with the Englehardt Institute in Moscow, where SHOM was demonstrated initially.

Informatics: Data Collection and Analysis

Explosive growth of information and the challenges of acquiring, representing, and providing access to data pose continuing monumental tasks for the large public databases. Over the last 3 years, the Genome Database (GDB), the major international repository of human genome mapping data, has made extensive changes culminating in the enhanced representation of genomic maps and gene information in GDB V6.0. Major issues for the Genome Sequence DataBase (GSDB), established in 1994, are to capture and annotate the sequence data and to represent it in a form capable of supporting complex, ad hoc queries. Both GDB and GSDB have been restructured recently to handle the increasing flood of data and make it more useful for downstream biology (see Research Narratives, GDB, p. 49, and GSDB, p. 55. *[http://www.gdb.org* and *http://www.ncgr.org/gsdb]*

Victor Markowitz, formerly of LBNL, has developed a suite of database tools allowing substantial modifications of underlying data structures while the biologists' query tools remain stable. *[http://gizmo.lbl.gov/DM_TOOLS/DMTools.html]*

The Genome Annotation Consortium (based at ORNL) was initiated in 1997 to be a modular, distributed informatics facility for analyzing and processing (e.g., annotating) genome-scale sequence data.

The many improvements in World Wide Web software now enable maps to be downloaded simply by using a browser with accessory software provided by GDB. Computers sift stretches of DNA sequence for patterns that identify such biologically important features as protein-coding regions (exons), regulatory areas, and RNA splice sites. Other computer tools are used to compare a new sequence (i.e., a putative gene) against all other database entries, retrieve any homologous sequences that already have been entered, and indicate the degree of similarity.

The Gene Recognition and Analysis Internet Link (GRAIL) at ORNL localizes genes and other biologically important sequence features (see box, p. 17).

Another analytical service that returns informative, annotated data is MAGPIE, provided through ANL by Terry Gaasterland. MAGPIE is designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated, with automated continued reevaluation as search databases grow (*http://www.mcs.anl.gov/home/gaasterl/magpie.html*). Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact. To this end, the WIT (What is There) system, which succeeds PUMA, supports the construction of metabolic pathways. Such constructions or models are based on sequence data, the clearly established biochemistry of specific organisms, and an understanding of the interdependencies of biochemical mechanisms. WIT, which was developed by Evgenij Selkov and Ross Overbeek at ANL, offers a particularly valuable tool for testing current hypotheses about microbial biology. *[http://www.cme.msu.edu/WIT]*

Researchers at the University of Colorado have developed another approach for predicting coding regions in genomic DNA, combining multiple types of evidence into a single scoring function, and returning both optimal and ranked suboptimal solutions. The approach is robust to substitution errors but sensitive to frameshift errors. The group is now exploring methods for predicting other classes of sequence regions, especially promoters. [software