EBI (EMBL) European Bioinformatics Institute of the European Molecular Biology Laboratory

Swiss Institute of Bioinformatics

Protein Information Resource

PRESS RELEASE

Embargoed until 2pm US EST, Oct. 23, 2002

Establishing a universal knowledgebase of proteins

Major funding from NIH will help combine some of the world's best resources for biological information

Scientific contacts:

Rolf Apweiler
European Bioinformatics Institute
Hinxton Hall, Hinxton
Cambridge CB10 1SD
Great Britain
Phone: +44 1223 49- 4435
Email: apweiler@ebi.ac.uk

Amos Bairoch
Swiss Institute of Bioinformatics
CMU, 1 Michel-Servet
CH-1211 Geneve 4 / Switzerland
Tel: +41-22-702 5860
Email: info@isb-sib.ch

Cathy H. Wu
Director, Protein Information Resource
Georgetown University Medical Center
Box 571455, 3900 Reservoir Road, NW
Washington, DC 20057-1455 USA
Phone: +001 202 687 1039
Email: wuc@georgetown.edu

EMBL Press Office:
Office of Information and Public Affairs
Meyerhofstr. 1
D-69117 Heidelberg, Germany
Tel: +49-6221-387252/452
Fax: +49-6221-387525
email: info@embl.de
www.embl.de

Establishing a universal knowledgebase of proteins

Major funding from NIH will help combine some of the world's best resources for biological information

Oct. 23, 2002
Hinxton (UK), Geneva (Switzerland), Heidelberg (Germany), Washington DC (USA)

Today the U.S. National Institutes of Health (NIH) announced that they will provide major funding to establish a "unique, universal knowledgebase" of protein molecules. The award, totaling $15 million over three years, will go to the establishment of a new resource called the *United Protein Databases* (UniProt). It will be managed by European and American groups.

Until nearly four decades ago, there was no systematic, world-wide collection of protein data. Today this information is a key to all biological research because of the functions that these molecules carry out in cells and their crucial roles in disease processes.

The intervening years have seen the creation of *SWISS-PROT* and *TrEMBL*, operated by researchers from Switzerland and the European Molecular Biology Laboratory (EMBL), as well as the *Protein Information Resource* (PIR) at Georgetown University Medical Center and the National Biomedical Research Foundation (US). These groups will share the grant, using it to combine the strengths of each of their databases into a central public resource.

Rolf Apweiler of the European Bioinformatics Institute will head the UniProt project.

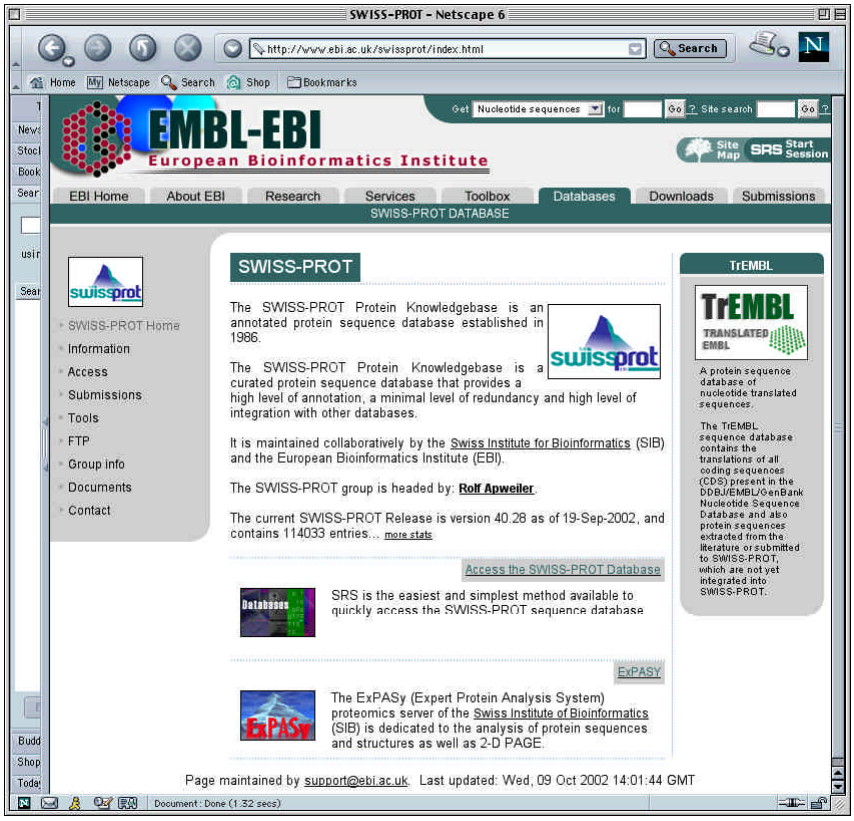
photo by: Marietta Schupp, EMBL

The combined project will be headed by Rolf Apweiler at the European Bioinformatics Institute (EBI) situated near Cambridge (UK). The EBI, which is a unit of the EMBL, operates and maintains several of the world's largest databases of biological information. Co-Principal Investigators are Cathy Wu, Director of PIR, and Amos Bairoch, director of the SWISS-PROT group at the Swiss Institute of Bioinformatics (SIB).

"With the increasing volume and variety of protein sequences and functional information that have become available," Apweiler says, "UniProt will serve as the central database of protein sequence and function. It will become a cornerstone for a wide range of scientists active in modern biological research, especially in the field of proteomics. Each UniProt entry will be a central hub for the data available about the protein."

SWISS-PROT was created in 1986 and is a collaboration between founder Amos Bairoch, at the SIB, and Rolf Apweiler's group at the EBI. Researchers regard it as the world's highest-quality protein database because the information it contains has been painstakingly reviewed and is continually updated by a staff of scientist/curators.

Early on it became apparent that SWISS-PROT alone, however, could not meet all scientists' needs. Improvements in DNA sequencing technology were generating information on hundreds of thousands of new proteins, and that number has risen astronomically since genome projects entered the high-throughput age. The hands-on approach that guaranteed high-quality information in SWISS-PROT was unable to cope with the flood of new information. In response the European teams created *TrEMBL*, which uses increasingly sophisticated methods to annotate the entries automatically.



SWISS-PROT is already one of the world's most highly-regarded and frequently-used sources of scientific information.

"The result is a hugely valuable resource that is as near to SWISS-PROT quality as possible," says Graham Cameron, Associate Director of the EBI. "When protein entries have gone through the process of annotation, they're checked by a curator and can be made available to the scientific community very quickly."

Researchers at the PIR have also made great strides in automating the use of computers to analyze proteins. The PIR was founded by Margaret Dayhoff, who published the first collection of protein sequences as a series of books from 1965 to 1978. Called the *Atlas of Protein Sequence and Structure*, the collection also included research on evolutionary relationships. As the quantity of data grew, it was made available electronically and on-line as the PIR-International Protein Sequence Database. Collaborators in the project have included the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID).

Central to quality and consistent sequence annotation at PIR is the protein family classification approach. "The approach, inspired by the superfamily concept pioneered by Margaret Dayhoff, is widely recognized as one of the most effective means for functional annotation of protein sequences," says Cathy Wu. "The collaboration will allow us to extend our methods to the annotation of all UniProt proteins."

"Combining the resources will give us even more new tools to automate and improve the process of annotation," Bairoch says. "It will give us the chance to combine everything into a unique central resource, and we're particularly excited that we'll be able to integrate some of the tools developed by Cathy Wu and her team at PIR."

Richard Roberts, winner of a 1993 Nobel Prize for his work on DNA, has been an enthusiastic supporter of the Uniprot project. "Given on the one hand the excellence of previous efforts to create SWISS-PROT and PIR, it makes a lot of sense to combine the efforts," he stated. "Since databases of these sorts are now of the utmost importance to molecular biologists worldwide, it would be most welcome to have a truly unified database that could be maintained in an up-to-date fashion and that combined the strength of each of the individual members of the consortium. I can think of no-one better qualified than these teams to pull it off."

SWISS-PROT currently holds entries on 116,000 proteins, TrEMBL contains 700,000, and PIR 283,000. Under the three years of the grant, Bairoch predicts the total number should reach well above the two million mark. "And that's a conservative estimate," he says.

- Russ Hodge, EMBL

About EMBL

The European Molecular Biology Laboratory is a basic research institute funded by public research monies from 16 member states, including most of the EU, Switzerland and Israel. Research at EMBL is conducted by approximately 80 independent groups covering the spectrum of molecular biology. The Laboratory has five units: the main Laboratory in Heidelberg, Outstations in Hinxton (the European Bioinformatics Institute), Grenoble, Hamburg, and an external research programme in Mouse Biology in Monterotondo near Rome. The cornerstones of EMBL's mission are: to perform basic research in molecular biology, to train scientists, students and visitors at all levels, to offer vital services to scientists in the member states, and to develop new instruments and methods in the life sciences. The Laboratory also sponsors an active Science and Society programme. Visitors from the press and public are welcome. For more information see the EMBL website at:

http://www.embl-heidelberg.de

or contact:

the Office of Information and Public Affairs
European Molecular Biology Laboratory (EMBL)
Tel: +49 (0)6221 387252
Fax: +49 (0)6221 387525
email: info@embl-heidelberg.de

The EMBL member states are:
Austria, Belgium, Denmark, Finland, France, Germany, Greece, Israel, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom

About the EBI

The European Bioinformatics Institute (EBI) is an Outstation of the EMBL, located on the Wellcome Trust Human Genome Campus in Hinxton near Cambridge (UK). The EBI grew out of EMBL's pioneering work in providing public biological databases to the research community. It currently hosts some of the world's oldest and largest collections of biological data, including EMBLBank, SWISS-PROT/TrEMBL, Ensembl, the Macromolecular Structures Database, and ArrayExpress. The EBI hosts several research groups and scientists continually develop new tools for the biocomputing community.

European Bioinformatics Institute, EMBL
Hinxton Hall, Hinxton
Cambridge CB10 1SD
Great Britain
Phone: +44 1223 49- 4435
Email: apweiler@ebi.ac.uk