanding of gene–toxicant interactions will provide significant new opportunities for protecting public health. But there's a catch: these toxicogenomics promises lie hidden in mountains of data.

Thanks to technology advances, the nucleotide sequences that make up DNA, in addition to the amino acid sequences that make up proteins, are collected with robotic automation and stored by the millions in vast, expanding databases throughout the world. Microarrays, which provide snapshots of thousands of expressed genes simultaneously, are also data-intensive. Years ago, when sequencing was slow and tedious, scientists could study the output manually—no more. By necessity, they now need computers and sophisticated algorithms to wade through it all.

In recent years, the field of bioinformatics has emerged to meet these challenges. By definition, bioinformatics is the process by which informatics—the science of turning data into information—is applied to biology. A combination of computer science, information technology, and molecular biology, bioinformatics allows researchers to quickly access and interpret a rising tide of genomic information. This is critical for the genomic era: scientists are sequencing the genomes of many species, but they know little about how great regions of these genomes and the proteins they give rise to actually function.

In a basic application, bioinformatics allows researchers to search online databases such as GenBank for a given gene's composition, proteins, mutations, coverage in the scientific literature, and many other relevant parameters that are collectively termed "annotation." With more advanced applications, scientists use bioinformatics techniques to model chemical networks in living cells, including those stressed by disease or toxicity.

No researcher can possibly be familiar with all the known interactions in a cell, says Trey Ideker, a computational biologist with the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. Bioinformatics allows scientists to access, display, and interpret systems-level information. Fueled by bioinformatics, toxicogenomics is becoming an *in silico* science, with computerized data mining a key source of new discoveries.

Core Repositories

The rise of modern bioinformatics is rooted in the history of protein and nucleotide sequencing. The timeline arguably dates back to 1955, the year a Nobel Prize–winning British biochemist named Frederick Sanger first sequenced the protein bovine insulin. The first completed genome, sequenced in 1980, was that of a virus called phiX174. In subsequent years, scientists have gone on to sequence the genomes of higher organisms, including the human genome, which was completed in April 2003.

At first, sequencing was a slow and tedious process. The traditional technique—which involved gel electrophoresis and autoradiography—allowed scientists to manually sequence a single DNA fragment of 300–500 base pairs in about a day. This technique has been replaced almost entirely by automated high-throughput technologies to process DNA samples to determine the arrangement of nucleotides. The Applied Biosystems sequencers used in the decoding of the human genome, for example, are roughly 6,000 times faster than earlier approaches.

Today, sequencing is an international phenomenon. Entire consortia are devoted to sequencing the genomes of many species, including the human, the rat, the mouse, and many types of fish, birds, and microbes. Most of these sequences eventually wind up in a few publicly available databases. For nucleotides, the chief database in the United States is GenBank, maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine of the NIH. GenBank was actually started by the late physicist Walter Goad of the Los Alamos National Laboratory, who began compiling sequences there in 1979 while initiating efforts to create a national DNA/RNA database. The NIH created GenBank from Goad's original compilation, and the database was transferred to the NCBI from Los Alamos in 1992.

Today, all of GenBank's content is tightly integrated with two other databases, one (the EMBL Nucleotide Sequence Database) maintained by the European Molecular Biology Laboratory in Heidelberg, Germany, and the other (the DNA Data Bank of Japan) by the Center for Information Biology of the Japanese National Institute of Genetics in Mishima. GenBank's place in the U.S. research community is pivotal; most journals won't publish new sequences that GenBank has yet to accept. "GenBank is designed as a repository for all publicly available nucleotide data," explains NCBI staff scientist David Wheeler. "Anyone can come here [via the Internet] and pick what they need in terms of primary sequences."

Another publicly available source of nucleotide data is at The Institute for Genomic Research (TIGR), a nongovernmental research group based in Rockville, Maryland. Unlike GenBank, the TIGR database is populated with data produced by TIGR researchers in addition to data collected from bacterial sequencing projects going on around the world. Initially a pioneer in the field of bacterial genomics (scientists there sequenced the first bacterial genomes in 1995), TIGR has more recently broadened its scope to include nonbacterial species, including the parasites that cause malaria and sleeping sickness. It was also a major contributor to the sequencing of the human genome. The TIGR database is complementary to GenBank, in that it tracks all ongoing bacterial genome sequencing projects, in addition to those that have already been completed.

For protein sequences, the critical database is Swiss-Prot, which is a collaboration of the Geneva-based Swiss Institute of Bioinformatics and the European Bioinformatics Institute of the European Molecular Biology Laboratory. (Within the next three years, the United Protein Database, or UniProt, will combine Swiss-Prot and two other databases; see "Putting Proteins in One Place," p. A336 this issue.) With respect to microarrays, the database options are quite diverse. Among the public databases are the NCBI's Gene Expression Omnibus and ArrayExpress, which is maintained by the European Bioinformatics Institute. Various research organizations maintain a host of smaller "core" databases, including the holdings of the Microarray Center at the NIEHS National Center for Toxicogenomics (NCT) and the Stanford Microarray Database at Stanford University. And finally, the key public database for single-nucleotide polymorphisms, or SNPs, which are simple gene mutations, is dbSNP, maintained by the NCBI.

Side-by-Side Sequences

Bioinformatics has traditionally been focused on "sequence comparisons" performed with an evolving set of computational algorithms. With this process, scientists compare known and unknown sequences in an attempt to infer the properties of the latter. The underlying assumption is that similar sequences are homologous, meaning they are ancestrally related with similar properties across a variety of species. Screening a newly sequenced protein for homologues in Swiss-Prot, for example, provides predictive information about the protein's function, three-dimensional structure, and organization.

Because these predictions are based on sequence homology, they must be confirmed experimentally. Ideker says the ability to find new opportunities for experimentation is fueling a paradigm shift in biology. Because of bioinformatics, he says, biology is becoming a predictive rather than merely descriptive science.

Like sequencing itself, sequence comparisons have evolved from their tedious origins. The first algorithms, such as the Needleman–Wunsch algorithm introduced in 1971, were designed to allow "global alignment." These algorithms align every amino acid or nucleotide in a sequence of interest to a known counterpart in a search for homologous regions.

Current sequencing approaches favor "local alignment" strategies that look for short regions of nearly perfect matches. The most widely used of these is the Basic

Bioinformatics Resources

Databases

ArrayExpress
http://www.ebi.ac.uk/arrayexpress/

Biomolecular Interaction Network Database
http://www.bind.ca/

Chemical Effects in Biological Systems
http://www.niehs.nih.gov/nct/cebs.htm

dbSNP
http://www.ncbi.nlm.nih.gov/SNP/

DNA Data Bank of Japan
http://www.ddbj.nig.ac.jp/

EMBL Nucleotide Sequence Database
http://www.ebi.ac.uk/embl/Access/index.html

GenBank
http://www.ncbi.nlm.nih.gov/Genbank/index.html

Gene Expression Omnibus
http://www.ncbi.nlm.nih.gov/geo/

Stanford Microarray Database
http://genome-www5.stanford.edu/MicroArray/SMD/

Swiss-Prot
http://us.expasy.org/sprot/

Transcription Factor Database
http://transfac.gbf.de/TRANSFAC/

Institutes and Centers

European Bioinformatics Institute
http://www.ebi.ac.uk/embl/index.html

European Molecular Biology Laboratory
http://www.embl-heidelberg.de/

Genome Canada
http://www.genomecanada.ca/

National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov/

National Center for Toxicogenomics, Microarray Center
http://dir.niehs.nih.gov/microarray/home.htm

National Institute of Genetics (Japan), Center for Information Biology
http://www.cib.nig.ac.jp/

North Carolina State University, Bioinformatics Research Center
http://statgen.ncsu.edu/bioinformatics/

Swiss Institute of Bioinformatics
http://www.isb-sib.ch/

The Institute for Genomic Research
http://www.tigr.org/

The Wellcome Trust Sanger Institute
http://www.sanger.ac.uk/

Toxicogenomics Research Consortium
http://www.niehs.nih.gov/nct/trc.htm

Whitehead Institute for Biomedical Research
http://www.wi.mit.edu/

University of California, Santa Cruz, Genome Bioinformatics Group
http://genome.ucsc.edu/

Software and Other Tools

Basic Local Alignment Search Tool (BLAST)
http://www.ncbi.nlm.nih.gov/BLAST/

Bioinformatics Organization
http://bioinformatics.org/

BLAST-Like Alignment Tool (BLAT)
http://genome.ucsc.edu/cgi-bin/hgBlat?command=start&org=human

Ensembl
http://www.ensembl.org/

Entrez
http://www.ncbi.nlm.nih.gov/Entrez/

Digital Vision, PhotoDisc, Matt Ray/EHP

Local Alignment Search Tool (BLAST®) software, available from the NCBI. By running BLAST, researchers quickly scan novel sequences against up-to-date content from GenBank and a host of other relevant databases. “BLAST was just amazing to us when it was released in the early nineteen-nineties,” recalls Fran Lewitter, director of biocomputing at the Whitehead Institute. “[Before BLAST,] it could take hours to compare sequences. But with BLAST you could enter a sequence into a computer, hit ‘return,’ and you’d get your answer immediately.”

Global and local alignments are often performed sequentially. Researchers will run a sequence through BLAST to identify short regions of high similarity and then run global alignments to identify a wider range of sequences around those alignments. Thus, it is possible to observe evolutionary changes around the more highly conserved surrounding regions.

BLAST is typically the first step for someone consulting GenBank to evaluate a novel sequence. Upon entry of the sequence, BLAST returns lists of accession numbers for other, similar sequences. Researchers click on these accession numbers and through the GenBank interface—known as Entrez—connect to databases of annotated information for the sequence matches. “Entrez is our general search system,” Wheeler explains. “It covers data contained in a variety of databases including GenBank, Swiss-Prot, PubMed, and many others.”

Another useful search tool for obtaining sequence annotation is Ensembl, offered by the European Bioinformatics Institute and The Wellcome Trust Sanger Institute, a biomedical research organization near Cambridge, United Kingdom. Like Entrez, Ensembl allows users to run BLAST searches and link results to annotated databases by accession number. And the University of California, Santa Cruz, offers a genome browser that is particularly well suited for novel RNA sequences. This particular browser runs sequence comparisons with a program called BLAST-Like Alignment Tool (BLAT). According to Jim Kent, a research scientist with the university’s Genome Bioinformatics Team, BLAT maps RNA sequences to the genome at a speed roughly 50 times faster than BLAST.

George Bell, a bioinformatics scientist at the Whitehead Institute, says users are best served by employing a variety of search tools. “It’s like searching for movie reviews,” he explains. “You don’t want to go to just one site; you want as much information as you can get.” There are several good reasons to consult multiple sources for sequence matching and information, Bell says. No one site is definitive—the number of published sequences changes every day, as does the amount and quality of associated annotations. Furthermore, automated algorithms are all prone to error. Comparing the output of several sites provides a maximal amount of information. The question of which output to use, Bell emphasizes, is best answered using the researcher’s own scientific judgment.

Mining Microarrays

The bioinformatic techniques used to evaluate microarray data differ entirely from those used to compare nucleotides and proteins. In a toxicogenomics experiment with microarrays, fluorescent dyes are used to differentially label RNA from unexposed versus exposed animals. Results are measured in terms of relative fluorescence intensity, a continuous variable that Mike Waters, the NCT’s assistant director for database development, says is best compared using classical statistics for measurable outcomes, such as analysis of variance. These analyses can be run using standard desktop software, says Bruce Weir, director of the Bioinformatics Research Center at North Carolina State University in Raleigh. Such programs allow scientists to approximate which genes have been activated or inactivated by chemical exposure.

Multivariate statistics are then applied to microarray data to identify groups of genes that respond concurrently to chemical exposures. There are many techniques for grouping genes in this way, including gene clustering, a statistical method developed by Michael B. Eisen, a scientist with the Life Sciences Division at the Lawrence Berkeley National Laboratory in Berkeley, California.

Identifying chemically induced gene clusters is of high value to toxicogenomics. Modern microarrays show the expression of hundreds to thousands of genes simultaneously. Clustering of highly expressed genes provides structure to these voluminous data. “It allows you to find genes that are regulated in the same way,” Kent explains. “You may find these clusters are tissue-specific. Clustering basically allows you to create groups of gene families as we do with sequence homology. Therefore, we can infer something about the gene’s function according to the family to which it belongs.”

An effort to apply microarrays to toxicogenomics is currently under way at the Microarray Center at the NIEHS. Pierre Bushel, bioinformatics manager at the center, says data generated there are shared with public repositories such as Gene Expression Omnibus and ArrayExpress, in addition to a new “knowledge base” at the NCT called Chemical Effects in Biological Systems. With this knowledge base, the NIEHS aims to provide the ultimate international resource for all toxicogenomics data. Bushel says most of the microarray chips currently used by the NCT are prepared in-house.

A key objective, says Waters, is to ensure that annotation for all of the center’s microarrays is current. This is a tall order, he admits. Annotated information in the public domain is continually updated. Ultimately, Waters says, the NCT wants to automate its annotation, using distributed annotation servers that track GenBank, Swiss-Prot, and other major databases, pulling in new information as it becomes available.

Presently, the NCT is working with Agilent Technologies on a mouse microarray for toxicogenomics studies. According to James Selkirk, deputy director of the NCT, this “ToxChip” is being designed in cooperation with the NIEHS-funded Toxicogenomics Research Consortium, a group of five academic research centers plus the Microarray Center. The intention, he says, is to produce a chip containing a large number of genes thought to be relevant to the toxicity of environmental agents. “This should be something that is of wide interest to the microarray profiling public,” Selkirk says.

Computing Biology

At a certain point, the knowledge gained from studying sequences and microarrays sets the stage for investigations of cellular networks and pathways. Toxicity is manifested by a stunningly complex array of cellular events. The nature of these complex systems is studied with an extension of bioinformatics called computational biology. Whether the two fields are actually distinct is a matter of debate. One view suggests that bioinformatics deals with the acquisition, storage, and presentation of data, whereas computational biology applies the data to biological models. But in general, both fields cover the spectrum of computer-related activities in biological research.

In some ways, computational biology is more applicable to proteomics—the study of protein function in biological systems—where experts say the biomedical benefits of genomic knowledge will ultimately be found. “The actual network of molecular interactions is elucidated with proteomics,” says Ideker. “Researchers in this field are asking two key questions: what are the protein–protein interactions, and what are the protein–DNA interactions? These are the fundamental iterations that we're concerned with."

According to Ideker, a number of experimental methods predominate in this type of research. These methods are currently focused mainly on studies in yeast. For protein–protein interactions, Ideker says, a key method is the 2-hybrid system (also known as the yeast 2-hybrid system). This experimental system allows researchers to screen for interactions in large numbers of yeast proteins simultaneously.

A high-throughput method for assessing protein–DNA interactions has been developed by Richard Young, a biology professor at the Massachusetts Institute of Technology and a member at the Whitehead Institute. Young's method is based on a technique known as immunoprecipitation. In brief, the technique involves tagging proteins, cross-linking them with DNA in a cell, and then purifying the protein–DNA linkages. By uncrossing the linkages, scientists are able to evaluate the nature of the protein–DNA interactions.

These interactions can then be made publicly available via a number of online repositories. According to Ideker, one of the best repositories for protein–protein interactions is the Biomolecular Interaction Network Database, coordinated in part by Genome Canada, a genomics research organization based in Ottawa. This database is specifically designed for studies in computational biology. An important repository for protein–DNA interactions is the Transcription Factor Database, coordinated by Research Group Bioinformatics of Germany.

According to Ideker, computational biologists mine these repositories to model cell networks. It's now possible to construct models of "virtual cells" that are broad although not detailed, he says. "It's also possible to really nail a particular pathway," he adds. Ideker is currently collaborating with Leona Samson, a professor of toxicology at the Massachusetts Institute of Technology, on computational studies investigating pathways of DNA repair following exposure to chemical mutagens.

Eventually, scientists hope to pull all the available genomic data into complete models that also address the influence of genetic mutations such as SNPs. These models will allow researchers to assess how genomic variations contribute to disease or the response to toxicants. But many difficult challenges remain. For instance, database information must be maintained in compatible formats for global searches. Databases must also be updated with respect to the ever-increasing body of biological knowledge. And of course, scientists still need to extrapolate the results of experiments in lower organisms such as yeast to mammalian systems, humans in particular. "We're dealing with a level of exceeding complexity," Waters says. "These are not advances that are going to come overnight."

Charles W. Schmidt