EDITOR'S CORNER

The Biology Workbench—A Seamless Database and Analysis Environment for the Biologist

Bioinformatics is emerging as the "emperor's new clothes" in biology. While it is perceived differently by different subdisciplines in biology, the unifying theme is one of dealing with the enormous volume of data accruing from genome sequencing. The challenge, then, is to annotate, curate, and store the data and to enable easy access and analysis of the data by the biologist at large. In addition to being seamless, this data environment should also be all-encompassing in terms of the heterogeneity and diversity of data types and formats. Using the omnipresent World Wide Web technology, we have software engineered an infrastructure (unpublished), titled the "Biology Workbench," (http://biology.ncsa.uiuc.edu), which provides all of the publicly available databases and analysis tools in a "point-and-click" mode to a biologist with the only access requirement being a networked computer.

Biology is rapidly becoming an information-driven science. With the genomes of a plethora of organisms being sequenced in automated high-throughput sequencers, the problem becomes one of deciphering and processing the myriad data emerging. There are two major issues associated with bioinformatics. The first is the creation, management, and accession of large repositories of sequence, structure, and functional data; the second deals with trying to make biological sense of the data.

The debate on the organization of genome databases is only beginning. First, there is the problem associated with curating genome data to ensure integrity and consistency. Then comes the choice of objects to represent in the database, and the format in which the representation is made. At least two very diverse appraoches have been used for the database schema in biology. In the more simple approach, the database contains flat file objects where each object is represented in a raw form and schemes such as efficient indexing make the access to the database easy. In this form, complex queries are difficult to pose and cross-access to multiple databases is rendered more difficult. In the "computer science" approach, object-relational database schema are used, where extensive relational links and tables are built between objects. The enduser (biologist) can pose complex queries and work across databases, if metarelationships between databases are preestablished. The shortcoming is the relative slowness of query operations on such a database, as well as the need to build the entire relational schema de novo each time new data enters the databank. Until a few years ago, large users of these databases obtained the data in the form of tapes from repositories at periodic intervals. The advent of the worldwide web has enabled direct and instantaneous access of the data from the repositories.

In order to make biological sense of the data, biologists use a wide variety of analysis tools. Search tools such as BLAST, FASTA, and GenScan extract entries from sequence databases, then sequence tools such as PSA, and Grail are used to analyze a new sequence, and programs such as MSA or ClustalW create alignments between multiple sequences. Tools for interacting with higher levels of biological macromolecular organization exist as well, ranging from programs for protein structure prediction and protein interaction modeling to metabolic pathway analysis. Several of these tools are written in computer languages as diverse as Fortran, Pascal, and C, and the enduser needs to compile these on specific platforms. Given the lack of uniformity in the databases and program inputs, there is the need for multiple conversions of formats. In the past there has been a steep learning curve for the biologist who wishes to use these tools.

A complete biology analysis environment would shorten much of the time spent accessing and analysing the information. An analysis environment consists of a set of related databases, between which information can be shared, and a set of tools to manipulate the data. An ideal biology analysis environment would utilize the extant computational and communications infrastructure, provide interoperability between programs and databases to present the user with a single unified interface, and make available computational resources which were previously unavailable. The current level of communication infrastructure available via the World Wide Web (WWW), provides access to many of the important databases such as Genbank, Medline, SwissProt, and species-specific genome databases from a multimedia browser. Interoperability between databases and programs and a single user interface are important to save time for researchers. In the past, a researcher could connect to multiple databases and generate queries in many different formats. Results from these queries were returned separately, producing duplicated results and necessitating the hand collation of results. Then, conversion of the result formats was necessary to submit the data to various software. An ideal analysis environment alleviates these problems by interpretation of a standard query into database-specific queries, then interpretation of the results into a unified presentation and format suitable for submission across multiple programs.

Also, the ideal biology analysis environment will be accessible by all researchers, no matter how modest their computational power. Currently, most analysis programs require at least the computational power of a Unix workstation, and as biological information accumulates, the programs become even more time- and cpu-intensive. This situation presents a problem to the average researcher, who cannot afford to purchase an expensive workstation or to maintain the system properly. The current computation and communication infrastructure is capable of supporting a truly integrated analysis environment. To this end, we developed the Biology Workbench[1], a unique and versatile biological analysis environment. The Workbench consists of a WWW-based user interface, an interoperable interface to a variety of biological sequence and structure databases, an evolving collection of biological sequence and structure analysis tools, and easy-access links to large biological sequence and structure databases.

The Biology Workbench is object-centric, where the object can be databases, tools, sequences, alignments, structures, or other predefined biological objects. The Workbench is built around the cgi (common gateway interface) core, with its links to databases, analysis programs, and interfaces. Databases which have commonly describable objects are federated in the Biology Workbench. This enables simultaneous search across several databases for a single object. Several operations performed by the user are directly targeted at performing some task with one or more sequences. The tasks range from housekeeping (e.g., retrieving a sequence from adatabase or deleting from the work area) to analysis (e.g., sending several sequences to a multiple alignment program).

Scientists can use the Biology Workbench to investigate biological questions in detail. For instance, a research team needs to analyze a new, proprietary DNA sequence. The scientists begin by importing the sequence to the Workbench. Using point-and-click options, the scientists first run a BLAST search over GenBank. This reveals homologies to a small number of DNA sequences associated with a specific protein. The scientists translate the initial DNA sequence into a protein sequence and then click on more options to perform further homology searches over a federation of protein databases. Close matches are retrieved in seconds and, with a few mouse clicks, imported to the workbench for further analysis. Again, by clicking a mouse the scientists can use analysis packages—such as Multiple Sequence Alignment and MSAShade—to visualize how the proteins retrieved relate to the initial sequence. The scientists can also learn more about potential structure/function relationships by predicting secondary structures, which can be enhanced by viewing related 3D structures obtained using RasMol, which displays three-dimensional structures in HTML pages.

With increasing interests in comparative genome sequence analysis, the creation of orthologous and paralogous protein families and relating sequences to function, the Biology Workbench can aid the bench-biologist to easily traverse a variety of databases and tools with point-and-click access. The Biology Workbench is available freely to the academic community by accessing the URL, http://biology.ncsa.uiuc.edu.

Shankar Subramaniam
Departments of Biochemistry, Molecular & Integrative Physiology, and Chemical Engineering,
Beckman Institute for Advanced Science and Technology & National Center for Supercomputing Applications,
University of Illinois at Urbana-Champaign,
Urbana, Illinois

REFERENCE

1. Unwin, R., Fenton, J., Whitsitt, M., Jamison, C., Stupar, M., Jakobsson, E., Subramaniam, S. Biology Workbench: A WWW-based virtual computing and analysis environment for the biological sciences. In: "Bioinformatics." Letovsky, S., ed. 1998: in press.