The ExPASy proteome WWW server in 2003

Amos Bairoch, Elisabeth Gasteiger, Alexandre Gattiker, Christine Hoogland, Corinne Lachaize, Khaled Mostaguir, Ivan Ivanyi and Ron D. Appel
Swiss Institute of Bioinformatics; 1, Rue Michel Servet, 1211 Geneva 4; Switzerland

Email contacts: amos.bairoch@isb-sib.ch, elisabeth.gasteiger@isb-sib.ch, alexandre.gattiker@isb-sib.ch, christine.hoogland@isb-sib.ch, corinne.lachaize@isb-sib.ch, khaled.mostaguir@isb-sib.ch, ivan.ivanyi@isb-sib.ch, ron.appel@isb-sib.ch

Version of January 22, 2003.
The latest version of this document is available on http://www.expasy.org/doc/expasy.pdf

Introduction

ExPASy [1,2] (the **Ex**pert **P**rotein **A**nalysis **Sy**stem) is a World Wide Web server (http://www.expasy.org) which is provided as a service to the Life Sciences community. Its main focus is on proteins. It provides access to a variety of databases and analytical tools dedicated to what is now known as *proteomics*. It is developed at the **Swiss Institute of Bioinformatics (SIB)** (http://www.isb-sib.ch) by a multidisciplinary team. It first started to operate in August 1993 and has been running without interruption since that date. It seems to have been the first WWW server to be established in the field of life sciences. In December 2002 it had been accessed 242 million times by a total of more than 3 million computer hosts from 185 countries. We describe here what are the information resources and tools available on ExPASy.

Databases

ExPASy is the main host for the following databases that are partially or completely developed in Geneva:

- **The Swiss-Prot knowledgebase** [3,4] (http://www.expasy.org/sprot/); a curated protein sequence database which strives to provide high quality annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and a high level of integration with other databases. Swiss-Prot is supplemented by **TrEMBL** which contains computer-annotated entries for all sequences not yet integrated in Swiss-Prot.
- **SWISS-2DPAGE** [5] (http://www.expasy.org/ch2d/); a database of proteins identified on two-dimensional polyacrylamide gel electrophoresis (2-D PAGE). SWISS-2DPAGE contains data from a variety of human and mouse biological samples as well as from *Arabidopsis thaliana, Escherichia coli*, *Saccharomyces cerevisiae* and *Dictyostelium discoideum*.
- **PROSITE** [6,7] (http://www.expasy.org/prosite/); a database of protein domains and families. PROSITE contains biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs to.
- **ENZYME** [8] (http://www.expasy.org/enzyme/); a repository of information relative to the nomenclature of enzymes.
- **SWISS-3DIMAGE** [9] (http://www.expasy.org/sw3d/); a database of high quality annotated images of biological macromolecules with known three-dimensional structure.
- **SWISS-MODEL Repository** [10] (http://www.expasy.org/swissmod/smrep.html); a database of automatically generated protein structural models.
- **CD40Lbase** [11] (http://www.expasy.org/cd40lbase/); a collection of clinical and molecular data on the CD40 ligand defects leading to Hyper-IgM syndrome.
- **SeqAnalRef** [12] (http://www.expasy.org/seqanalref/); a bibliographic reference database relative to papers dealing with sequence analysis.

A variety of access options are available from the home pages of each of the above databases. These options allow the users to display and retrieve specified subsets of the database. For example, from the home page of Swiss-Prot and TrEMBL, there are options that allow searching by description, accession number, author, citation or by full text search. To complement these options, we have also implemented a SRS [13] server that allows complex searches to be made on any fields of the combination of Swiss-Prot and TrEMBL databases. PROSITE, ENZYME and SWISS-2DPAGE can also be queried using SRS.

A large variety of documents (user's manual, release notes, indices, nomenclature documents, etc.) are available with Swiss-Prot; these documents are all browsable from ExPASy and are enhanced by a variety of hyper-links.

All the databases available on ExPASy are extensively cross-referenced to other molecular biology databases or resources all over the world. For example Swiss-Prot is cross-referenced to more than 50 different databases (such as: EMBL/GenBank/DDBJ, PDB, MEDLINE/PubMed, EcoGene, FlyBase, GeneCards, Genew, Leproma, MaizeDB, MGD, MIM, MypuList, SGD, SubtiList, TubercuList, WormPep, ZFIN, InterPro, Pfam, PRINTS, ProDom, PROSITE, SMART, TIGRFAMs, HSC-2DBASE, HSSP, MEROPS, REBASE, TRANSFAC, etc.).

Swiss-Prot is updated at a frequency of about every two weeks. Most of the other ExPASy databases are frequently updated. Swiss-3DIMAGE, CD40Lbase and SeqAnalRef are no longer maintained.

All the ExPASy databases data and associated documentation files can be copied locally by anonymous FTP (ftp.expasy.org). We also distribute the files to build up a non-redundant and complete protein sequence database (ftp://ftp.expasy.org/databases/sp_tr_nrdb/) consisting of three components: Swiss-Prot, TrEMBL and new entries to be later integrated into TrEMBL (known as TrEMBL_New). These files are completely rebuilt every time Swiss-Prot is updated. They are also available in the "fasta" format used by many sequence similarity search programs such as FASTA and BLAST.

Thanks to hardware provided by HP/Compaq, it is possible to run extremely high-speed BLAST similarity searches on the non-redundant protein sequence database from any Swiss-Prot or TrEMBL entry on ExPASy.

The use of all ExPASy databases is free for academic users. However, we implemented in September 1998 a system of annual subscription fee for commercial users of the Swiss-Prot, PROSITE and SWISS-2DPAGE databases. A new company - **Geneva Bioinformatics** (**GeneBio**) (http://www.genebio.com) - was mandated to conclude the necessary license agreements and to levy the fees. The funds raised are used to bring these databases up to date, to keep them up to date, and to further enhance their quality. Further information on this new funding scheme is available at http://www.expasy.org/announce/

Software tools

We have developed over the years an extensive collection of software tools most of which are either targeted toward the access and display of the databases mentioned above or which are used to analyze protein sequences and proteomics data originating from 2D-PAGE and Mass Spectrometric experiments. These tools can all be accessed from ExPASy:

- **AACompIdent** [14]; identifies a protein by its amino acid composition.
- **AACompSim** [14]; compares the amino acid composition of a Swiss-Prot entry with all other entries in the database.
- **Compute pI/MW**; computes the theoretical isoelectric point (pI) and molecular weight (MW) from a Swiss-Prot or TrEMBL entry or for a user sequence.
- **FindMod** [15]; predicts potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence. Mass differences are used to better characterize the protein of interest.

- **FindPept** [16]; identifies peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage.
- **GlycoMod** [17]; predicts possible oligosaccharide structures that occur on proteins from their experimentally determined masses. This is done by comparing the mass of a potential glycan to a list of pre-computed masses of glycan compositions.
- **NiceProt**; provides a user-friendly tabular view of Swiss-Prot entries. Similar tools are available for PROSITE (**NiceSite** and **NiceDoc**), ENZYME (**NiceZyme**) and SWISS-2DPAGE (**Nice2DPage**).
- **PeptIdent, TagIdent**, **MultiIdent** [18,19,20]; these three related programs identify proteins using a variety of experimental information such as the pI, the MW, the amino acid composition, partial sequence tags and peptide mass fingerprinting data.
- **PeptideCutter,** predicts potential protease cleavage sites and sites cleaved by chemicals in a given protein sequence.
- **PeptideMass** [21]; calculates the theoretical masses of peptides generated by the chemical or enzymatic cleavage of proteins so as to assist in the interpretation of peptide mass fingerprinting.
- **ProtParam**; calculates physico-chemical parameters of a protein sequence such as the composition, the pI, the atomic composition, the extinction coefficient, etc.
- **ProtScale**; computes and represents the profile produced by any amino-acid scale on a selected protein. Some 50 predefined scales are available, the default being the Doolittle and Kyte hydrophobicity scale.
- **RandSeq**; a random protein sequence generator.
- **ScanProsite** [22]; scans a sequence against all of the patterns, profiles and rules in PROSITE or scans a pattern, profile or rule against all sequences in Swiss-Prot and/or TrEMBL.
- **Sulfinator** [23]; a program to predict tyrosine sulfation sites within protein sequences.
- **SWISS-MODEL** [24,25]; an automated knowledge-based protein modelling server. It is able to build models of the three-dimensional structure of proteins whose sequence is closely related to that of proteins with known 3D structure.
- **Swiss-Shop**; an automated sequence alerting system which allows users to obtain new Swiss-Prot entries relevant to their field(s) of interest. Keyword-based and sequence/pattern-based requests are possible. Every time a weekly Swiss-Prot release is performed, all new database entries matching the user-specified search keywords or patterns and the entries showing sequence similarities to the user-specified sequence are automatically sent to the user.
- **Translate**; translates a nucleotide sequence to a protein.

A very important feature of the ExPASy proteomics tools (such as PeptIdent, TagIdent, MultIdent, PeptideMass, FindPept, or FindMod) is that they use the annotations of the Swiss-Prot entries to take into account post-translational modifications as well as sequence variants to perform their predictions.

Some of the above tools (such as SWISS-MODEL, Swiss-Shop or AACompSim) report their results back by email while the others display them directly on-line.

These tools are all listed on a page of ExPASy (http://www.expasy.org/tools/) that also offers links to many other useful programs for the analysis of protein sequences available elsewhere on the Web. We notably have links to the tools provided by our colleagues from the bioinformatics group at ISREC in Lausanne (http://www.isrec.isb-sib.ch). They have developed a BLAST similarity search server, TMpred (to predict transmembrane regions) and interfaces to the SAPS (Statistical Analysis of Protein Sequences), COILS (prediction of coiled coil regions), CLUSTAL and T-COFFEE (multiple sequence alignment) programs.

ExPASy as a portal to other life sciences resources

The mass of information available to life scientists on the Web has completely changed the way that biological data is accessed and processed. It has created many opportunities but also brought new dangers. The most critical problem being the difficulties for researchers to distinguish useful and up-to-date sources of information from sites that provide either ‘fossilized‘ or low quality data. To partially address this problem, we have developed a series of lists and tools:

- **Amos' WWW links page** (http://www.expasy.org/alinks.html); a list that contains links to over a thousand information resources for the life sciences. This list is updated very frequently and is organized in a number of sections that correspond to specific topics.
- **WORLD-2DPAGE** (http://www.expasy.org/ch2d/2d-index.html); a list of all known 2-D PAGE database WWW servers and related services.
- **BioHunt** (http://www.expasy.org/BioHunt/); a service to help search the Internet for molecular biology information. BioHunt is built by Marvin, a software robot which automatically roams the web to search and index life science information. Currently BioHunt indexes more than 20'000 documents.
- **2DHunt** (http://www.expasy.org/ch2d/2DHunt/); a specialized index for 2-D PAGE-related sites.

Other interesting ExPASy features

Biochemical pathways (http://www.expasy.org/cgi-bin/search-biochem-index); an indexed, digitized and clickable version of the Boehringer Mannheim's 'Biochemical Pathways' poster
is available on the server). It allows the user to navigate through the graphical representation of metabolic pathways and is linked to the ENZYME database.

DeepView (formerly **SWISS-PdbViewer**) [25] (http://www.expasy.org/spdbv/); an application that runs on the Microsoft Windows, Mac and Unix (SGI and Linux) platforms and that offers a wide range of options to visualize and manipulate protein structures. It can also be used as a WWW helper application for the display of PDB format entries. DeepView complements the previously described SWISS-MODEL homology-modeling tool. It can be downloaded from ExPASy.

LALNVIEW [26] (http://www.expasy.org/tools/lalnview.html); an application that runs on the Microsoft Windows, Mac and Unix platforms. LALNVIEW is a graphical viewer for pairwise sequence alignments. It can be used to display the results of a pairwise alignment carried out with the SIM software also installed on ExPASy (http://www.expasy.org/tools/sim-prot.html).

2-D PAGE; a wide variety of information concerning 2-D PAGE is available from ExPASy. This includes the full description of experimental protocols as well as an overview of the Melanie 3 2-D PAGE analysis software package. You can also download a 2-D gel viewer.

Protein Spotlight (http://www.expasy.org/spotlight/); a periodical review centered on a specific protein or group of proteins.

Recreational. One must not forget that science can also have a lighter side. So we hope that users will take the time to take a small pause from the hectic pace of modern research and visit **Swiss-Quiz** (http://www.expasy.org/swiss-quiz/) or **Swiss-Jokes** (http://www.expasy.org/cgi-bin/sw-jokes.pl). With Swiss-Quiz one can have a chance to win some Swiss chocolate (real, not virtual!) after having successfully answered a quiz in the field of molecular biology. Swiss-Jokes provides access to a collection of jokes and aphorisms from the fields of life and computer sciences.

Mirror sites

Network congestion is a growing problem. To help address this issue we decided to implement mirror sites of ExPASy in various countries around the world. Such sites will help users to access the ExPASy databases and tools more rapidly in locations that do not have a fast connection to Switzerland. The mirror sites are computers that host exact copies of the information available from the Geneva ExPASy server. They are updated every week.

ExPASy mirror sites are located in academic institutions that have shown an active interest in hosting such sites. As of today seven sites are operational (see below for their addresses and the names of the host institutions).

The ExPASy mirror sites are located in:

Australia
http://au.expasy.org/
at the Australian Proteome Analysis Facility (APAF), Sydney

Bolivia
http://bo.expasy.org/
at the Universidad Católica Boliviana (UCB), Cochabamba

Canada
http://ca.expasy.org/
at the Canadian Bioinformatics Resource (CBR), Halifax

China
http://cn.expasy.org/
at the Center of Bioinformatics, Peking University, Beijing

South Korea
http://kr.expasy.org/
at the Yonsei Proteome Research Center

Taiwan
http://tw.expasy.org/
at the National Health Research Institutes (NHRI), Taipei

United States
http://us.expasy.org/
at the North Carolina Supercomputing Center (NCSC)

Conclusions

The team that develops ExPASy is committed to bring to its users top quality information services in the field of proteomics. We hope that in the next years we will be able to add many new features to those that are already available.

To keep track of new developments on ExPASy do not forget to subscribe to **Swiss-Flash** (http://www.expasy.org/swiss-flash/), a service that allows users to automatically obtain email bulletins that report new and updated ExPASy features. Finally, we want to thank all the users of ExPASy that over the years have sent us feedback that has led to the improvement of existing services and to the development of new ones.

References

[1] Appel R.D., Bairoch A., Hochstrasser D.F.
Trends Biochem. Sci. 19:258-260(1994).

[2] Bairoch A., Appel R.D., Peitsch M.C.
Protein Data Bank Quat. Newsletter 81:5-7(1997).

[3] Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M.
Nucleic Acids Res. 31:354-370(2003).

[4] O'Donovan C., Martin M.-J., Gattiker A., Gasteiger E., Bairoch A., Apweiler R.
Briefings Bioinform. 3:275-284(2002).

[5] Hoogland C., Sanchez J.-C., Tonella L., Binz P.-A., Bairoch A., Hochstrasser D.F., Appel R.D.
Nucleic Acids Res. 28:286-288(2000).

[6] Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C.J., Hofmann K., Bairoch A.
Nucleic Acids Res. 30:235-238(2002).

[7] Sigrist C.J.A. Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A., Bucher P.
Briefings Bioinform. 3:265-274(2002).

[8] Bairoch A.
Nucleic Acids Res. 28:304-305(2000).

[9] Peitsch M.C., Wells T.N.C., Stampf D.R., Sussman J.L.
Trends Biochem. Sci. 20:82-83(1995).
[10] Peitsch M.C.
ISMB 5:234-236(1997).
[11] Notarangelo L.D., Peitsch M.C.
Immunol. Today 17:511-516(1996).
[12] Bairoch A.
Comput. Appl. Biosci. 7:268-268(1991).
[13] Etzold T., Ulyanov A.V., Argos P.
Meth. Enzymol. 266:114-128(1996).
[14] Wilkins M.R., Pasquali C., Appel R.D., Ou K., Golaz O., Sanchez J.C., Yan J.X.,
Gooley A.A., Hughes G., Humphery-Smith I., Williams K.L., Hochstrasser D.F.
Bio/Technology 14:61-65(1996).
[15] Wilkins M.R., Gasteiger E., Gooley A.A., Herbert B.R., Molloy M.P., Binz P.-A.,
Ou K., Sanchez J.-C., Bairoch A., Williams K.L., Hochstrasser D.F.
J. Mol. Biol. 289:645-657(1999).
[16] Gattiker A., Bienvenut W.V., Bairoch A., Gasteiger E.
Proteomics 2:1435-1444(2002).
[17] Cooper C.A., Gasteiger E., Packer N.
Proteomics 1:340-349(2001).
[18] Wilkins M.R., Gasteiger E., Sanchez J.-C., Appel R.D., Hochstrasser D.F.
Curr. Biol. 6:1543-1544(1996).
[19] Wilkins M.R., Gasteiger E., Tonella L., Ou K., Tyler M., Sanchez J.-C., Gooley A.A.,
Walsh B.J., Bairoch A., Appel R.D., Williams K.L., Hochstrasser D.F.
J. Mol. Biol. 278:599-608(1998).
[20] Wilkins M.R., Gasteiger E., Wheeler C., Lindskog I., Sanchez J.-C., Bairoch A.,
Appel R.D., Dunn M.D., Hochstrasser D.F.
Electrophoresis 19:3199-3206(1998).
[21] Wilkins M.R., Lindskog I., Gasteiger E., Bairoch A., Sanchez J.-C., Hochstrasser D.F.,
Appel R.D.
Electrophoresis 18:403-408(1997).
[22] Gattiker A., Gasteiger E., Bairoch A.
Applied Bioinform. 1:107-108(2002).
[23] Monigatti F., Gasteiger E., Bairoch A., Jung E.
Bioinformatics 18:769-770(2002).
[24] Peitsch M.C.
Biotechnology 13:658-660(1995).
[25] Guex N., Peitsch M.C.
Electrophoresis 18:2714-2723(1997)
[26] Duret L., Gasteiger E., Perrière G.
Comput. Appl. Biosci. 12:507-510(1996).