Europe Opens Institute to Deal With Gene Data Deluge

CAMBRIDGE, U.K.—An 18th century manor park 13 kilometers south of Cambridge may seem an unlikely setting for the latest in 20th century biological science and technology. But the tree-covered grounds of Hinxton Hall have recently sprouted several new buildings that will be home to Europe's largest center for genome research. And this month marks another milestone in Hinxton's transformation when the European Bioinformatics Institute (EBI) formally opens its doors there for the first time.

Bioinformatics has become a buzzword in biology as the explosion in protein- and gene-sequence data turns molecular biology into an information science. With complete or partial sequence data now available for 300,000 genes and 100,000 proteins across all species, a researcher can quickly glean clues to the function of a newly sequenced gene or the structure of a freshly isolated protein—if she or he can find the relevant data amid the swarm of literature citations and widely dispersed information stored in computers. That's where bioinformatics' computational methods and databases come in.

The mission of EBI, an outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, is to find ways to stitch these information fragments into a coherent tapestry of easily accessible data. Other centers in other countries—notably the United States and Japan—have similar goals, but the focus at EBI will be on Europe. "It makes sense to expand and develop European databases because of the explosion of activity and existence of strong, cohesive research groups in Europe," says Guy Dodson, a protein structure chemist at the University of York. The institute will directly administer two databases, the EMBL nucleic-acid sequence database and SWISS-PROT, a protein amino-acid sequence database; it will also provide an access point for researchers to explore more than 40 other databases scattered around the world, allowing European labs to develop software links to them. Says Graham Cameron, head of services at EBI: "There is a huge awareness now of the importance of databases. I was appalled at the lack of computer awareness when we began 13 years ago. Now there are many professional and robust databases."

Indeed, the institute has come a long way from its cramped beginnings at EMBL in 1982. Then it had a staff of two researchers, and its initial database contained just 0.5 megabases (a megabase is a million base pairs) of nucleotide sequence. EBI research coordinator and Cambridge University biology professor Michael Ashburner says this database now contains more than 300 megabases and is growing by 75% each year. And institute head Paulo Zanella now supervises a staff of 40, which will be expanded to 70 over the next couple of years. EBI's projected annual budget is $8.5 million, supplied by EMBL and the European Union.

While the proliferation of databases has created an abundance of information for geneticists, it has also given them some major headaches. "A key challenge is how to link databases, and expectations are rising all the time," says Cameron. Those expectations have been fueled by growing access to all these databases through gopher sites, the World Wide Web, and other manifestations of the Internet: Biologists trawling the Net for sequence information naturally want to hop from database to database, but this is only possible if the databases have compatible structures, says Cameron. And right now, says Peter Karp, a biological database specialist at SRI International in Menlo Park, California, it's not very feasible to get from proteins in one metabolic pathway to the genes encoding these proteins, because the data are stored in incompatible formats.

So the EBI intends to make the incompatible mesh. "We are working to link our core databases with others developed by our collaborators to make them fully interoperable," says Cameron. EBI is working with the U.S. National Center for Biotechnology Information, part of the National Library of Medicine in Bethesda, Maryland, to improve integration of protein and nucleotide sequence data. In addition to administering SWISS-PROT, EBI is starting a project with bioinformatics researchers at Brookhaven National Laboratory in New York to develop standards for representing protein structure data.

The institute is also closely involved with many European research groups as a result of links created by EMBnet—a 20-node network initially set up by EMBL to coordinate access to data at the national level for users in Europe. The experience of creating this network will help EBI programmers develop access through other channels, such as the World Wide Web. "We are seeking a loose federation of databases," says Cameron.

Service isn't EBI's only mission, however. The institute already has two research groups working to develop software tools to validate protein structure data and has plans to fund two additional research groups. The research mission, scientists say, is a natural, given EBI's neighborhood. The Hinxton campus is also home to the Sanger Center, Europe's largest gene-sequencing laboratory. The Medical Research Council's Human Genome Mapping Project has also relocated to the site. "The potential for exchange of biological knowledge at Hinxton is quite unique," says Rodrigo Lopez, an EMBnet manager at the Biotechnology Centre of Oslo in Norway, noting that the combination of sequencing expertise and database design should give EBI some real advantages in genome database development and ensure it a prominent role in the international scene. "With the EBI, Europe really now has something to bang on the table," Lopez says.

Ashburner and his colleagues hope the banging will soon be under way. "There's enough information for genome explorers to begin to ask powerful questions about function and evolution," he says. With the advent of EBI, the researchers think, such explorations will yield information treasures to match investigators' wildest dreams.

–Nigel Williams

SELECTED EBI DATABASES

EMBL	Nucleic acid sequence database
SWISS-PROT	Protein amino-acid sequence database
dbEST	Expressed Sequence Tag database
dbSTS	Sequence Tagged Sites database
flybase	The *Drosophila melanogaster* set of databases
haema	Mutations in the factor VIII gene hemophilia A
haemb	Mutations/deletions associated with hemophilia B
hla	Alignments of HLA class I and II nucleotide and protein sequences
p53	Database of *p53* somatic mutations in human tumors and cell lines
pir	Protein sequences database
rebase	Restriction enzymes
transfac	Eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors

EBI-administered databases
Popular external databases accessible through EBI

SOURCE: EBI

REEVE PHOTOGRAPHY

Data on demand. The European Bioinformatics Institute will be open for business this month.