Published online 7 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D771–D779***
doi:10.1093/nar/gkl784

MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes

Yanling Zhang[1,2], Yong Zhang[1,2], Jun Adachi[1,3], Jesper V. Olsen[1], Rong Shi[1], Gustavo de Souza[1], Erica Pasini[4], Leonard J. Foster[5], Boris Macek[1], Alexandre Zougman[1], Chanchal Kumar[1], Jacek R. Wiśniewski[1], Wang Jun[2,6] and Matthias Mann[1,*]

[1]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany, [2]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China, [3]Graduate School of Global Environmental Studi, Kyoto University, Yoshida-Honmachi Sakyo-Ku, Kyoto 606-8501, Japan, [4]Biomedical Primate Research Centre, Lange Kleiweg 139, 2288 GJ Rijswijk, The Netherlands, [5]Department of Biochemistry and Molecular Biology, Centre for Proteomics, University of British Columbia, Vancouver, BC V6T 1Z4, USA and [6]Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense M, Denmark

Received August 15, 2006; Revised September 22, 2006; Accepted September 29, 2006

ABSTRACT

Mass spectrometry (MS)-based proteomics has become a powerful technology to map the protein composition of organelles, cell types and tissues. In our department, a large-scale effort to map these proteomes is complemented by the Max-Planck Unified (MAPU) proteome database. MAPU contains several body fluid proteomes; including plasma, urine, and cerebrospinal fluid. Cell lines have been mapped to a depth of several thousand proteins and the red blood cell proteome has also been analyzed in depth. The liver proteome is represented with 3200 proteins. By employing high resolution MS and stringent validation criteria, false positive identification rates in MAPU are lower than 1:1000. Thus MAPU datasets can serve as reference proteomes in biomarker discovery. MAPU contains the peptides identifying each protein, measured masses, scores and intensities and is freely available at http://www.mapuproteome.com using a clickable interface of cell or body parts. Proteome data can be queried across proteomes by protein name, accession number, sequence similarity, peptide sequence and annotation information. More than 4500 mouse and 2500 human proteins have already been identified in at least one proteome. Basic annotation information and links to other public databases are provided in MAPU and we plan to add further analysis tools.

INTRODUCTION

The availability of genome sequences, in conjunction with spectacular advances in mass spectrometric (MS) technology for protein identification have now made it possible to quickly determine large numbers of proteins in complex mixtures (1–5). One early application of MS-based proteomics has been the mapping of various proteomes—that is, the identification of their constituent proteins.

Partial proteomes of microorganisms have been reported, for instance the malaria parasite proteome in various stages of its life cycle (6,7) and international consortia are studying the liver and brain proteome in mice and men. The proteomes of body fluids, such as the plasma proteome, the urinary proteome and many others may have potential diagnostic utility. The proteins expressed in specific cell types and cell lines provide clues to functions of these cells and are useful resource for researchers employing them as models. Finally, 'organellar proteomes' are the proteins constituting subcellular structures such as mitochondria or non-membrane enclosed structures such as the nucleolus (8,9).

Despite its obvious utility, proteome mapping faces several technological and some conceptual challenges. Because of the finite dynamic range and sequencing speed of MS, it is difficult to exhaustively map proteomes with the current state of technology (10). Therefore, proteomes will remain

*To whom correspondence should be addressed. Tel: +49 89 8578; Fax: +49 89 8578 2219; Email: mmann@biochem.mpg.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joined first Authors

'in progress' for some time. Proteomes are not static (i.e. body fluid proteomes change with the state of the organism), organellar proteomes vary between cell types (11) and generally as a function of cell state (12). Biochemical purification of an organelle is never 100% successful, and additional steps need to be incorporated into the proteomic analysis to distinguish genuine members of the proteome from co-purifying ones. For these and other reasons, constructing databases of proteomes is not as straightforward as constructing sequence databases and proteome databases have to include much additional information concerning the technology employed in mapping and the state of the proteome. Of more immediate concern for proteome database construction is the fact that MS technology can mis-identify proteins, particularly when low-resolution technology is employed (2). Anderson *et al*. (13) have noted that four studies of the blood plasma proteome identified together 1175 proteins but only had an overlap of 4% between them.

We have embarked on the mapping of a large number of different proteomes. We employ high resolution mass spectrometry and peptide masses are typically measured within a few p.p.m. Typically, proteins have to be identified with at least two peptides, or peptides have to have been sequenced by MS^3 [two subsequent stages of mass spectrometry, (14)]. Identified peptides generated by enzymatic cleavage have to obey strict enzyme specificity. For our reference proteomes, criteria are chosen such that a search against a nonsense database consisting of reversed sequence entries (15) indicates error rates of less than one in a thousand. Our goal is to eventually cover most important organelles, cell types and tissues as well as body fluids in MAPU, accompanied by a set of unified analysis tools.

DATA GENERATION AND VALIDATION

Our typical work flow to map a proteome is as follows (Figure 1). Protein mixtures are obtained by homogenization of tissue, centrifugation of a body fluid, lysis of cultured cells or sub-cellular fractionation. These mixtures are then solubilized in SDS buffer and subjected to one dimensional gel electrophoresis. 1D gels are Coomassie stained and cut into ~10–20 gel slices. These slices are in-gel digested with trypsin or endoproteinase Lys-C and peptides are extracted (16). Resulting peptide mixtures are automatically loaded onto a 75 μm chromatography column and eluted using a 2 h gradient. The MS platform used for nearly all proteomes reported in MAPU is the linear ion trap—Fourier transform ion cyclotron resonance mass spectrometer (LTQ-FT ICR MS) or the linear ion trap—orbitrap (LTQ-Orbitrap). Both instruments are manufactured by Thermo Electron and are capable of extremely high mass accuracy (17,18) very fast sequencing speeds and very high sensitivity. Several hundred of the top scoring peptides are then used as internal calibrants to remove any systematic mass errors by recalibrating all masses (7,10). Resolution of peptide spectra (termed survey or MS^1 scans) is typically 50 000–100 000, which is sufficient to resolve nearly all co-eluting peptides. The n most intense ions—with n typically 2–10 depending on the complexity of the proteome—are fragmented and detected in the linear ion trap, while the high resolution survey spectrum is acquired in the FTICR or orbitrap part of the instrument.

Data are processed and analyzed using an automated pipeline involving peak recognition, database search with the Mascot search engine (19), and optional manual validation using MSQuant, an open-source program developed by our group and available at http://msquant.sourceforge.net.

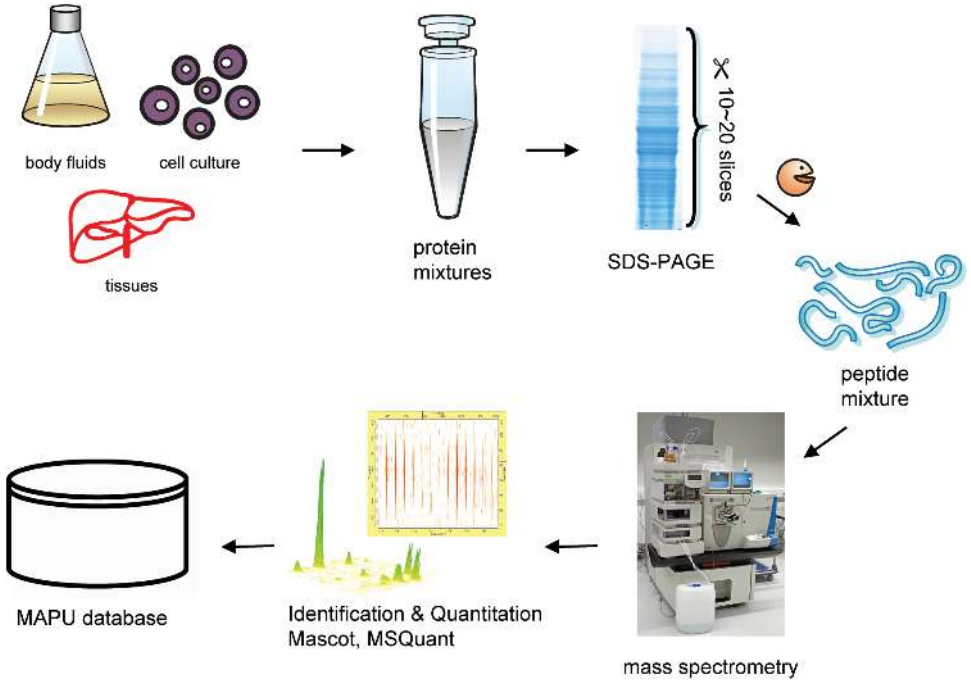

Figure 1. Workflow for protein identification and validation.

We typically search all MS data against the International Protein Index (IPI) database (20), which we find to be an acceptable compromise between inclusiveness (presence of most protein coding sequences) and redundancy (several entries for the same protein). Depending on the project, datasets are joined and checked for overlap using ProteinCenter, a proteomics software suite developed by Proxeon Biosystems (http://www.proxeon.com). ProteinCenter also allows us to directly distinguish which isoforms of a protein are present in a proteome, provided that distinguishing peptide sequences have been detected.

For proteins entered into a MAPU reference proteome, very stringent identification criteria are applied. It is not possible to fix these criteria permanently for all projects because technology is evolving rapidly, because different amounts of manual validation are employed and because different projects deal with different proteome and database sizes. The following parameters are typical: (i) Peptides must have a length of at least seven amino acids; (ii) mass accuracy after recalibration is typically better than 2 p.p.m. and no peptides with mass deviation greater than 5 p.p.m. are allowed; (iii) depending on the mass accuracy achieved and the complexity of the proteome, different cut-offs for the Mascot peptide database scores are required; (iv) we require that peptides strictly obey enzymatic cleavage specificity (17). Generally, at least two peptides are required for unambiguous protein identification. In addition to a minimum Mascot score for each peptide, total protein identification score is also required. This minimum protein score is normally set so that a search of a decoy database does not result in any hits. This assures that the reported proteome contains no or very few false positives. Single peptide identifications are allowed only in special cases, when supplementary information is available. Currently, this supplementary information can be provided by a second stage of peptide fragmentation (MS^3), by a high resolution and high mass accuracy fragmentation spectrum recorded in the orbitrap or by the sequencing of SILAC pairs [SILAC is a method employed in quantitative proteomics and stands for Stable Isotope Labeling with Amino acids in Cell culture, (21)]. MAPU also estimates a *P*-value for the identification for each protein where possible.

SUB-DATABASE ACCESS AND CONTENT

MAPU consists of several sub-databases containing different proteomes. Common templates are used in creation of the databases. All data are created in a single laboratory facilitating a common data standard, user interface and user experience. At the top level MAPU is organized into four branches: body fluids, tissues, cell types, and organelles. Some of these branches contain a clickable map to access the relevant proteomes.

Body fluid databases

Body fluids are of special interest in proteomics because they are an easily obtainable source of biomarkers. To discover such biomarkers involves measurement of the proteomes of healthy and diseased individuals and the construction of a comprehensive protein catalog for each fluid is a necessary first step. MAPU attempts to provide a 'gold standard' set of reference proteomes for diagnostically important body fluid proteomes from normal individuals. So far, the following human body fluids have been analyzed by advanced proteomics methods in our department: Plasma, urine, cerebrospinal fluid (CSF), tear fluid, saliva, seminal fluid and breast milk.

Urinary proteome database

Urine is the second most important diagnostic body fluid and we performed an in depth analysis from healthy donors (22). The MAPU urinary proteome database (http://www.mapuproteome.com/urine) contains 1543 proteins measured with extremely high confidence, a much higher number than all previous urinary proteome studies combined. Textbook knowledge suggests that proteins of MW > 45 kDa should be retained by the glomerular barrier in the kidney, but we found many proteins with MW of >100 kDa. Furthermore, we found a large percentage of membrane proteins, usually migrating at the apparent MW expected for the full length protein. These proteins are probably present in urine embedded in small vesicles. While this is the largest high stringency body fluid proteome to date, our previous model study on the yeast proteome suggests that improved methods should be able to at least double the number of proteins detectable in urine in the future (10).

Tear fluid database

Recently, tear fluid has become a subject of interest due to several reports that demonstrate differences in protein content in disease states such as diabetic dry-eye syndrome and Sjogren's syndrome (23–25). Tear fluid was collected from a single donor and analyzed in depth by MS^3 on an LTQ-FT instrument and MS^2 on an LTQ-Orbitrap instrument (26). Combined analysis resulted in the high confidence identification of 491 proteins (http://www.mapuproteome.com/tear). Notable features of this proteome are the high number of proteins involved in protection against oxygen and pathogens.

Seminal fluid database

Seminal fluid is a little studied but important body fluid. It has buffering properties and contains many distinct proteins, contributing to the functioning and survival of spermatozoa and is thus crucial to successful fertilization. The human seminal fluid proteome data may thus be useful in fertility related research and may also be a starting point for future quantitative analysis in diseases including prostate and testis cancer. We used LTQ-FTICR with MS^3 to identify a total of 923 proteins in seminal fluid from a single individual (27). The seminal fluid database is located at http://www.mapuproteome.com/seminal.

Milk database

Breast milk was collected from a human volunteer and separated into water soluble, fat soluble and insoluble fractions and each fraction analyzed by 1D gel electrophoresis followed by MS (J.V. Olsen, G.M. Sowa *et al*. manuscript in preparation). A total of more than 500 proteins were identified. Apart from clinical use, this knowledge may be helpful in designing infant formula.

Milk is the first body fluid proteome in which we have analyzed a post-translational modification on a large scale. The

milk proteome was digested and phospho-peptides were enriched using titanium dioxide beads in the presence of 2,5-dihydroxy benzoic acid (DHB) according to a published protocol (28). Many proteins were found to be phosphoproteins. This information is also contained in MAPU.

Cerebrospinal fluid database

Produced in amounts averaging 500 ml per day, CSF provides buoyancy and protection to the brain and spinal cord. CSF can be obtained by lumbar puncture and is the only readily accessible fluid in direct contact with the brain, carrying proteins, protein fragments and regulatory peptides from perfused tissues. Consequently, there exists an enormous interest in using CSF in diagnostics of neurodegenerative, inflammatory, psychiatric and neoplastic disorders. We analyzed indepth a CSF sample of one person, and a CSF sample pooled from five individuals. The MAPU CSF database (http://www.mapuproteome.com/csf) contains 798 proteins identified with near certainty in these samples (A. Zougmann, B. Pilch *et al*., manuscript in preparation). We also characterized the low molecular weight range of the CSF proteome. The 'peptidome' turned out to contain a number of known and novel factors. The neuropeptides were investigated bioinformatically for the presence of characteristic features such as N- and C-terminal pro-hormone enzymatic cleavage sites, cysteine content, C-terminal amidation, and N-terminal pyroglutamination. Several of the peptides were found to be post-translationally modified by *O*-linked glycosylation and phosphorylation.

Cell type and cell line database

The human body consists of more than 200 different cell types, each with its distinct function, morphology and protein expression pattern. An example of a cell type represented in the MAPU database is the red blood cell proteome described below.

Cell lines are the main experimental model of a large community of researchers. These are cells originally isolated from human donors or animals and subsequently transformed or otherwise immortalized. They may serve as generic cells (for example, HeLa cells are used to study basic cell biology in microscopy studies), they are chosen for desirable experimental attributes (HEK 293 cells for transfectability), or they are selected to recapitulate essential aspects of the tissue of interest (3T3-L1 mouse fat cells for adipose tissue or HEPA 1-6 hepatocytes for the liver). It is important to know which proteins are actually expressed in the model cell employed. For example, when performing protein interaction studies, only binding partners that are actually expressed in the cell line can be found.

Human Red Blood Cell Database

The human Red Blood Cell Database (hRBCD, http://www.mapuproteome.com/rbc) contains 587 proteins, divided into membrane and soluble proteins (29). In additional, we also provide information ranging from identification of specific isoforms to the class, metabolic status of identified proteins and categorization by sub-cellular localization. MAPU supplies information on the biochemical characteristics of the membrane proteins and related statistical peptide information. The hRBCD can be used in confirming the presence of a protein in the red blood cell and to obtain further information on specific proteins and their biochemical behavior.

3T3-L1 cell line proteome database

We describe the proteome of a mouse fat cell line, 3T3-L1, as an example of a cell line proteome. 3T3 L1 cells are adipocyte precursors that are differentiated over the course of eight days to adipocytes. They are the most common model in adipocyte biology. In the Diabetes Genome Anatomy Project, (http://www.diabetesgenome.org/) microarray studies have already been performed. We fractionated differentiated 3T3-L1 cells into nuclear, microsomal (membrane), mitochondrial and cytosolic fractions and analyzed each of them by 'GeLCMS'. Together we found more than 3287 unique proteins (J. Adachi, C. Kumar *et al*. submitted) (http://www.mapuproteome.com/adipo). From this and the microarray information we conclude that the fat cell proteome is likely to have a very complex proteome of more than 5000 different proteins. This is a further indication that fat cells are not just passive storage container for lipids but participate in a network of complex regulatory functions.

Organellar proteomes

The eukaryotic cell has a sophisticated sub-cellular organization, which is a main subject of cell biology. Best known are the membrane-enclosed organelles in the cytoplasma: Golgi apparatus, endoplasmic reticulum, mitochondria and others. The nucleus also has organelles but the mechanisms for maintaining their organization are less clear. We have previously described the mouse mitochondrial proteome in various tissues (11). Furthermore,—in collaboration with the Angus Lamond laboratory—we have characterized the human nucleolar proteome. These data are already accessible in the nucleolar database at http://www.lamondlab.com/nopdb/ (12,30).

Organellar databases—and other proteome databases to a lesser degree—have another source of erroneous identification besides the potential for misidentification by MS. Even when the protein is correctly determined by MS, it may just have been co-purifying in the organellar preparations and may not be a genuine member of the organellar proteome. When determining the proteome of the centrosome, the microtubule organizing center of animal cells and a structure involved in chromosome segregation, we distinguished centrosomal proteins from co-purifying ones using an algorithm called Protein Correlation Profiling (PCP), in which the sedimentation profile of an organellar protein distinguishes it from unrelated 'background' proteins (31). We briefly describe the Organellar Map Database (ORMD) in the next section, a part of MAPU that contains the results of applying the same principle to all membrane enclosed organelles in mouse liver tissue.

Organellar Map Database

Protein localization to membrane-enclosed organelles is a central feature of cellular organization. Using a linear ion-trap Fourier transform mass spectrometer (LTQ-FTICR MS, Thermo Finnigan) combined with the PCP method (31), we identified 2197 proteins in mouse liver homogenate (32). Peptides identifying these proteins were quantified over 32

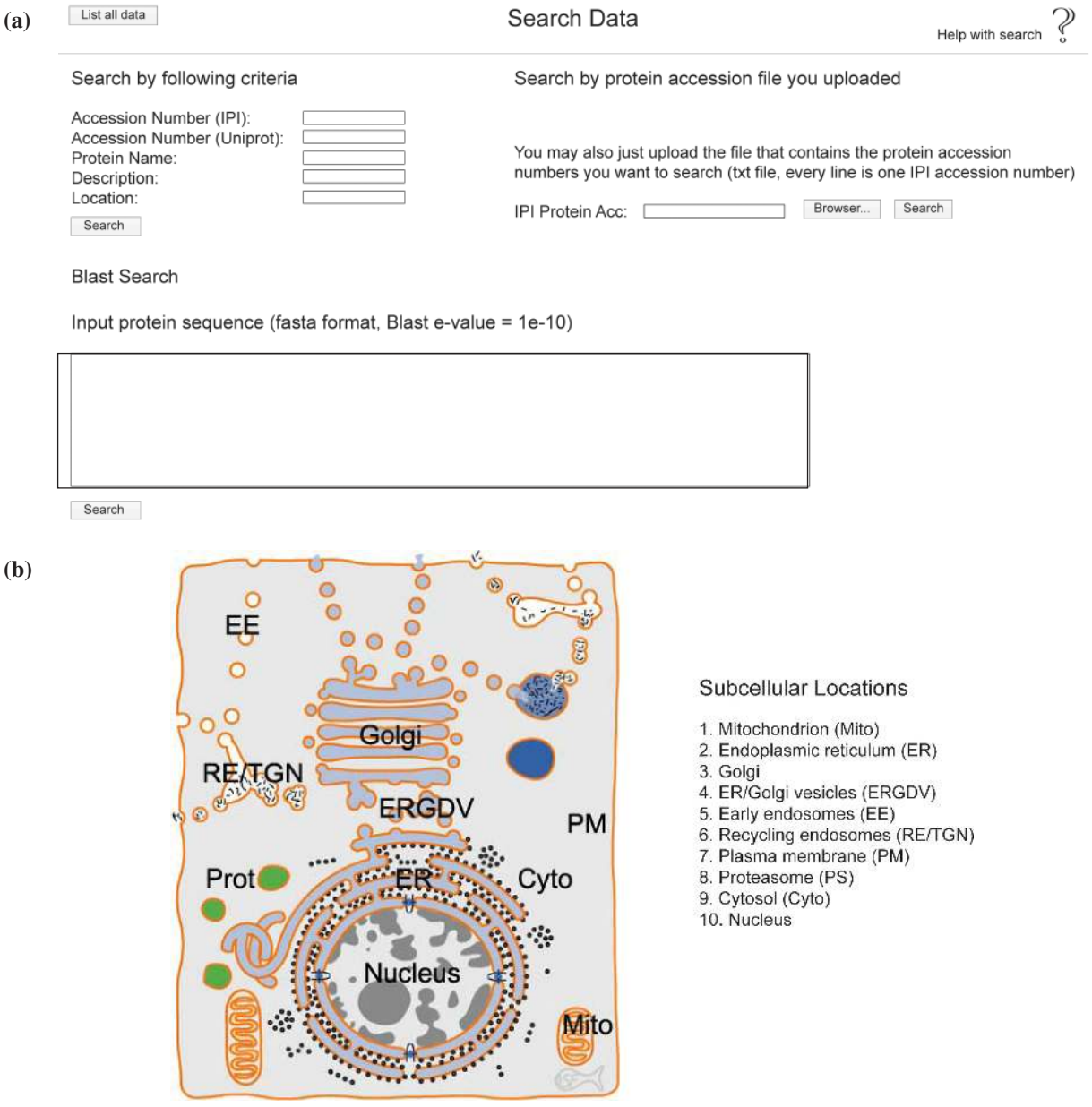

Figure 2. Four search facilities in the MAPU database. (**a**) Search sections in ORMD. At the top of the section, there is a button 'List all data' for list-query. The left side is the advanced search query section, which includes several search terms. Some of these are specific to different sub-databases. The right side is so-called batch search module, but only in ORMD. The bottom of search section is BLAST search section. The input protein sequence should be in fasta format and *E*-value is 1e−10. (**b**) Cell sub-cellular map in ORMD. The picture is clickable and presents sub-cellular location name when the mouse is moving over them. We also list all selectable sub-cellular locations in ORMD on the right side. User can also click them directly to go to the protein list report. The same idea will be applied to other sub-databases, such as body fluid database, in the future.

centrifugation fractions. Ten subcellular locations were apparent in the data and over 1400 proteins were mapped to them (32). The ORMD (http://www.mapuproteome.com/ormd/) lists all the identified proteins with their sub-cellular localization. ORMD presents basic information for the identified mouse liver proteins. (i) protein name, IPI accession, Uniprot accession and protein description, (ii) sub-cellular locations of proteins, (iii) information of identified peptides which support proteins, including peptide sequence, peptide delta score, location on proteins, (iv) a link of the protein to the IPI and Uniprot (33) databases. As in the other databases, homology search is supported using protein sequences by a BLAST module (34). The user may upload an IPI accession list to do a batch search and obtain all relevant information in the ORMD database. ORMD currently archives 2197 proteins in total.

Interestingly, 39% of all proteins with localization information belong to more than one organelle. Data in ORMD comes from endogenous proteins in tissue and therefore is a good complement to localization information obtained by overexpressed fluorescent fusion proteins in cell lines. Information in ORMD has also been transferred to the MINT

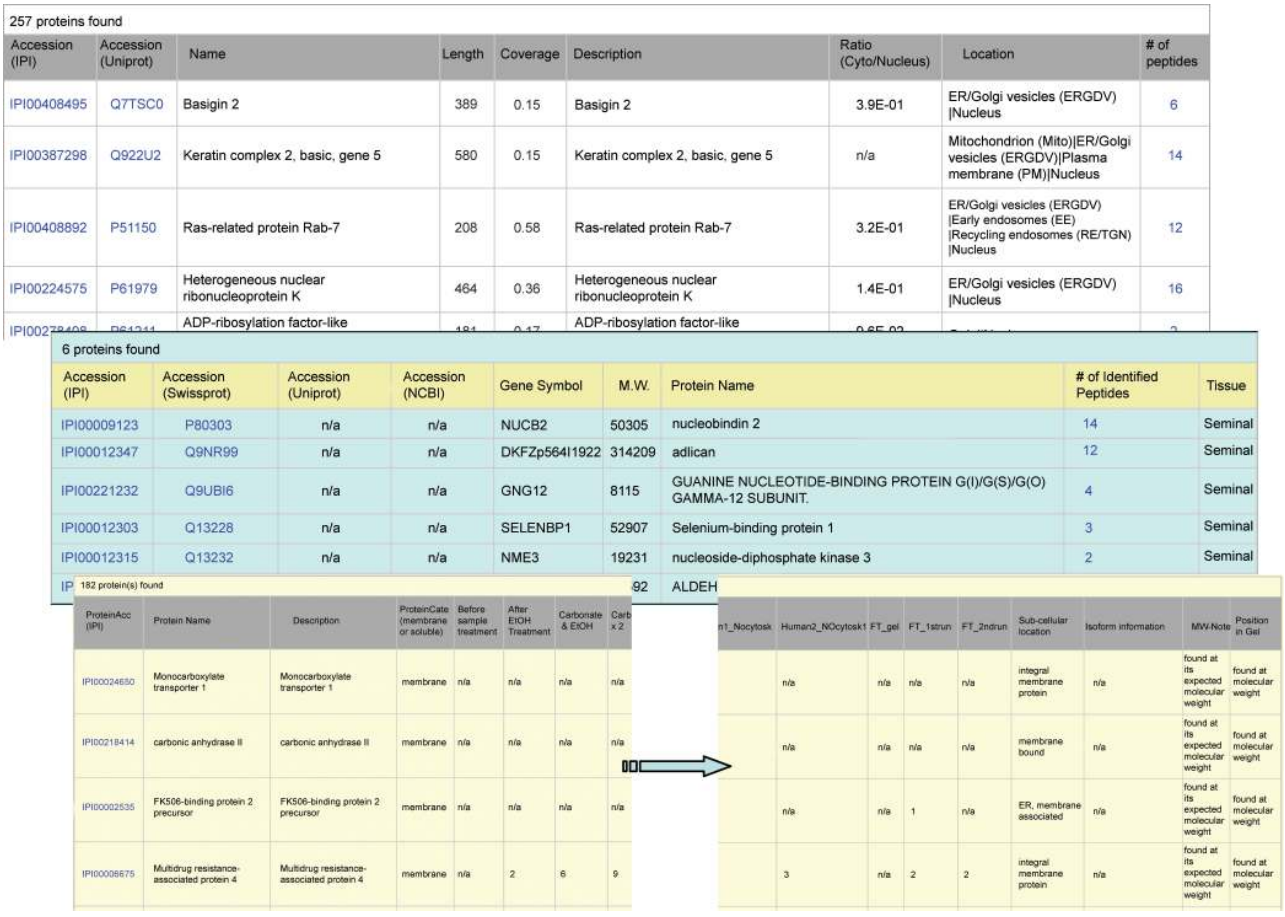

Figure 3. Report pages in the MAPU database. Protein list report page (in ORMD, Seminal fluid Database and Red Blood Cell Database). All proteins in our proteome database are hyperlinked in the BLAST result page. User can navigate to the protein report in the relevant sub database using these hyperlinks.

protein interaction database (35). Integration of proteomic and genomic data enabled identification of networks of co-expressed genes, *cis*-regulatory motifs, and putative transcriptional regulators involved in organelle biogenesis (11,32,36). Large-scale sub-cellular proteomics thus ties biochemistry, cell biology and genomics into a common framework for organelle analysis.

WEB QUERY/SEARCH INTERFACE

We strove to develop search interfaces that are sufficiently flexible to handle the different sub-databases, allow easy retrieval of information and—as much as possible—enforce a common user interface across the databases. MAPU provides four search facilities (Figure 2): list-query button, advanced search query section, BLAST search section, and clickable figure-mapped search section. Usually queries are performed in the sub-databases but a common search interface for all MAPU databases and for each of the four main branches will also be provided.

List-query button

At the top of each search section, we provide quick access to all of the proteome data in the database, using a button termed 'List all data'. This operation retrieves all identified proteins with relevant information for each individual database in a protein list report page. This simple module provides access to all the data for integrative bioinformatic analysis by other researchers.

Advanced search query section

The sub-databases differ in terms of the type of data and they support project-depended information. We provide advanced search query sections with the information that the proteome data have as the search terms (Figure 2a). Thus specialized query interfaces have been built for different databases with commonly-used search terms. (e.g. IPI accession number and protein name, particularly with protein description, peptide sequence in the seminal fluid database, biochemical methods and sub-cellular localization in the red blood cell database, and UniProt accession number, sub-cellular location in ORMD.) In ORMD, we also provide a batch search module (Figure 2a). Users can use this module to retrieve a protein list report page with a series of IPI accession numbers. Two report pages are implemented: protein list report and peptide list report (Figure 3). All protein IPI accession numbers (and UniProt accession numbers) have been linked to IPI (and UniProt) database so that users can obtain more

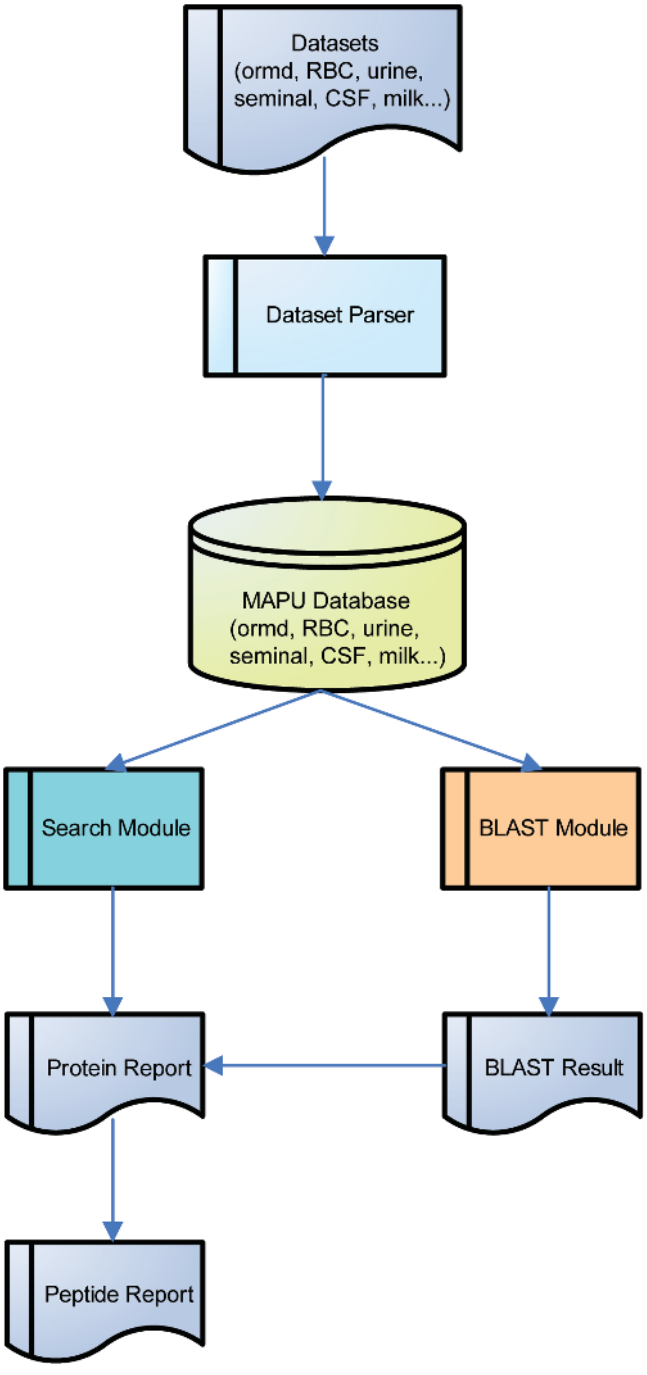

Figure 4. Data work flow in our MAPU database. To generate the branches databases more easily and flexibly, we developed dataset parsers/database generators as assistant tools. The whole work flow is from original data to data in database with several functional modules by our parser, generator tools and common templates.

annotation information from those public databases. In the seminal fluid database, the protein list report page contains protein IPI accession, Swissprot accession, gene symbol, molecular weight, number of identified peptides and tissue name; the peptide list report page contains peptide sequence, modification, MS^2 score, MS^2 + MS^3 score, MS^3 precursor, peptide length, and delta mass (difference between calculated and measured mass in p.p.m.). In the red blood cell database, the protein list report page contains protein IPI accession, protein name, protein description, protein category, biochemical methods, sub-cellular location and some other relevant information. In ORMD, the protein list report page contains IPI accession, UniProt accession, protein name, protein description, sub-cellular location, number of identified peptides and related information. The peptide list report page shows peptide sequence and peptide score, and associated proteins with relevant protein information, such as protein IPI accession, UniProt accession, name, length, description, sub-cellular location. The peptide position in the proteins is also displayed.

BLAST search section

The protein sequences can be searched via BLAST (Figure 2a) to identify sequence similarities between user-submitted sequences and the protein sequences in MAPU (*E*-value 1e−10). For user friendliness, all proteins in our proteome database are hyperlinked in the BLAST result report.

Clickable figure-mapped search section

In the ORMD database, we have a picture of a generic cell and users can directly click different sub-cellular locations to search proteins located there (Figure 2b).

SYSTEM DESIGN AND IMPLEMENTATION

The MAPU database adopts the popular Browser/Server model and consists of a database server and an application server. The application system is based on the open source MySQL relational database and runs on the open source Tomcat application server. JSP and Java Bean are served as client requests.

We also developed some common templates that can be used for generating branch databases. For example, the common search module and the BLAST search module are implemented from templates. With a view to increase search performance, we created different instances for different databases, and indexes on some search parameters. We also implemented scripts to obtain statistics from all branch databases.

The data work flow in the MAPU database is illustrated in Figure 4. A database specialist (Y.Z.) and the proteomics researchers that create the original data agree on a set of data items to be reported in the database. This information is used to create formats that are kept as similar as possible across projects. Dataset parsers are then created to extract the relevant information from those datasets and automatically upload them to the database. All of the above searches functionalities give users a protein report with detail information. In most protein list report pages, the item 'number of identified peptides' is linked to the peptide list report with details.

PERSPECTIVE

We have created a family of proteome databases, which we hope will become an important resource for biology and biomedicine. We chose to integrate only our own data to escape the error accumulation effect currently seen in the

proteomic literature. Our data can be freely downloaded and incorporated into more universal databases such as UniProt and SwissProt, as long as their source is acknowledged. Similarly, some of the data has already been uploaded into the MINT database and dedicated proteomics databases such as PeptideAtlas (37) and PRIDE (38)

The next version of the MAPU database will (i) include more body fluid proteomes and cell line proteomes (ii) have comparative analysis, summary and difference between different body fluid samples, e.g. GO-categories distribution and comparison (39,40), (iii) incorporate more complex annotation information for each protein from other informative database, (iv) integrate more powerful search modules, (v) provide a platform for user to compare their own data with our public proteome data, (vi) alternatively, if necessary, combine relevant genome information as a reference. In summary, the MAPU database provides a powerful and user friendly resource for scientists to explore the proteome.

ACKNOWLEDGEMENTS

We thank the database group at the Beijing Genomics Institute for help, discussion and providing database templates. Funding to pay the Open Access publication charges for this article was provided by the Max-Planck Society for the Advancement of Science.

Conflict of interest statement. None declared.

REFERENCES

1. Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
2. Steen,H. and Mann,M. (2004) The abc's (and xyz's) of peptide sequencing. *Nature Rev. Mol. Cell Biol.*, **5**, 699–711.
3. Sadygov,R.G., Cociorva,D. and Yates,J.R. (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature Methods*, **1**, 195–202.
4. Peng,J. and Gygi,S.P. (2001) Proteomics: the move to mixtures. *J. Mass. Spectrom*, **36**, 1083–1091.
5. Link,A.J., Eng,J., Schieltz,D.M., Carmack,E., Mize,G.J., Morris,D.R., Garvik,B.M. and Yates,J.R., III (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.*, **17**, 676–682.
6. Florens,L., Washburn,M.P., Raine,J.D., Anthony,R.M., Grainger,M., Haynes,J.D., Moch,J.K., Muster,N., Sacci,J.B., Tabb,D.L. *et al.* (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature*, **419**, 520–526.
7. Lasonder,E., Ishihama,Y., Andersen,J.S., Vermunt,A.M., Pain,A., Sauerwein,R.W., Eling,W.M., Hall,N., Waters,A.P., Stunnenberg,H.G. *et al.* (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature*, **419**, 537–542.
8. Yates,J.R., 3rd, Gilchrist,A., Howell,K.E. and Bergeron,J.J. (2005) Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.*, **6**, 702–714.
9. Andersen,J.S. and Mann,M. (2006) Organellar proteomics: from inventory to insight. *EMBO Rep.*, **7**, 874–879.
10. de Godoy,L.M., Olsen,J.V., de Souza,G.A., Li,G., Mortensen,P. and Mann,M. (2005) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.*, **7**, R50.
11. Mootha,V.K., Bunkenborg,J., Olsen,J.V., Hjerrild,M., Wisniewski,J.R., Stahl,E., Bolouri,M.S., Ray,H.N., Sihag,S., Kamal,M. *et al.* (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, **115**, 629–640.
12. Andersen,J.S., Lam,Y.W., Leung,A.K., Ong,S.E., Lyon,C.E., Lamond,A.I. and Mann,M. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77–83.
13. Anderson,N.L., Polanski,M., Pieper,R., Gatlin,T., Tirumalai,R.S., Conrads,T.P., Veenstra,T.D., Adkins,J.N., Pounds,J.G., Fagan,R. *et al.* (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell Proteomics*, **3**, 311–326.
14. Olsen,J.V. and Mann,M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl Acad. Sci. USA*, **101**, 13417–13422.
15. Elias,J.E., Haas,W., Faherty,B.K. and Gygi,S.P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods*, **2**, 667–675.
16. Shevchenko,A., Wilm,M., Vorm,O. and Mann,M. (1996) Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. *Anal. Chem.*, **68**, 850–858.
17. Olsen,J.V., Ong,S.E. and Mann,M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell Proteomics*, **3**, 608–614.
18. Olsen,J.V., de Godoy,L.M., Li,G., Macek,B., Mortensen,P., Pesch,R., Makarov,A., Lange,O., Horning,S. and Mann,M. (2005) Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell Proteomics*, **4**, 2010–2021.
19. Perkins,D.N., Pappin,D.J., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
20. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
21. Ong,S.E., Blagoev,B., Kratchmarova,I., Kristensen,D.B., Steen,H., Pandey,A. and Mann,M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*, **1**, 376–386.
22. Adachi,J., Kumar,C., Zhang,Y., Olsen,J.V. and Mann,M. (2006) The human urinary proteome contains more than 1500 proteins including a large proportion of membranes proteins. *Genome Biol.*, **7**, R80.
23. Grus,F.H. and Augustin,A.J. (2001) High performance liquid chromatography analysis of tear protein patterns in diabetic and non-diabetic dry-eye patients. *Eur. J. Ophthalmol.*, **11**, 19–24.
24. Johnson,M.E. and Murphy,P.J. (2004) Changes in the tear film and ocular surface from dry eye syndrome. *Prog. Retin. Eye Res.*, **23**, 449–474.
25. Kassan,S.S. and Moutsopoulos,H.M. (2004) Clinical manifestations and early diagnosis of Sjogren syndrome. *Arch. Intern. Med.*, **164**, 1275–1284.
26. de Souza,G.A., Godoy,L.M. and Mann,M. (2006) Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors. *Genome Biol.*, **7**, R72.
27. Pilch,B. and Mann,M. (2006) Large-scale and high-confidence proteomic analysis of human seminal plasma. *Genome Biol.*, **7**, R40.
28. Larsen,M.R., Thingholm,T.E., Jensen,O.N., Roepstorff,P. and Jorgensen,T.J. (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell Proteomics*, **4**, 873–886.
29. Pasini,E.M., Kirkegaard,M., Mortensen,P., Lutz,H.U., Thomas,A.W. and Mann,M. (2006) In-depth analysis of the membrane and cytosolic proteome of red blood cells. *Blood*, **108**, 791–801.
30. Leung,A.K., Trinkle-Mulcahy,L., Lam,Y.W., Andersen,J.S., Mann,M. and Lamond,A.I. (2006) NOPdb: Nucleolar Proteome Database. *Nucleic Acids Res.*, **34**, D218–D220.
31. Andersen,J.S., Wilkinson,C.J., Mayor,T., Mortensen,P., Nigg,E.A. and Mann,M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, **426**, 570–574.
32. Foster,L.J., de Hoog,C.L., Zhang,Y., Xie,X., Mootha,V.K. and Mann,M. (2006) A mammalian organelle map by protein correlation profiling. *Cell*, **125**, 187–199.
33. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
34. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

35. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
36. Mootha,V.K., Lepage,P., Miller,K., Bunkenborg,J., Reich,M., Hjerrild,M., Delmonte,T., Villeneuve,A., Sladek,R., Xu,F. *et al*. (2003) Identification of a gene causing human cytochrome *c* oxidase deficiency by integrative genomics. *Proc. Natl Acad. Sci. USA*, **100**, 605–610.
37. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. and Aebersold,R. (2006) The PeptideAtlas project. *Nucleic Acids Res*., **34**, D655–D658.
38. Jones,P., Cote,R.G., Martens,L., Quinn,A.F., Taylor,C.F., Derache,W., Hermjakob,H. and Apweiler,R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res*., **34**, D659–D663.
39. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.
40. Ye,J., Fang,L., Zheng,H., Zhang,Y., Chen,J., Zhang,Z., Wang,J., Li,S., Li,R. and Bolund,L. (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res*., **34**, W293–W297.