The Path to Enlightenment: Making Sense of Genomic and Proteomic Information

Martin H. Maurer

Department of Physiology and Pathophysiology, University of Heidelberg, 69120 Heidelberg, Germany.

Whereas genomics describes the study of genome, mainly represented by its gene expression on the DNA or RNA level, the term proteomics denotes the study of the proteome, which is the protein complement encoded by the genome. In recent years, the number of proteomic experiments increased tremendously. While all fields of proteomics have made major technological advances, the biggest step was seen in bioinformatics. Biological information management relies on sequence and structure databases and powerful software tools to translate experimental results into meaningful biological hypotheses and answers. In this resource article, I provide a collection of databases and software available on the Internet that are useful to interpret genomic and proteomic data. The article is a toolbox for researchers who have genomic or proteomic datasets and need to put their findings into a biological context.

Key words: proteomics, genomics, Internet databases, web databases, structural bioinformatics

Introduction

The term genomics describes the study of the genome, representing the whole set of genes. Their expression is studied on the DNA or RNA level using high-throughput systems such as microarray technology. Proteomics is defined as the study of the proteome that describes the protein complement of the genome and subsumes the whole set of proteins of a cell, an organ, or the whole organism, respectively. In recent years, there has been a tremendous increase in scientific publications describing proteomic approaches (Figure 1).

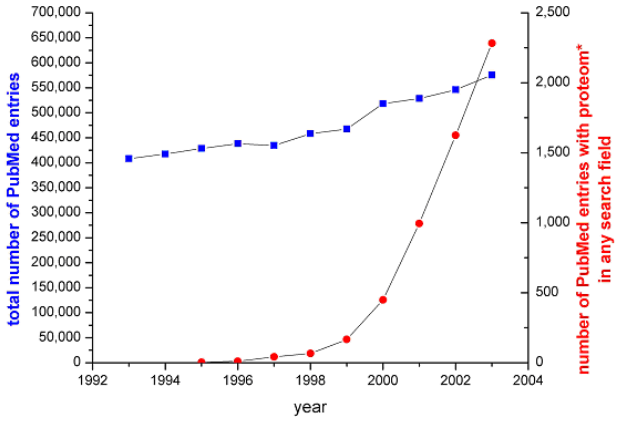

Fig. 1 Since its introduction into the scientific literature in 1995, the term "proteome" saw a tremendous increase in its number of PubMed entries. Whereas there was a steady increase of about 3%–5% in the total number of PubMed entries (squares), the number of publications with the term "proteome" (and its derivations) in any PubMed search field grew exponentially (dots). In 2003, it occurred in almost 2,300 publications, representing about 0.4% of all PubMed entries.

Proteomic analysis includes not only the description of all preferential proteins in a given compartment, but also the set of protein isoforms and modifications, as well as interaction and structural information (*1*). In principle, two main areas in the field of proteomics have been developed, "profiling" and "functional" proteomics (*2*). Proteomic profiling means that the whole set of proteins of a biological sample—which could be an organism, an organ, a tissue, a cell, or an organelle—is indexed and described. Moreover, differential protein expression levels under certain experimental conditions, or different types of materials analyzed, are subsumed under the term of proteomic profiling. In contrast, functional proteomics investigates protein activity, protein interactions, and post-translational modifications.

The typical proteomic experiment consists of three distinct steps, including (1) protein isolation from biological samples, (2) protein separation and fraction-

*** Corresponding author.**
E-mail: maurer@uni-hd.de

ation, and (3) protein quantitation and identification (*3*). Since large information datasets are created in proteomic experiments, proteome informatics involves not only collection, storage, search, analysis, and classification of experimental results, but also integration, management, and retrieval of data from large-scale databases. Thus, the whole proteomic technology is mainly based on computer analysis not only during the step of data acquisition, but also for translating experimental results into meaningful biological answers, which allow accomplishing the generation of new hypotheses.

In this article, I describe bioinformatic tools for help in the analysis of proteomic data, mostly based on the use of the Internet and free access to the academic community. Its purpose is to serve as a link collection and a practical approach how to handle and interpret proteomic data. It is not intended for primary data analysis such as comparison of gel images or identification of mass spectra. Currently available commercial software for two-dimensional electrophoresis gel handling is reviewed by Raman *et al* (*4*). This article is intended as a toolbox for researchers who have primary genomic or proteomic data and need to interpret their findings in the biological context.

Toolbox

This section contains Internet hypertext links organized in a workflow suitable for the analysis of genomic and proteomic data (Figure 2). The URLs of the mentioned links are mainly shown in Table 1.

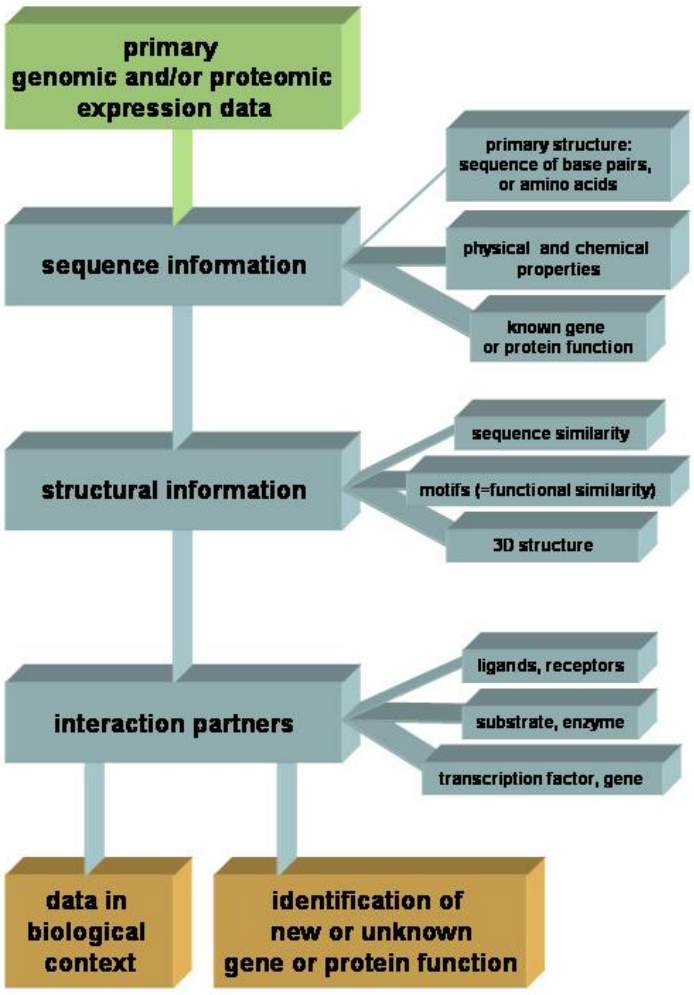

Fig. 2 Basic flowchart of the information retrieval process. The organization of the flowchart is also reflected by the structure of the manuscript and subsumes the suggested proceeding when analyzing genomic or proteomic data.

Sequence information

After creating result lists from genomic or proteomic experiments, researchers are interested in gene and protein functions of the specific items in their result lists. First, the primary gene and protein sequences have to be determined, as they are the basis for structural comparisons. Second, physical and chemical properties of the sequences are retrieved such as molecular weight and isoelectric points for proteins. The theoretical data are then compared to the measured experimental values.

Proposed starting point of the analysis is the **Entrez** resource provided by the National Institutes of Health (NIH) of the USA. It contains mainly molecular data and bibliographic citations. The Entrez concept is to integrate different databases such as GenBank, EMBL, DDBJ for nucleic acids; or Swiss-Prot, PIR, PRF, and PDB for proteins (see below).

The **Ex**pert **P**rotein **A**nalysis **Sy**stem (**ExPASy**) is a proteomics server of the Swiss Bioinformatic Institute (SBI; http://www.isb-sib.ch/) and provides databases and software tools for nearly all aspects of protein analysis. On datasets, for example, Swiss-Prot and TrEMBL are two protein knowledgebases; PROSITE contains protein families and domains (*5*); SWISS-2DPAGE contains reference maps of two-dimensional polyacrylamide gel electrophoresis experiments, protocols and additional software tools. ENZYME is a database of enzyme nomenclature. SWISS-3DIMAGE comprehends 3D images of proteins and other biological macromolecules, whereas SWISS-MODEL is a repository for automatically generated protein models. The GermOnLine knowledgebase holds information about germ cell differentiation. These databases are cross-linked to other information databases such as PRINTS (http://bioinf.man.ac.uk/dbbrowser/PRINTS/; ref. *6*) and BLOCKS (http://blocks.fhcrc.org; ref. *7*, *8*). On software tools, the site includes tools and software packages for identifying and characterizing proteins by mass spectrometric data, translating DNA into protein sequences, searching for sequence similarities, protein sequence patterns and profiles, predicting post-translational modifications and protein topology, analyzing the protein's primary, secondary, or tertiary structure, and aligning sequences (*9*, *10*). Also, tools for the analysis of biological literature have been added.

The journal *Frontiers in Biosciences*, which is thought as virtual library for protein information, provides a link collection for a wide variety of protein analysis tools (http://www.bioscience.org/urllists/protdb.htm).

In the next step, the known gene and protein functions are searched, as the identified functions define the affiliation to specific metabolic or signaling pathways. With regard to a given genomic or proteomic experiment comparing two different biological conditions, this allows finding activated or repressed pathways.

It is not easy to describe all functions of a given protein. For example, the metabolic function of glyceraldehyde-3-phosphate dehydrogenase in glycolysis is well known, but recent years have shown its role in membrane fusion, microtubule bundling, phosphotransferase activity, nuclear RNA export, DNA replication and DNA repair, apoptosis, age-related neurodegenerative disease, prostate cancer and viral pathogenesis (*11*). In this light, one must take into account that protein function databases are never complete. The current scientific literature has to be searched for novel findings.

The **Reactome** project is a collaboration among Cold Spring Harbor Laboratory (http://www.cshl.org/), the European Bioinformatics Institute (http://www.ebi.ac.uk/), and the Gene Ontology (GO) consortium (http://www.geneontology.org/) to develop a curated resource of core pathways and reactions in human biology (*12*). It enables the researcher to locate gene products in a biochemical reaction.

Another program that allows displaying and modifying pathway diagrams is **GenMAPP** (*13*, *14*).

The **Kyoto Encyclopedia of Genes and Genomes (KEGG)** was developed by the Kanehisa Laboratory of the Kyoto University Bioinformatics Center, Japan, in order to create a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic information (*15*).

Once a gene or protein function is found, it is of interest whether the protein is involved in disease processes. **Genecards** is a database of human genes, their products and their involvement in diseases. It offers information about the functions of nearly all human genes (*16*, *17*).

The **Human Protein Reference Database (HPRD)** is provided by the Johns Hopkins University and contains information relevant to the function of human proteins in health and disease. Data pertain

to thousands of protein-protein interactions, post-translational modifications, enzyme/substrate relationships, disease associations, tissue expression, and subcellular localization (*18*, *19*).

Structural information

Once the basic information of a gene or protein is found, several kinds of structural information need to be retrieved. Are there sequence similarities to other known sequences? Does the sequence contain motifs, as these common patterns of nucleotides or amino acids indicate a functional similarity? Third, has the three-dimensional structure of the molecule been determined?

As a starting point, sequence homologies can be found using the **BLAST** (Basic Local Alignment Search Tool, programs BLASTP and PSI-BLAST; ref. *20*). Another search tool for sequence comparison is **SRS** (Sequence Retrieval System; ref. *21*), which is part of the EMBL database and information system.

For searching protein domain families, **Prodom** provides a comprehensive set of protein domain families automatically generated from the Swiss-Prot and TrEMBL sequence databases (*22*).

In the same context, **Pfam** is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. It can be used to view the domain organization of proteins (*23*).

SMART, a Simple Modular Architecture Research Tool, allows the identification and annotation of genetically mobile domains and the analysis of domain architectures (*24*, *25*).

The A. N. Belozersky Institute of Physico-Chemical Biology of the Moscow State University provides multiple alignment tools at http://www.genebee.msu.su/genebee.html.

An Internet resource for visualizing gene splicing from EST data, called **spliceNest**, is provided at http://splicenest.molgen.mpg.de/.

The **Transcript Assembly Program** (**TAP**; ref. *26*) is an EST-based gene finder software that infers the predominant and alternative gene structures in genomic sequences.

Molecular interactions

Structural similarity may lead to new insights to predict protein function from primary, secondary, or tertiary structure. This is highly desirable for proteins with hitherto unknown function, or to discover novel interrelations between proteins.

The Research Collaboratory for Structural Bioinformatics (RCSB) provides the **Protein Data Bank (PDB)**, which contains three-dimensional structures of biological macromolecules. A similar project is hosted by the EMBL-EBI server at http://www.ebi.ac.uk/msd/.

The European Bioinformatics Institute, as part of the European Molecular Biology Laboratories (EMBL) family, provides the **Dali** server, a tool to find neighbors based on the three-dimensional structure of proteins (*27*). The Holm Group at the Institute of Biotechnology, University of Helsinki, Finland, provides a local version of the Dali/FSSP tool as well as other structure-function related computer programs.

The **CATH** program provides a hierarchic classification of protein domain structures (*28*, *29*) and several other structural related software like the Dictionary of Homologous Superfamilies (DHS), the sequence database Gene3D, and Impala, a software package for comparing a single query sequence with a database of PSI-BLAST generated PSSMs.

The **SCOP** database provides a detailed and comprehensive description of the structural and evolutionary relationships between all proteins with known structures. It provides a survey of all known protein folds and detailed information about the close relatives of any particular protein (*30*).

3Dee contains structural domain definitions for all protein chains in the PDB that have 20 or more residues and are not theoretical models (*31*, *32*).

HSSP is a database of Homology-derived Secondary Structures of Proteins and contains information about aligned sequences, secondary structure, sequence variability and sequence profile of known protein structures (*33*).

Phylogenetic trees allow inter-species and inter-molecule sequence comparison along the evolutionary time scale. A set of phylogenetic programs can be found at the websites in Table 1.

Protein superstructures may allow finding similarities between protein families, such as transcription factors and transmembrane domains. **COILS** is a program that compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score (*34*). By comparing this score to the distribution of scores in globular and coiled-coil proteins, the program then calculates the probability

that the sequence will adopt a coiled-coil conformation.

Comparing one-, two-, and three-dimensional structures of genes and proteins is essential for *in silicon* prediction of possible interaction partners. Tools for comparison include **JPRED:TNG**, a consensus method for protein secondary structure prediction; **Jalview**, a Java multiple alignment editor; **PredictProtein**, a service for sequence analysis and structure prediction; **GenThreader** and **3D-PSSM**.

Tools for comparative modeling of protein structures include **COMPOSER**, **DRAGON**, **WhatIf**, a versatile protein structure analysis program that can be used for mutant prediction, structure verification, and molecular graphics, and **SwissModel**.

Once protein structures and possible interactions have been predicted, the models need to be validated. Tools for assessing modeled protein structures can be found in the **Biotech Validation Suite**, **Joy**, and **WhatIf**.

MolTalk (*35*) is a computational environment for structural bioinformatics. It interprets PDB-formatted files and creates an object representation of the structure-chain-residue-atom hierarchy. Thus, the PDB becomes an object-oriented database. **BioWeb** is a collaborative website produced by faculty members from several universities and centers of the University of Wisconsin System. It provides several tools for protein structure analysis, such as modeling algorithms. All information is presented in units and is useful for self-study.

The Sali lab at the University of California at San Francisco provides software for protein modeling and a collection of additional links for programs and Internet servers useful in comparative protein modeling.

The Sean Eddy lab at the Washington University in St. Louis, MO, provides software for hidden Markov models and phylogenetic trees as well as for structural RNA analysis.

Visualization software

Whereas most sequence comparing algorithms and structural bioinformatics software rely on linear description of the nucleotide or amino acid sequences, there are several programs that allow drawing and manipulating three-dimensional molecular structures.

Protein Explorer, formerly known as RasMol, is a molecular visualization software tool for looking at macromolecular structure and its relation to function. From the same site, the web browser visualization plug-in Chime can also be downloaded.

Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service.

ProteinShop is an interactive tool for manipulating protein structures. It was designed to quickly create a diverse set of initial configurations for a given sequence of amino acids. Download is free for academic users.

Miscellaneous links

The following list contains a description of other helpful resources. They are mainly huge collections of integrated web pages with several subcategories of hypertext links. Since these collections are comprehensive and not focused on a single purpose, or group of software functions, they will also contain aspects of the proposed workflow in addition to the discussed items of this compilation.

The Genomics and Bioinformatics Group of the Weinstein lab at the NIH Laboratory of Molecular Pharmacology offers a web site at the URL http://discover.nci.nih.gov useful in genomic and proteomic research. The available software includes **MatchMiner** (*36*), a program that translates among many types of gene and protein identifiers; **GoMiner** (*37*), a tool for the biological interpretation of microarray data using the Gene Ontology databases; **MedMiner** (*38*), a program searching and organizing the PubMed literature on genes and drugs; **CIMminer**, which produces Clustered Image Maps (CIMs; that is, clustered heat maps) of gene expression data; and **MIMminer**, a tool that electronically navigates the Kohn Molecular Interaction Maps. An additional resource, **AbMiner**, is planned as a database of available monoclonal antibodies.

Ron Shamir's group at the School of Computer Sciences at the Tel Aviv University provides bioinformatic resources for cluster analysis of gene expression data, tools for the analysis of gene promoters and the design of degenerate primers, software for the graph realization problem, and programs for the visualization of protein interactions.

With regard to the neurosciences, the Society for Neuroscience of the USA provides a website with a database gateway to databases of experimental data, knowledgebases, and software tools for neuroscience.

A huge collection of bioinformatic links can be found at the website of the Canadian Bioinformatics Workshop Series sponsored by the Canadian Genetic

Diseases Network (CGDN), where many hyperlinks for proteomics can also be found.

Pedro M. Coutinho from the Technical University of Lisbon, Portugal, provides a vast list of molecular biology links also useful for proteome research.

The Bioinformatics and Biocomputing group at the Weizmann Institute of Science, Israel, has collected multiple tools and links at http://bioinformatics.weizmann.ac.il/.

The NIH Computational Molecular Biology section can be found at http://molbio.info.nih.gov/molbio/.

A huge collection of bioinformatic links including software, journal access, technology resources and providers, institutions and careers, is maintained at http://www.bioinformatik.de/ (the website is available in English language).

A resource for links and online protein services can be found at http://www.ch.embnet.org/index.html.

The **PeKing University BIOinformatics Server (PKUBIOS)** at Beijing, China, is a comprehensive collection of bioinformatic network services and can be found at http://mdl.ipc.pku.edu.cn/mirror/mirror.html.

The open directory http://dmoz.org/Science/Biology/Software/ lists a variety of links to software and ongoing projects. The website http://sourceforge.net/ encourages the user to contribute to the development of open access software. Already working projects such as alpha and beta versions can be downloaded. Some of the projects encompass the Biomolecule Naming Service (http://openbns.sourceforge.net) with the main purpose to quickly and easily convert between different name and identifier schemes commonly used for specifying gene and protein sequences.

A huge directory of bioinformatics, genomics, proteomics, biotechnology and molecular biology is provided by the Oxford researcher Dr. Jonathan D. Rees at http://www.bioinformatics.vg/index.shtml. Bioinformatics.Net is a unified gateway to store, search, retrieve and update information about bioinformatics tools, databases and resources.

The Jeffrey Skolnick's research group from the Center of Excellence in Bioinformatics at the University at Buffalo, NY, provides a set of online services including **Prospector** (*39*), a threading approach which defrosts frozen approximation; **TM-score** (*40*), a scoring function for the automated assessment of protein structure template quality; **SAL** (*41*), a Structural Alignment service; the enzyme function prediction tool **EFP-FSSI** (*42*); **SCAR**, a clustering algorithm bundling user-submitted protein structures and computing the representative centroids and substructures of each cluster (*43*); and the keyword search engine **BIOMOLQUEST** for protein structure, function, and classification.

The **Generic Model Organism Project (GMOD)** is a joint effort by the model organism system databases including WormBase for *Caenorhabditis elegans*, FlyBase for *Drosophila melanogaster*, MGI for *Mus musculus*, SGD for *Saccharomyces cerevesiae*, Gramene for grains and grasses, Rat Genome Database for *Rattus norvegicus*, EcoCyc for *Escherichia coli*, and TAIR for *Arabidopsis thaliana* to develop reusable components suitable for creating new community databases of biology.

Table 1 Internet Links for the Analysis of Genomic and Proteomic Information

Name	URL
Sequence Information:	
Entrez	http://www.ncbi.nlm.nih.gov/Entrez/index.html
ExPASy	http://www.expasy.org
Protein analysis tools	http://www.bioscience.org/urllists/protdb.htm
Reactome	http://www.reactome.org
GenMAPP	http://www.genemapp.org
KEGG	http://www.genome.jp/kegg/
Genecards	http://bioinfo.weizmann.ac.il/cards/index.shtml
HPRD	http://www.hprd.org
Structural Information:	
BLAST	http://www.ncbi.nlm.nih.gov/blast/
SRS	http://srs.ebi.ac.uk
Prodom	http://protein.toulouse.inra.fr/prodom.html

Table 1 ***Continued***

Name	URL
Pfam	http://www.sanger.ac.uk/Software/Pfam
	Mirror sites: http://pfam.wustl.edu; http://www.cgr.ki.se/Pfam
SMART	http://smart.embl-heidelberg.de
A. N. Belozersky Institute	http://www.genebee.msu.su/genebee.html
spliceNest	http://splicenest.molgen.mpg.de/
TAP	http://sapiens.wustl.edu/∼zkan/TAP/
Molecular Interactions:	
PDB	http://www.rcsb.org/pdb/; http://www.ebi.ac.uk/msd/
Dali	http://www.ebi.ac.uk/dali/
The Holm Group	http://ekhidna.biocenter.helsinki.fi:8080/templates/software.html
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
3Dee	http://www.compbio.dundee.ac.uk/3Dee
HSSP	http://www.sander.embl-heidelberg.de/hssp/
Phylogenetic programs	http://genome.cs.iastate.edu/supertree/introduction/links.html
	http://evolution.genetics.washington.edu/phylip/software.html
COILS	http://www.ch.embnet.org/software/COILS_form.html
JPRED:TNG	http://www.compbio.dundee.ac.uk/∼www-jpred/
Jalview	http://www2.ebi.ac.uk/∼michele/jalview/contents.html
PredictProtein	http://www.embl-heidelberg.de/predictprotein/predictprotein.html
GenThreader	http://insulin.brunel.ac.uk
3D-PSSM	http://www.bmm.icnet.uk
COMPOSER	http://www-cryst.bioc.cam.ac.uk/
DRAGON	http://www.imtech.res.in/pub/pss/dragon/unix/dragon.html
WhatIf	http://www.cmbi.kun.nl/gv/whatif/
SwissModel	http://www.expasy.ch/swissmod/SWISSMODEL.html
Biotech Validation Suite	http://biotech.embl-heidelberg.de:8400/ or http://biotech.ebi.ac.uk:8400/
Joy	http://www-cryst.bioc.cam.ac.uk/cgi-bin/joy.cgi
MolTalk	http://www.moltalk.org/
BioWeb	http://bioweb.uwlax.edu
The Sali lab	http://salilab.org
The Sean Eddy lab	http://selab.wustl.edu/cgi-bin/selab.pl?mode=software
Visualization Software:	
Protein Explorer	http://www.umass.edu/microbio/rasmol/
Cn3D	http://www.ncbi.nih.gov/Structure/CN3D/cn3d.shtml
ProteinShop	http://proteinshop.lbl.gov//proteinshop-bin/generate.pl?doc=Home
Miscellaneous Links:	
The Genomics and Bioinformatics Group	http://discover.nci.nih.gov
Ron Shamir's group	http://www.cs.tau.ac.il/∼rshamir/
The Society for Neuroscience	http://big.sfn.org/NDG/site/
The Canadian Bioinformatics Workshop Series	http://www.bioinformatics.ca/links_directory/
Pedro M. Coutinho	http://www.public.iastate.edu/∼pedro/research_tools.html
The Bioinformatics and Biocomputing group	http://bioinformatics.weizmann.ac.il/
The NIH Computational Molecular Biology section	http://molbio.info.nih.gov/molbio/

Table 1 ***Continued***

Name	URL
bioinformatic links	http://www.bioinformatik.de/
links and online protein services	http://www.ch.embnet.org/index.html
PKUBIOS	http://mdl.ipc.pku.edu.cn/mirror/mirror.html
open directory	http://dmoz.org/Science/Biology/Software/
Dr. Jonathan D. Rees	http://www.bioinformatics.vg/index.shtml
Jeffrey Skolnick's group	http://www.bioinformatics.buffalo.edu/current_buffalo/skolnick/services.html
GMOD:	
WormBase	http://www.wormbase.org
FlyBase	http://www.flybase.org
MGI	http://www.informatics.jax.org
SGD	http://genome-www.stanford.edu/Saccharomyces/
Gramene	http://www.gramene.org
RatBase	http://rgd.mcw.edu/
EcoCyc	http://ecocyc.org
TAIR	http://www.arabidopsis.org/

Conclusion

The presented hypertext links may serve as entry level for further information retrieval by following the hypertext links in the documents provided. Other publications collected bioinformatic tools available at the Internet and may be consulted for additional resources (*44–46*). Many of the resources described require registration free of charge for academic users. Although I checked all of the hyperlinked sites personally, I do not take any responsibility for the content of hyperlinked pages.

Acknowledgements

I thank Wolfgang Kuschinsky and my colleagues of the Heidelberg Proteome Network, Robert E. Feldmann, Jr., Jens O. Brömme, Heinrich F. Bürgers, Benjamin Gross, and Armin Kalenka for fruitful discussions on proteomic data analysis.

References

1. Tyers, M. and Mann, M. 2003. From genomics to proteomics. *Nature* 422: 193-197.
2. Choudhary, J. and Grant, S.G. 2004. Proteomics in postgenomic neuroscience: the end of the beginning. *Nat. Neurosci.* 7: 440-445.
3. Freeman, W.M. and Hemby, S.E. 2004. Proteomics for protein expression profiling in neuroscience. *Neurochem. Res.* 29: 1065-1081.
4. Raman, B., *et al.* 2002. Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. *Electrophoresis* 23: 2194-2202.
5. Bairoch, A. 1993. The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Res.* 21: 3097-3103.
6. Attwood, T.K., *et al.* 2003. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31: 400-402.
7. Henikoff, J.G., *et al.* 2000. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* 28: 228-230.
8. Henikoff, S., *et al.* 1999. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15: 471-479.
9. Wilkins, M.R., *et al.* 1999. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* 112: 531-552.
10. Hoogland, C., *et al.* 1999. Two-dimensional electrophoresis resources available from ExPASy. *Electrophoresis* 20: 3568-3571.
11. Sirover, M.A. 1999. New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. *Biochim. Biophys. Acta* 1432: 159-184.
12. Joshi-Tope, G., *et al.* 2003. The Genome Knowledgebase: A Resource for Biologists and Bioinformaticists. In *Cold Spring Harbor Symposia on Quantitative Biology,* pp. 237-244. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, USA.

13. Dahlquist, K.D., *et al.* 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* 31: 19-20.
14. Doniger, S.W., *et al.* 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* 4: R7.
15. Kanehisa, M. and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27-30.
16. Safran, M., *et al.* 2002. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* 18: 1542-1543.
17. Rebhan, M., *et al.* 1997. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* 13: 163.
18. Peri, S., *et al.* 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32 (Database issue): D497-501.
19. Peri, S., *et al.* 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13: 2363-2371.
20. Altschul, S.F., *et al.* 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
21. Kulikova, T., *et al.* 2004. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 32 (Database issue): D27-30.
22. Servant, F., *et al.* 2002. ProDom: automated clustering of homologous domains. *Brief Bioinform.* 3: 246-251.
23. Bateman, A., *et al.* 2004. The Pfam protein families database. *Nucleic Acids Res.* 32 (Database issue): D138-141.
24. Letunic, I., *et al.* 2002. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30: 242-244.
25. Schultz, J., *et al.* 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* 95: 5857-5864.
26. Kan, Z., *et al.* 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* 11: 889-900.
27. Holm, L. and Sander, C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* 25: 231-234.
28. Orengo, C.A., *et al.* 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5: 1093-1108.
29. Pearl, F.M., *et al.* 2000. Assigning genomic sequences to CATH. *Nucleic Acids Res.* 28: 277-282.
30. Murzin, A.G., *et al.* 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247: 536-540.
31. Siddiqui, A.S., *et al.* 2001. 3Dee: a database of protein structural domains. *Bioinformatics* 17: 200-201.
32. Dengler, U., *et al.* 2001. Protein structural domains: analysis of the 3Dee domains database. *Proteins* 42: 332-344.
33. Holm, L. and Sander, C. 1999. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* 27: 244-247.
34. Lupas, A., *et al.* 1991. Predicting coiled coils from protein sequences. *Science* 252: 1162-1164.
35. Diemand, A.V. and Scheib, H. 2004. MolTalk—a programming library for protein structures and structure analysis. *BMC Bioinformatics* 5: 39.
36. Bussey, K.J., *et al.* 2003. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.* 4: R27.
37. Zeeberg, B.R., *et al.* 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4: R28.
38. Tanabe, L., *et al.* 1999. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* 27: 1210-1214, 1216-1217.
39. Skolnick, J. and Kihara, D. 2001. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins* 42: 319-331.
40. Zhang, Y. and Skolnick, J. 2004. A scoring function for the automated assessment of protein structure template quality. *Proteins.* In press.
41. Kihara, D. and Skolnick, J. 2003. The PDB is a covering set of small protein structures. *J. Mol. Biol.* 334: 793-802.
42. Tian, W. and Skolnick, J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* 333: 863-882.
43. Betancourt, M.R. and Skolnick, J. 2001. Finding the needle in a haystack: educing native folds from ambiguous *ab initio* protein structure predictions. *J. Comput. Chem.* 22: 339-353.
44. Gottesman, I.I. 2002. A brief guide to Internet addresses for psychiatric genetics and genomics. In *Psychiatric Genetics & Genomics* (ed. McGuffin, P., *et al.*), pp. 461-463. Oxford University Press, Oxford, UK.
45. Yaron, Y. and Orr-Urtreger, A. 2000. Computer resources for the clinical and molecular geneticist. *Methods Mol. Biol.* 132: 291-299.
46. Edwards, Y.J. and Cottage, A. 2001. Prediction of protein structure and function by using bioinformatics. *Methods Mol. Biol.* 175: 341-375.

This work was supported by a grant of the National Genome Research Network (NGFN-2), an initiative of the German Ministry of Education and Research (BMBF).