JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH

How to cite this article:
RAHIM F. BIOINFORMATICS AND PROTEOMIC APPROACHES TO DISEASE: IN VIVO AND IN SILICO PROTEOME ANALYSIS TOOLS. Journal of Clinical and Diagnostic Research [serial online] 2008 June [cited: 2008 June 2];3:879-886
Available from
http://www.jcdr.net/back_issues.asp?issn=0973-709x&year=2008&month= June &volume=2&issue=3&page=879-886&id=181

REVIEW ARTICLE

Bioinformatics And Proteomic Approaches To Disease: In Vivo And In Silico Proteome Analysis Tools

RAHIM F

Abstract

The availability of human genome sequences and transcriptomic, proteomic, and metabolomic data provides us with a challenging opportunity to develop computational approaches for systematic analysis of metabolic disorders. Mass spectrometry represents an important set of in vivo technologies for protein expression measurement. Among them, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI TOF-MS), because of its high throughput and on-chip sample processing capability, has become a popular tool for clinical proteomics. Bioinformatics plays a critical role in the analysis of SELDI data, and therefore, it is important to understand the issues associated with the analysis of proteomic data. A variety of protein sequence databases exist, ranging from simple sequence repositories, which store data with little or no manual intervention in the creation of the records, to expertly curated universal databases that cover all species, and in which the original sequence data are enhanced by the manual addition of further information in each sequence record. As the focus of researchers moves from the genome to the proteins encoded by it, these databases play an even more important role as central comprehensive resources of protein information. In this review, we discuss such issues and the bioinformatics strategies and several leading protein sequence databases used for proteomic in silico analysis technologies associated with in vivo techniques.

Keywords: proteinchip, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS), bioinformatics, proteomic, in vivo, in silico

Corresponding Author: Fakher Rahim Msc. Bioinformatics, Physiology research Center, Ahwaz Jondishapur University of Medical Sciences, Ahwaz, Iran.

Introduction

One of the major goals of the post-genomic era understands the structures, interactions, and functions of all cell proteins. Since the cellular proteome is a dynamic profile, subject to change in response to various signals through posttranslational modification, translocation, and protein-protein and protein-nucleic acid interactions, the task becomes even more complex, looming to a million or more modification events. Proteomics encompasses the study of expressed proteins, including identification and elucidation of the structure-tion interrelationships, which define healthy disease conditions. Information at the level of the proteome is critical to understand the function of the cellular phenotype and its role in health and disease. Since posttranslational events, and indeed, an accurate assessment of protein expression levels cannot always be predicted by mRNA analysis, proteomics, used in concert with genomics, can provide a holistic understanding of the biology underlying the disease process. The challenge in deciphering the proteome is the development and integration of analytical instrumentation combined with bioinformatics, that provide rapid, high-throughput, sensitive, and reproducible tools. Continual advancement in proteome research has led to an influx of protein sequences from a wide range of species, representing a challenge in the field of Bioinformatics. Genome sequencing is also proceeding at an increasingly rapid rate, and this has led to an equally rapid

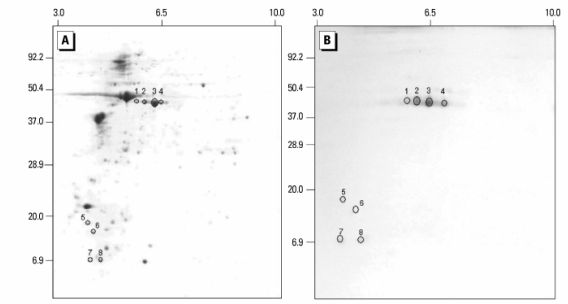
increase in predicted protein sequences. All these sequences, both experimentally derived and predicted, need to be stored in comprehensive, non-redundant protein sequence databases. Moreover, they need to be assembled and analyzed to represent a solid basis for further comparisons and investigations. Especially the human sequences, but also those of the mouse and other model organisms, are of interest for the efforts towards a better understanding of health and disease. An important instrument is the in silico proteome analysis. The term "proteome" is used to describe the protein equivalent of the genome. Most of the predicted protein sequences lack a documented functional characterization. The challenge is to provide statistical and comparative analysis, and structural and other information for these sequences as an essential step towards the integrated analysis of organisms at the gene, transcript, protein, and functional levels. Especially, whole proteomes represent an important source for meaningful comparisons between the species, and furthermore, between individuals of different health states. To fully exploit the potential of this vast quantity of data, tools for in silico proteome analysis are necessary. In this article, some important sources for proteome analysis like sequence databases and analysis tools will be described, which represent highly useful proteomics tools for the discovery of protein function and protein characterization.

In vivo Techniques

Now that the human genome is completed, the characterization of the proteins encoded by the sequence remains a challenging task. The study of the complete protein complement of the genome, the "proteome," referred to as proteomics, will be essential if new therapeutic drugs and new disease biomarkers for early diagnosis are to be developed. Research efforts are already underway to develop the technology necessary to compare the specific protein profiles of diseased versus non-diseased states.

2D gel electrophoresis:

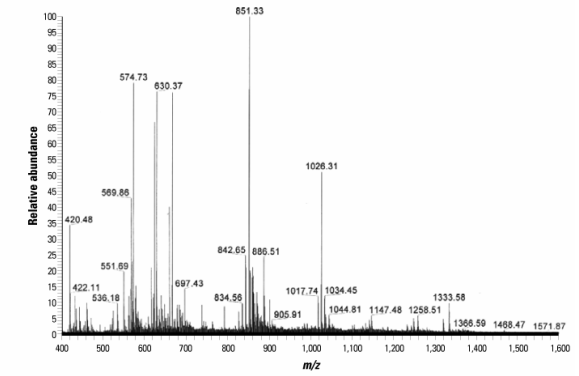
Two-dimensional gel electrophoresis (2DE) is by far, the most widely used tool in proteomics hes for more than 25 years [1]. This ıe involves the separation of complex mıxtures of proteins, first on the basis of isoelectric point (pI) using isoelectric focusing (IEF), and then, in a second dimension, based on molecular mass. The proteins are separated by migration in a polyacrylamide gel. By use of different gel staining techniques such as silver staining [2], Coomassie blue stain, fluorescent dyes [3], or radiolabels, few thousands proteins can be visualized on a single gel. Fluorescent dyes are being developed to overcome some of the drawbacks of silver staining, in making the protein samples more amenable to mass spectrometry [4, 5]. The data can be analyzed with software such as PDQuest by Bio-Rad Laboratories (Hercules, Calif, USA) [6], Melanie 3 by GeneBio (Geneva, Switzerland), Imagemaster 2D Elite by Amersham Biosciences, and DeCyder 2D Analysis by Amersham Biosciences (Buckinghamshire, UK) [7]. Ratio analysis is used to detect quantitative changes in proteins between two samples. 2DE is currently being adapted to high-throughput platforms [8]. Periplaneta americana is the predominant cockroach (CR) species and a major source of indoor allergens in Thailand. Nevertheless, data on the nature and molecular characteristics of its allergenic components are rare. There was a study to identify and characterize the *P. americana* allergenic protein. Two-dimensional gel electrophoresis, liquid chromatography, mass spectrometry, and peptide mass fingerprinting were used to identify the *P. americana* protein containing the MAb-specific epitope that show in figures 1, 2, and 3 [9].



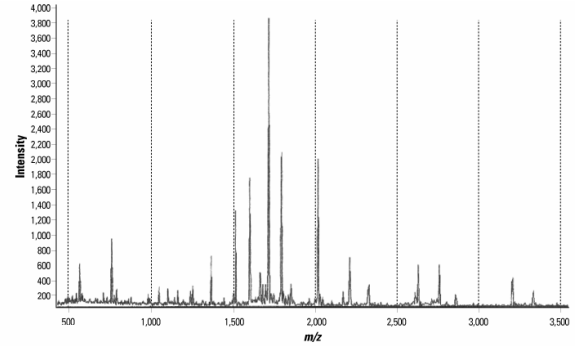
Table/Fig 1 Two-dimensional gel electrophoresis of crude extract of *P. americana* stained by CBB (*A*) and the blot probed with MAb38G6 (*B*). Circles and the indicated numbers in *A* are protein spots that were subjected to LC-MS. Circles in *B* are proteins that reacted to MAb38G6.

ProteinChips:

Unique ionization techniques, such as electrospray ionization and matrix-assisted laser desorption-ionization (MALDI), have facilitated the characterization of proteins by mass



Table/Fig 2. Mass spectra of protein spot 2 (see Figure 1) obtained from LC/MS analysis. The mass spectrometer operated in full-scan mode; the total ion chromatogram was collected over a range of *m/z* 400–1,600.



Table/Fig 3 MALDI-TOF MS profile of a tryptic digest of proteins in the gel plug of spot 2. The protein was digested by trypsin; the peptides were analyzed by MALDI-TOF Pro, and the prominent mass peaks were chosen for database searches.

spectrometry (MS) [10, 11]. Hence,a spectrum is generated with the molecular mass of individual peptides, which are used to search databases to find matching proteins. A minimum of three peptide molecular weights is necessary to minimize false-positive matches. The principle behind peptide mass mapping, is the matching of experimentally generated peptides with those determined for each entry in a sequence. The alternative process of ionization, through the electrospray ionization, involves dispersion of the sample through a capillary device at high voltage [12]. Recent developments have led to the MALDI quadrupole TOF instrument, which combines peptide mapping with peptide sequencing approach [13, 14, 15]. An important feature of tandem MS (MS-MS) analysis is the ability to accurately identify posttranslational modifications such as phosphorylation and glycosylatio, through the measurement of mass shifts. Another MS-based proteinChip technology, surface enhanced laser desorption-ionization time of flight mass spectrometry (SELDI-TOF-MS), has been successfully used to detect several disease-associated proteins in complex biological specimens, such as cell ly 881 minal plasma, and serum [16, 17, 18].

Other technologies:

Arrays of peptides and proteins provide another biochip strategy for parallel protein analysis. Protein assays using ordered arrays have been explored through the development of multipin synthesis [19]. Arrays of clones from phage-display libraries can be probed with antigencoated filters for high-throughput antibody screening [20]. Proteins which are covalently attached to glass slides through aldehyde-containing silane reagents have been used to detect protein-protein interactions, enzymatic targets, and protein small molecule interactions [21]. Other methods of generating protein microarrays are by printing the proteins (ie, purified proteins, recombinant proteins, and crude mixtures) or antibodies using a robotic arrayer and a coated microscope slide in an ordered array. Protein solutions to be measured are labeled by covalent linkage of a fluorescent dye to the amino groups on the proteins [22].

In Silico Techniques

In the growing field of proteomics, tools for the in silico analysis of proteins, and even of whole proteomes, are of crucial importance to make best use of the accumulating amount of data. To utilize this data for healthcare and drug development, first, the characteristics of proteomes of entire species—mainly the human—have to be understood, before differentiation between individuals can be surveyed. Specialized databases about nucleic acid sequences, protein sequences, protein tertiary structure, genome analysis, and proteome analysis, represent useful resources for analysis, characterization, and classification of protein sequences. Different from most proteomics tools focusing on similarity searches like structure analysis and prediction, detection of specific regions, alignments, data mining, 2D PAGE analysis, or protein modeling, respectively, comprehensive databases like the proteome analysis database benefit from the information stored in different databases, and make use of different protein analysis tools to provide computational analysis of whole proteomes.

Protein sequence databases:

In protein sequence databases, information on proteins is stored. Two categories of universal protein sequence databases can be discerned: simple archives of sequence data and annotated databases, where additional information has been added to the sequence record. Especially,

the latter are of interest for the needs of proteome analysis.

PIR, the protein information resource [23] (http://www-nbrf.georgetown.edu/), has been the first protein sequence database which was established in 1984 by the National Biomedical Research Foundation (NBRF), as a successor of the original NBRF Protein Sequence Database. Since 1988, it has been maintained by PIR-International, collaboration between the NBRF, the Munich Information Center for Protein Sequences (MIPS), and the Japan International Protein Information Database (JIPID). The PIR release 71.04 (March 1, 2002) contains 283 153 entries. It presents sequences from a wide range of species, not especially focusing on those of humans.

SWISS-PROT [24] is an annotated protein sequence database established in 1986 and maintained since 1988, collaboratively, by the Swiss Institute of Bioinformatics (SIB) (http://www.expasy.org/) and the EMBL Outstation-The European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/swissprot/). It strives to provide a high level of annotation such as the description of the function of a protein, its domain structure, posttranslational modifications, variants, and so forth, and a minimal level of redundancy. More than 40 cross references— about 4 000000 individual links in total—to other biomolecular and medical databases, such as the EMBL/GenBank/DDBJ international nucleotide sequence database [25], the PDB tertiary structure database [26] or Medline, are providing a high level of integration. Human sequence entries are linked to MIM [27], the "Mendelian Inheritance in Man" database that represents an extensive catalogue of human genes and genetic disorders. SWISSPROT contains data that originates from a wide variety of biological organisms. Release 40.22 (June 24, 2002) contains a total of 110 824 annotated sequence entries from 7459 different species; 8294 of them are human sequences. The annotation of the human sequences is part of the HPI project, the human proteomics initiative [28], which aims at the annotation of all known human proteins, their mammalian orthologues, polymorphisms at the protein sequence level, posttranslational modifications, and at providing tight links to structural information and clustering and classification of all known vertebrate proteins. Seven hundred sixty-one human protein sequence entries in SWISS-PROT contain data relevant to genetic diseases. In these entries, the biochemical and medical basis of the diseases are outlined, as well as information on mutations linked with genetic diseases or polymorphisms, and specialized databases concerning specific genes or diseases, are linked [29].

TrEMBL (translation of EMBL nucleotide sequence database) [24] is a computer-annotated supplement to SWISS-PROT, created in 1996, with the aim to make new sequences available as quickly as possible. It consists of entries in the SWISS-PROT-like format, derived from the translation of all coding sequences (CDSs) in the EMBL nucleotide sequence database, except the CDSs already included in SWISS-PROT. TrEMBL release 21.0 (June 21, 2002) contains 671 580 entries, which should be eventually incorporated into SWISS-PROT, 32 531 of them human. Before the manual annotation step, automated annotation [30, 31] is applied to TrEMBL entries where sensible.

SP TR NRDB (or abbreviated SPTR or SWALL) is a database created to overcome the problem of the lack of comprehensiveness of single-sequence databases: it comprises both the weekly updated SWISS-PROT work release and the weekly updated TrEMBL work release. So SPTR provides a very comprehensive collection of human sequence entries, currently 45 629.

The *CluSTr* (clusters of SWISS-PROT and TrEMBL proteins) database [32] (http://www.ebi.ac.uk/clustr) is a specialized protein sequence database, which offers an automatic classification of SWISS-PROT and TrEMBL proteins into groups of related proteins. The clustering is based on analysis of all pairwise comparisons between protein sequences. Analysis has been carried out for different levels of protein similarity, thereby yielding a hierarchical organization of clusters.

Protein tertiary structure databases:

The number of known protein structures is increasing very rapidly, and these are available through *PDB*, the protein data bank [26] (http://www.rcsb.org/pdb/). There is also a database of structures of "small" molecules of interest to biologists concerned with protein-ligand interactions, available from the Cambridge Crystallographic Data Centre (http://www.ccdc.cam.ac.uk/). In addition, there are also a number of derived databases, which enable comparative studies of 3D structures as

well as to gain insight on the relationships between sequence, secondary structure elements, and 3D structure. *DSSP* (dictionary of secondary structure in proteins, http://www.sander.ebi.ac.uk/dssp/) [33] contains the derived information on the secondary structure and solvent accessibility for the protein structures stored in PDB. *HSSP* (homology-derived secondary structure of proteins, http://www.sander.ebi.ac.uk/hssp/) [34] is a database of alignments of the sequences of proteins with known structure, with all their close homologues. *FSSP* (families of structurally similar proteins, http://www.ebi.ac.uk/dali/fssp/) [35] is a database of structural alignments of proteins. It is based on an all against all comparison of the structures stored in PDB. Each database entry contains structural alignments of significantly similar proteins, but excludes proteins with high sequence similarity, since these are usually structurally very similar.

The *SCOP* (structural classification of proteins) database [36] (http://scop.mrc-lmb.cam.ac.uk/scop/) has been created by manual inspection, and has been abetted by a battery of automated methods. This resource aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds and detailed information about the close relatives of any particular protein. Another database, which attempts to classify protein structures in the PDB, is the *CATH* database [37] (http://www.biochem.ucl.ac.uk/bsm/cath new/), a hierarchical domain classification of protein structures in the PDB.

Proteome analysis databases:

The classic proteomics databases are those of 2D gel electrophoresis data, such as the *SWISS-2DPAGE* database (two-dimensional polyacrylamide gel electrophoresis database) [38] (http://www.expasy.ch/ch2d/). However, since the genome sequencing is proceeding at an increasingly rapid rate, this leads to an equally rapid increase in predicted protein sequences entering the protein sequence databases. Most of these predicted protein sequences are without a documented functional role. The challenge is to bridge the gap until functional data has been gathered through experimental research, by providing statistical and comparative analysis, and structural and other information for these sequences. This way of computational analysis can serve as an essential step towards the integrated analysis of organisms at the gene, transcript, protein, and functional levels. Proteome analysis databases have been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms. The *proteome analysis database* [39] (http://www.ebi.ac.uk/proteome) has the more general aim of integrating information from a variety of sources that will together facilitate the classification of the proteins in complete proteome sets. The proteome sets are built from the SWISS-PROT and TrEMBL protein sequence databases that provide reliable, well-annotated data, as the basis for the analysis. Proteome analysis data is available for all the completely sequenced organisms present in SWISSPROT and TrEMBL, spanning archaea, bacteria, and eukaryotes. In the proteome analysis effort, the Inter- Pro [40] (http://www.ebi.ac.uk/interpro/) and CluSTr resources have been used. Links to structural information databases like the HSSP and PDB are provided for individual proteins from each of the proteomes. A functional classification using gene ontology (GO; [41]) is also available. The proteome analysis database provides a broad view of the proteome data classified according to signatures, describing particular sequence motifs or sequence similarities, and at the same time, affords the option of examining various specific details like structural or functional classification. The *international protein index* (IPI) (http://www.ebi.ac.uk/IPI/IPIhelp.html) provides a top-level guide to the main databases that describe the human and mouse proteome, namely SWISS-PROT, TrEMBL, RefSeq [42], and EnsEMBL [43]. IPI maintains a database of crossreferences between the primary data sources with the aim of providing a minimally redundant, yet maximally complete set of human proteins (one sequence per transcript).

Discussion

The post-genomic era holds phenomenal promise for identifying the mechanistic bases of organismal development, metabolic processes, and disease, and we can confidently predict that bioinformatics research will have a dramatic impact on improving our understanding of such diverse areas as the regulation of gene expression, protein structure determination, comparative evolution, and drug discovery.

Software packages and bioinformatic tools have been, and are being developed to analyze 2D gel protein patterns. These software applications possess user-friendly interfaces that are incorporated with tools for linearization and merging of scanned images. The tools also help in segmentation and detection of protein spots on the images, matching, and editing [44]. Additional features include pattern recognition capabilities and the ability to perform multivariate statistics. New techniques and new collaborations between computer scientists, biostatisticians, and biologists are called for. There is a need to develop and integrate database repositories for the various sources of data being collected, to develop tools for transforming raw primary data into forms suitable for public dissemination or formal data analysis, to obtain and develop user interfaces to store, retrieve, and visualize data from databases, and to develop efficient and valid methods of data analysis.

In the past years, there has been a tremendous increase in the amount of data available concerning the human genome, and more particularly, the molecular basis of genetic diseases. Every week, new discoveries are being made, that link one or more genetic diseases to defects in specific genes. To take into account these developments, the SWISS-PROT protein sequence database, for example, is gradually enhanced by the addition of a number of features that are specifically intended for researchers working on the basis of human genetic diseases, as well as the extent of polymorphisms. The latter are very important too, since they may represent the basis for differences between individuals, which are particularly interesting for some aspects of medicine and drug research. Such comprehensive sequence databases are mandatory for the use of proteome analysis tools, like the proteome analysis database which combines the different protein sequences of a given organism to a complete proteome. This proteome can be regarded as a whole new unit, analyzable according to different points of view (like distribution of domains and protein families, and secondary and tertiary structures of proteins), and can be made comparable to other proteomes. In general, for using the proteomics data for healthcare and drug development, first, the characteristics of proteomes of entire –mainly the human— have to be understood before secondly differentiation between individuals can be surveyed. But although the number of proteome analysis tools and databases is increasing, and most of them are providing a very good quality of computational efforts and/or annotation of information, the user should not forget that automated analysis always can hold some mistakes. Data material in databases is reliable, but only to a certain point. Automatic tools which use data derived from databases can thus be error-prone, rules built on their basis can be wrong, and sequence similarities can occur due to chance and not due to relationship. Users of bioinformatics tools should in no way feel discouraged in their using, provided they keep in mind the potential pitfalls of automated systems and even of humans, be encouraged to check all data as far as possible, and not blindly rely on them.

References

[1] O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. J Biol Chem. 1975; 250(10):4007–4021.

[2] Merril CR, Switzer RC, Van Keuren ML. Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. Proc Natl Acad SciUSA. 1979; 76(9):4335–4339.

[3] patton WF. Making blind robots see: the synergy between fluorescent dyes and imaging devices in automated proteomics. Biotechniques. 2000; 28(5):944–957.

[4] Steinberg TH, Jones LJ, Haugland RP, Singer VL. SYPRO orange and SYPRO red protein gel stains:one-step fluorescent staining of denaturing gels for detection of nanogram levels of protein. AnalBiochem. 1996; 239(2):223–237.

[5] Chambers G, Lawrie L, Cash P, Murray GI. Proteomics: a new approach to the study of disease. JPathol. 2000; 192(3):280–288.

[6] Bergman AC, Benjamin T, Alaiya A, et al. Identification of gel-separated tumor marker proteins by mass spectrometry. Electrophoresis. 2000; 21(3):679–686.

[7] Chakravarti DN, Chakravarti B, Moutsatsos I. Informatic tools for proteome profiling. Biotechniques.2002; 32(Suppl):4–15.

[8] Lopez MF, Kristal BS, Chernokalskaya E, et al. High-throughput profiling of the mitochondrialproteome using affinity fractionation and automation. Electrophoresis. 2000; 21(16):3427–3440.

[9] Sookrung N, Chaicumpa W, Tungtrongchitr A, Vichyanond P, Bunnag C, Ramasoota P, Tongtawe P, Sakolvaree Y, Tapchaisri P. Periplaneta americana Arginine Kinase as a

Major Cockroach Allergen among Thai Patients with Major Cockroach Allergies. Environmental Health Perspectives.2006;114:875-880

[10] Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem. 1988; 60(20):2299– 2301.

[11] Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal Chem. 1991; 63 (24):1193A–1203A.

[12] Andersen JS, Mann M. Functional genomics by mass spectrometry. FEBS Lett. 2000; 480(1):25–31.

[13] Krutchinsky AN, Zhang W, Chait BT. Rapidly switchable matrix-assisted laser desorption/ionization and electrospray quadrupole-time-of-flight mass spectrometry for protein identification. J Am Soc Mass Spectrom. 2000; 11(6):493–504.

[14] Shevchenko A, Loboda A, Shevchenko A, Ens W, Standing KG. MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research. Anal Chem. 2000; 72(9):2132–2141.

[15] Merchant M,Weinberger SR. Recent advancements in surface-enhanced laser esorption/ionizationtime of flight-mass spectrometry. Electrophoresis. 2000; 21(6):1164–1177.

[16] Wright Jr GL, Cazares LH, Leung SM, et al. Proteinchip_ surface enhanced laser desorption ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. Prostate Cancer Prostatic Dis. 1999; 2(5- 6):264–276.

[17] Vlahou A, Schellhammer PF, Mendrinos S, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol. 2001; 158(4):1491–1502.

[18] Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res. 2002; 62(13):3609–3614.

[19] Geysen HM, Meloen RH, Barteling SJ. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. Proc Natl AcadSci USA. 1984; 81(13):3998–4002.

[20] De Wildt RM, Mundy CR, Gorick BD, Tomlinson IM. Antibody arrays for high-throughput screening of antibody-antigen interactions. Nat Biotechnol. 2000; 18(9):989–994.

[21] Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V, Mirzabekov A. Protein microchips:use for immunoassay and enzymatic reactions. Anal Biochem. 2000; 278(2):123–131.

[22] Haab BB, Dunham MJ, Brown PO. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biol. 2001; 2(2):1–13.

[23] Wu CH, Huang H, Arminski L, et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res. 2002; 30(1):35–37.

[24] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000; 28(1):45–48.

[25] Stoesser G, Baker W, van den Broek A, et al. The EMBL nucleotide sequence database. Nucleic Acids Res. 2001; 29(1):17–21.

[26] Bhat TN, Bourne P, Feng Z, et al. The PDB data uniformity project. Nucleic Acids Res. 2001; 29(1):214–218.

[27] Pearson PL, Francomano C, Foster P, Bocchini C, Li P, McKusick VA. The status of online Mendelian inheritance in man (OMIM) medio 1994. NucleicAcids Res. 1994; 22(17):3470–3473.

[28] O'Donovan C, Apweiler R, Bairoch A. The human proteomics initiative (HPI). Trends Biotechnol. 2001; 19(5):178–181.

[29] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J MolMed. 1997; 75(5):312–316.

[30] Fleischmann W, Moller S, Gateau A, Apweiler R. A novel method for automatic functional annotation of proteins. Bioinformatics. 1999; 15(3):228–233.

[31] Kretschmann E, Fleischmann W, Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISSPROT. Bioinformatics. 2001; 17(10):920–926.

[32] Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R. CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. Nucleic Acids Res. 2001; 29(1):33–36.

[33] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogenbonded and geometrical features. Biopolymers. 1983; 22(12):2577–2637.

[34] Dodge C, Schneider R, Sander C. TheHSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res. 1998; 26(1):313–315.

[35] Holm L, Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res. 1996; 24(1):206–209.

[36] Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2002:

refinements accommodate structural genomics. Nucleic Acids Res. 2002; 30(1):264–267.
[37] Pearl FMG, Martin N, Bray JE, et al. A rapid classification protocol for the CATH Domain Database to support structural genomics. Nucleic Acids Res. 2001; 29(1):223–227.
[38] Hoogland C, Sanchez JC, Tonella L, et al. The 1999 SWISS-2DPAGE database update. Nucleic Acids Res. 2000; 28(1):286–288.
[39] Apweiler R, Biswas M, Fleischmann W, et al. Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. Nucleic Acids Res. 2001; 29(1):44–48.
[40] Apweiler R, Attwood TK, Bairoch A, et al. The Inter- Pro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001; 29(1):37–40.
[41] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25– 29.
[42] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res.2001; 29(1):137–140.
[43] Hubbard T, Barker D, Birney E, et al. The Ensemble genome database project. Nucleic Acids Res.2002; 30(1):38–41.
[44] Ohler U, Harbeck S, Niemann H, Noth E, Reese MG. Interpolated Markov chains for eukaryotic promoter recognition. Bioinformatics. 1999; 5(5):362–369.