



Joint Press Release

Embargoed until 9 am GMT, December 15, 2003

UniProt consortium goes on-line

Press Contacts:

Cath Brooksbank
Scientific Outreach Officer
EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus Hinxton
Cambridge CB10 1SD, United Kingdom
Tel: +44 (0) 1223 492525
Fax: +44 (0) 1223 494468
cath@ebi.ac.uk

Vivienne Gerritsen
Science communication group of Swiss-Prot
Swiss Institute of Bioinformatics
CMU, 1 Michel-Servet
CH-1211 Geneva 4, Switzerland
Tel: +41 22 379 58 82
Fax: +41 22 379 58 58
Vivienne.Gerritsen@isb-sib.ch

Lindsey Spindle
Director of Media Relations
Georgetown University Medical Center
Box 571405, 3900 Reservoir Road, NW
Washington, DC 20057-1405, USA
Tel: +001 202 687 7707
las46@georgetown.edu

Geoff Spencer
National Human Genome Research Institute
National Institutes of Health
Bethesda, MD 20892, USA
Tel: +001 301 402 0911
spencerg@mail.nih.gov

Trista Dawson
EMBL Press Officer
Meyrhofstrasse 1
D-69117 Heidelberg, Germany
Tel: +49 6221 387 452
Fax: +49 6221 387 525
dawson@embl.de
www.embl.de

Today the EMBL-European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and Georgetown University Medical Center's Protein Information Resource (PIR) announce the launch of UniProt, a new universal protein resource that will be the world's most comprehensive catalogue of information on proteins. UniProt will provide a 'one-stop shop,' allowing easy access to all the publicly available information on proteins.

Protein sequence databases have become a crucial resource for molecular biologists, allowing them to analyse the proteomes of newly sequenced organisms, to make intelligent predictions about the functions of newly identified proteins, and to move towards understanding how proteins interact to create pathways, networks and entire systems. To do this efficiently they need access to a defined set of features describing all the proteins that are known to exist or have been predicted to exist by extrapolation from their gene sequences.

Until recently there have been two major efforts to make this information publicly available. One was a collaboration between SIB and EMBL-EBI that resulted in two complementary databases, Swiss-Prot (renowned for providing a great depth of information on proteins through high-quality manual curation) and TrEMBL (a much larger database in which information on protein function is derived computationally by comparison with other proteins). The other was the PIR-International Protein Sequence Database (PIR-PSD), the world's first database of classified and functionally annotated proteins. These databases held different, but overlapping, subsets of proteins. "The launch of UniProt is tremendously exciting because databases that have been running independently for years have come together for the benefit of their users," explains Maria-Jesus Martin, Sequence Database Group coordinator at the EBI.

This unification was made possible by funding from the U.S. National Institutes of Health, totalling US \$ 15 million over 3 years. The National Human Genome Research Institute (NHGRI) is the primary funding institute, contributing \$3 million annually. Other NIH participants are the National Institute of General Medical Sciences (\$1 million), the National Library of Medicine (\$460,000), the National Institute of Mental Health (\$300,000), the National Center for Research Resources (\$100,000) and the National Institute of Dental and Craniofacial Research (\$50,000).

"Scientists today must face the challenge of understanding an increasingly large amount of data generated by the Human Genome Project and related resources. The UniProt databases will be a critical resource for investigators trying to unlock the secrets in genome sequences, both to understand biology and to translate basic research into improvements in health care," says Peter Good, Ph.D., the NHGRI programme director in charge of the UniProt project.

The UniProt databases launched today are the result of a hectic but immensely productive year of collaboration among the three institutions that make up the UniProt Consortium. "UniProt's structure resembles that of a wedding cake," explains Rolf Apweiler, UniProt's Principal Investigator. "Each tier of the cake represents a different database, optimized for different uses."

Underpinning the entire project is the UniProt Archive (UniParc) – the most comprehensive publicly accessible non-redundant protein sequence database available. Protein sequences are loaded daily from the public databases, including not only Swiss-Prot, TrEMBL and PIR-PSD, but also the EMBL–Bank/DDBJ/GenBank nucleotide sequence databases, the Ensembl database of animal genomes, the International Protein Index (IPI), the Protein Data Bank (PDB), the NCBI's Reference Sequence Collection (RefSeq), model organism databases such as FlyBase and WormBase, and protein sequences from the European, American, and Japanese Patent Offices. UniParc provides cross-references to the source databases, sequence versions and status.

The next layer of the wedding cake – and the centerpiece of the UniProt Consortium's activities – is the UniProt Knowledgebase (UniProt) unified from Swiss-Prot, TrEMBL and PIR-PSD. "This is the place to go if you want to know everything there is to know about a specific protein," explains Maria-Jesus Martin. The Knowledgebase contains a non-redundant set of entries that include information on protein function and classification, as well as cross-references to more than 40 other resources. The UniProt Knowledgebase consists of two parts, one containing fully manually annotated records and another with computationally analysed records awaiting full manual annotation. Sequences for which new functional, structural and biochemical data have been published are prioritized for annotation. The two sections will continue to be referred to as Swiss-Prot and TrEMBL, respectively.

Researchers will also be able to submit protein sequences directly to the Knowledgebase using a new web-based submission tool called SPIN. SPIN replaces Swiss-Prot's e-

mail-based submission system, making it much easier for researchers to submit sequences. "SPIN's forms allow researchers to submit more information about a protein's features in a more structured way," explains Vincent Lombard, who coordinated the development of SPIN. "This improves the efficiency of submission for both submitters and curators."

The top tier of the wedding cake contains three sub-layers – UniRef100, UniRef90 and UniRef50 – collectively known as UniRef (for UniProt non-redundant reference). "The UniRef databases will use newly developed automatic procedures to combine closely related sequences into a single record," explains Cathy Wu, whose group at PIR is responsible for their creation. Wu continues, "UniRef100 is a non-redundant version of all the sequences in the Knowledgebase, UniRef90 collapses all the sequences that are 90% or more identical into a single record, and UniRef50 collapses sequences that are at least 50% identical. UniRef50 speeds up searching significantly and doesn't reduce the effectiveness of homology searching. The three UniRef databases allow the user to choose between a fast search and a truly comprehensive one."

"With UniProt we can address some aspects of the challenges that life scientists are currently facing," says Amos Bairoch, the founder of Swiss-Prot. "There has been a tremendous growth in the quantity of biomolecular information that has become available in the past 10 years, yet this is only the beginning!" He adds, "Thanks to UniProt we can continue to provide a wealth of knowledge on the fascinating universe of proteins." "Such integrated knowledge in UniProt will facilitate scientific discovery at various levels of biological organization from genes and proteins to metabolic pathways, cellular networks, and organisms," agrees Cathy Wu.

UniProt can be accessed at <http://www.uniprot.org>. The individual members of the UniProt consortium have their own web pages at

<http://www.ebi.uniprot.org>,

<http://expasy.uniprot.org>

<http://www.pir.uniprot.org>

- end -

Scientific contacts:

Rolf Apweiler
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD
United Kingdom
Tel: +44 1223 494435
apweiler@ebi.ac.uk

Amos Bairoch
Swiss Institute of Bioinformatics
CMU, 1 Michel-Servet
CH-1211 Geneva 4
Switzerland
Tel: +41 22 379 5050
amos.bairoch@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
Tel: +001 202 687 1039
wuc@georgetown.edu

About the UniProt consortium:

The UniProt Consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). EBI, located at the Wellcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, is the founding center of Swiss-Prot and maintains the ExpASY (Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases. PIR, hosted by the Georgetown University Medical Center (GUMC) in Washington, DC, USA, is heir to the oldest protein sequence database, Margaret Dayhoff's *Atlas of Protein Sequence and Structure* published by the National Biomedical Research Foundation (NBRF) from 1965-1978. In 2002, EBI, SIB and PIR joined forces as the UniProt Consortium. This unification was made possible by funding from the U.S. National Institutes of Health, totalling US \$ 15 million over 3 years. The primary mission of the consortium is to support biological research by maintaining a high quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence Knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community. UniProt will build upon the solid foundations laid by the consortium members over many years.

About the EBI:

The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL), and is 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 hosts some of the world's most important collections of biological data, including DNA sequences (EMBL-Bank), protein sequences (Swiss-Prot and TrEMBL), animal genomes (Ensembl), three-dimensional structures (the Macromolecular Structure Database) and data from microarray-based experiments (ArrayExpress). The EBI hosts several research groups and its scientists continually develop new tools for the biocomputing community.

About EMBL:

The European Molecular Biology Laboratory is a basic research institute funded by public research monies from 17 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, and Outstations in Hinxton (the European Bioinformatics Institute), Grenoble, Hamburg, and Monterotondo near Rome. The cornerstones of EMBL's mis-

sion 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. EMBL's international PhD Programme has a student body of about 170. The Laboratory also sponsors an active Science and Society programme. Visitors from the press and public are welcome.

About SIB:

The Swiss Institute of Bioinformatics (SIB) brings together Swiss experts in bioinformatics and provides high quality services to the national and international scientific community. Members of the SIB include more than ten research groups in Geneva, Lausanne and Basel. The Institute has three missions: research and development, education, and service. It maintains databases of international standing (such as Swiss-Prot, Prosite, EPD, Swiss-2Dpage, GermOnline, etc.). It supplies and develops software and services that can be accessed from the SIB web servers and in particular ExpASY (www.expasy.org). It supplies services to the Swiss biomedical research community within the framework of the international network EMBnet and is in the process of creating a high-performance informatics platform (Vital-IT). It undertakes specific research and development activities related to the databases and software developed within the Institute. Swiss-Prot is the largest group of the SIB. It is composed of about 70 collaborators, most of which are involved in the curation and development of the knowledgebase.

About PIR:

The Protein Information Resource (PIR) at the Georgetown University Medical Center (GUMC) and the National Biomedical Research Foundation (NBRF) in Washington, DC, USA, is an integrated public resource of protein informatics. For over three decades, PIR has provided many protein databases and analysis tools to support research on molecular evolution, functional genomics and proteomics, and computational biology. In addition to PIR-PSD, PIR maintains the PIRSF classification system that reflects evolutionary relationships of whole proteins; the iProClass integrated database of protein family, function, and structure, as well as a web site (pir.georgetown.edu) for information retrieval and knowledge discovery. The GUMC is comprised of a nationally ranked School of Medicine, School of Nursing & Health Studies, Lombardi Cancer Center and a \$120 million medical research enterprise. It has 710 full-time and 2,167 part-time faculty members from 8 basic science and 16 clinical departments. The NBRF is a non-profit institution chartered in 1960 and dedicated to scientific research in computational medical research.

17 EMBL Member States: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Israel, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom

