Chemical Design Automation News

Volume 8
Numbers 11 & 12
November/December
1993

Bioinformatics or Biology?

By George Michaels

The basic biological and medical sciences are on the brink of a major paradigm shift as the result of advances in biotechnology. Ten years ago, the idea that it was possible to completely determine the complete genetic code for a simple organism was considered science fiction. Now there is an international year effort to completely define the genetic code for a human being by the year 2005. Several smaller collaborations have also been formed to determine the complete sequence for few bacterial and yeast genome as well. These data sets should be completed by the end of the decade. As a result, we will have available the set of information that defines the metabolic machinery of those organisms. The technological and informational advances of these projects will have a major impact on basic biological research, medicine, computer science and related high technology industries.

What is Bioinformatics?

Lately, this is the one question that I have been asked most often. If you are a computer scientist, you would recognize that informatics deals with the management and extraction of information from data. This is the normal mode of operation for a scientist in any field. You have a question that you want answered. You design a method that will allow you to make an observation and gather data. Traditionally, these data were organized in some tabular form. Biologists have been doing this for centuries. However, the last 25 years have seen a tremendous explosion of biochemical and genetic data that describes various biological phenomena. Much of this data is currently in electronic form. The hard task is making the transition from data to information about a particular biological process. Szent-Györgyi expressed this sentiment well thirty-six years ago when he observed: "One of the characteristic features of present-day biochemistry is the coexistence of highlights with darkness, knowledge with ignorance. While we can perform reactions that amount to a "miracle" and, here and there, even improve on nature, we cannot answer many of the simplest and fundamental questions. We have, for instance, detailed information about the structure of the protein molecules but cannot tell why nature has put those atoms together in a specific way, what was the quality she was trying to achieve by doing so." [1] This is the challenge and science of bioinformatics, to answer those fundamental biological questions by identifying the knowledge embedded in the data available. We have enormous databases of chemical data, biochemical data, biophysical data, genetic data, metabolic data, and phenotypic or medical data available. The bioinformatician is a biologist whose primary tool for knowledge discovery is a computer and those databases. This is the "new paradigm" scientist that Gilbert described. [2] This new breed of scientist, the computational biologist, needs to be cultivated now in order to provide for future opportunities looming in the horizon.

The Challenges

The real goal for most bioinformatics research is to find a means of relating structure to biological function. This is true for the genome projects looking at the genetic information capacity of an organism as well as structural biology investigations that look at the biophysical ▸

Continued on page 34

Index

Continued from page 1

organization of proteins or larger subcellular complexes. There are several databases used for these investigations. The present limitations reflect the lack of ability to interoperate between databases. Let's suppose that we wanted to design a better enzyme. The databases available that would have relevant data for this work would include PIR for protein sequence and protein sequence alignments; PDB for crystallography and NMR three-dimensional structure coordinates; Genbank and EMBL for DNA sequence data; NCI and EPA for toxicology data; DBEMP, a database for metabolic enzymes and reaction mechanisms; and OMIM and GDB for human phenotypic and related genetic data that will soon be linked to similar mouse data. Each of these databases provides a wealth of data. However, if we wanted to determine the incremental metabolic efficiency improvement in any enzyme represented in more than three species, a custom set of tools would be needed because one of our current problems is that there is no standard set of semantics that will allow the cross mapping of information across these data collections. Each of the databases has a separate set of semantic relationships. There is little information included in the data describing the methods used to derive the data as well. Only the structural data from the PDB has any quality estimate associated with each record, and this data set is small. Each of these limitations provides an opportunity for research. I purposely haven't mentioned the molecular modelling and simulation tools that are well-known to this audience. This was not done to slight the computational chemists and their research, but to highlight the difficulties in dealing with the dissimilar data sources that could be brought to bear in addressing the biology. There is still an abundance of quantum mechanical and molecular simulation work to be done to provide the foundations for biology research on complex adaptive systems.

New Tools Are on the Way

Most of the data resources described above are available by computer via the internet. Similarly, there are several analytical tools that are currently available as e-mail services. For example, GRAIL is a neural net-based, exon-finding tool located at the Oak Ridge National Lab, that is optimized for human DNA sequences. A new DNA sequence can be e-mailed to the GRAIL server and a regional probability of coding is made. There are several similar tools available that are optimized for different species.

There is a need for an extension of this type of analysis that, with high degree of accuracy, can predict the secondary structure of the proposed exons found. Tertiary structure prediction from first principles is still beyond the existing capacity of current theory. On the other hand, reasonable approximations of the protein backbone folding pattern have been made using homology based modeling methods. As the number of different protein structures in PDB increases, the reliability of these methods should increase.

Client/server tools for interoperability are beginning to appear for biology databases. The National Center for Biotechnology Information (NCBI) has the network version of BLAST, the Basic Local Alignment Search Tool for rapid searching of the latest versions of the protein and nucleic acid databases. This is one of the most popular tools in use for similarity analysis of molecular sequence data. This system uses client software that runs on your local ▶

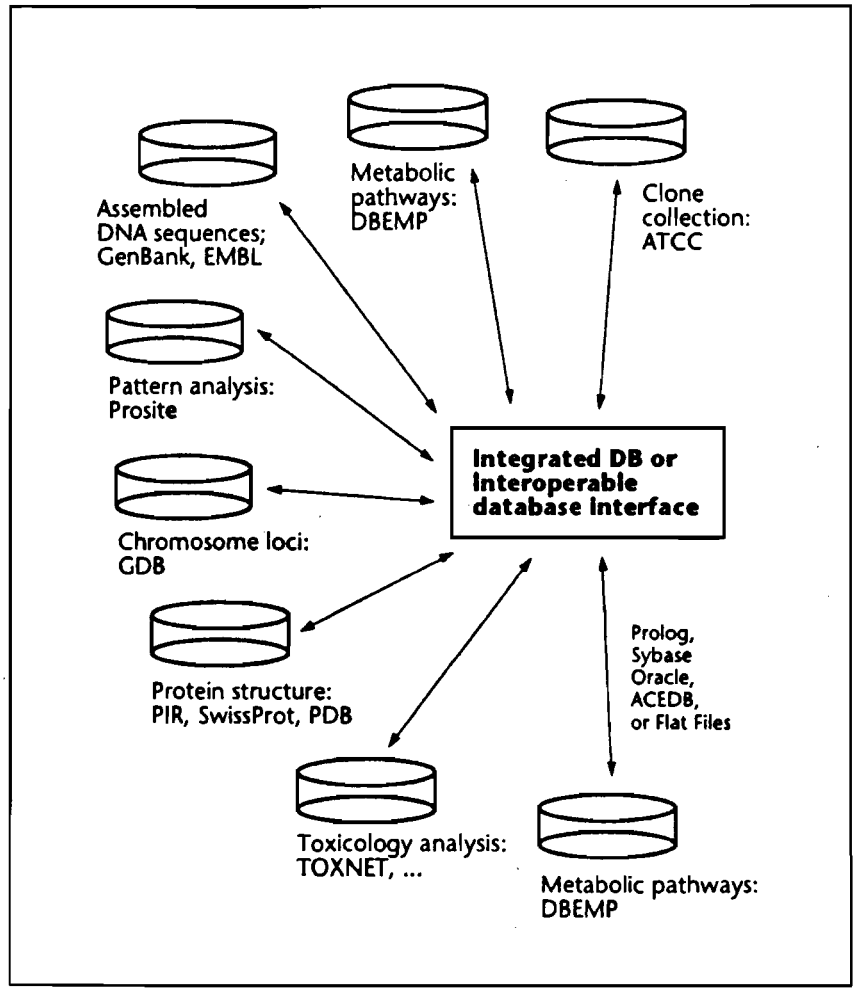

Figure 1. Bioinformatics Data for Analysis.

Continued from page 34

Unix machine that interacts with the high performance server at NLM that is the search engine. The client sends the query sequence via the internet to the NCBI server and the results are returned directly to your machine. The advantage is that NCBI manages and provides the latest data.

NCSA Xmosaic is a new network based information browser that works on Macintosh, DOS and Unix based computers. This tool allows anyone on the internet to take advantage of databases that have been translated in a hypertext format. Xmosaic permits browsing text, images and audio annotations. Genobase, an experimental database of molecular sequence data organized by organisms was developed at Argonne National Lab by Ross Overbeek and his collaborators. This database has been made available for use with Xmosaic at the NIH by Ron Taylor and John Powell. One advantage of the Xmosaic system is that personal annotations can be attached to any hypertext document. This would be an early step toward development of an extended community of annotators for databases

Collaborative community environments for biology that take advantage of the internet have seen limited use. One successful example is the Worm Community System (WCS) built by Bruce Schatz and his collaborators for the *C. elegans* research community. This system provides community access to a common lab notebook for the care and feeding of *C. elegans*, electronic publication and peer review, genetic maps, physical maps, DNA sequences and various tools for analyzing and querying the data.

With these computational and network based technologies coming online now, I expect to see even more innovations appear in the near future. The genome projects will spawn the development of several new automated technologies. A fully automated system for cloning DNA fragments should appear soon. Computerized DNA synthesizers are already common lab instruments. Clone processing machines are commercially available as well. Robots that run gels and extract fragments are already working prototypes. A network based technology approach toward protein design would take advantage of a database of DNA fragments from existing libraries and custom synthesized DNA fragments that were used by the cloning system to generate a library of permutations of known composition. This type of approach would be useful in exploring the rules for substrate recognition where multiple changes need to be made in a highly directed manner.

Gilbert[2] wrote in a recent editorial to Science: "The questions of science always lie in what is not yet known. Although our techniques determine what questions we can study, they are not themselves the goal. The march of science devises ever newer and more powerful techniques." In the end, the bioinformatics practitioner is a biologist whose more powerful techniques are computational in nature.

George Michaels, PhD, Institute for Computational Sciences and Infonnatics, George Mason University, Fairfax, VA 22030-4444 (703-993-1998).

References

1. Szent-Györgyi, A. (1957) *Bioenergetics*, v, Academic Press, New York.
2. Gilbert, W. (1991) *Science*, 349, 99. ■