

PRIDE: new developments and new datasets

Philip Jones^{1,*}, Richard G. Côté¹, Sang Yun Cho², Sebastian Klie³,
Lennart Martens¹, Antony F. Quinn¹, David Thorneycroft¹ and Henning Hermjakob¹

¹EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, ²Department of Biochemistry, Yonsei Proteome Research Center and Biomedical Proteome Research Center, Yonsei University, Seoul, Korea and ³Martin Luther University, Halle-Wittenberg, Halle-Saale, Germany

Received September 19, 2007; Revised October 24, 2007; Accepted October 27, 2007

ABSTRACT

The PRIDE (<http://www.ebi.ac.uk/pride>) database of protein and peptide identifications was previously described in the NAR Database Special Edition in 2006. Since this publication, the volume of public data in the PRIDE relational database has increased by more than an order of magnitude. Several significant public datasets have been added, including identifications and processed mass spectra generated by the HUPO Brain Proteome Project and the HUPO Liver Proteome Project. The PRIDE software development team has made several significant changes and additions to the user interface and tool set associated with PRIDE. The focus of these changes has been to facilitate the submission process and to improve the mechanisms by which PRIDE can be queried. The PRIDE team has developed a Microsoft Excel workbook that allows the required data to be collated in a series of relatively simple spreadsheets, with automatic generation of PRIDE XML at the end of the process. The ability to query PRIDE has been augmented by the addition of a BioMart interface allowing complex queries to be constructed. Collaboration with groups outside the EBI has been fruitful in extending PRIDE, including an approach to encode iTRAQ quantitative data in PRIDE XML.

INTRODUCTION

The PRIDE database has been developed to provide a standards-compliant repository for mass-spectrometry-based proteomics data comprising identifications of proteins, peptides and post-translational modifications, together with the mass spectra that provide evidence for these identifications. PRIDE has previously been

described in both Proteomics (1) and in the 2006 NAR Database Special Edition (2) that should be referred to for a description of the PRIDE data structure. PRIDE has been reviewed by Mead and co-workers as one of the most important proteomics data repositories in the field (3), and the specific infrastructure in PRIDE to support data privacy and anonymous peer reviewing has been well received by journals (4).

Several other databases exist for the purpose of capturing and disseminating proteomics data, some of which provide their own data analysis pipelines. The Global Proteome Machine Database (GPMDB) provides data from the GPM servers to support the validation of peptide MS/MS spectra and protein coverage patterns (5). PeptideAtlas provides a publicly accessible compendium of identifications from MS/MS that have been processed through PeptideProphet to provide a uniform score (6). The Human Proteinpedia (<http://www.humanproteinpedia.org>) is a portal for community annotation that is used as an addendum to the expert curated Human Protein Reference Database (HPRD) (7). The Open Proteomics Database (OPD) is a public database of mass spectrometry-based proteomics data (8). Tranche provides a secure distributed file system that is designed to handle the sharing of massive datasets (<http://www.proteomecommons.org/dev/dfs/>). PRIDE makes use of Tranche to allow the sharing of massive data files, currently including search engine output files and binary raw data from mass spectrometers that can be accessed via a hyperlink from PRIDE. The public proteomics data repositories are now poised to focus on collaboration and data sharing, through membership of the ProteomExchange consortium (9). This will allow each repository to share their data in a collaborative fashion while remaining independent and able to focus their efforts as they see fit.

This article describes the new data available in PRIDE, the infrastructure being developed to support submission to PRIDE and additions to the user interface that are intended to improve the utility of PRIDE as a query and analysis tool.

*To whom correspondence should be addressed. Tel: +44 1223 492610; Fax: +44 1223 494484; Email: pjones@ebi.ac.uk

Some of the new developments in PRIDE address long-standing requirements that have been present since the inception of the PRIDE project, such as the need for more sophisticated and user-configurable query that can return results in multiple file formats. Other developments, such as the PRIDE Wizard developed at the University of Manchester address requirements that are becoming increasingly important in proteomics, such as the use of quantitative labeling.

At the same time, proteomics journals are exerting increasing pressure to submit data supporting proteomics journal submissions to public repositories in standard formats (10).

DATABASE DESCRIPTION

New data content

A significant number of new public datasets are available in PRIDE. The majority of recent PRIDE submissions have accompanied corresponding journal submissions, indeed submission to PRIDE has regularly been used to support manuscript submissions. The accepted practice is to maintain data privacy until publication of the manuscript. PRIDE supports this by providing a mechanism for private data submission with the option of generating a random username and password that will grant access to the private dataset. This anonymous login can then be passed to peer reviewers, allowing them confidential access to the details of the proteomics experiment supporting the manuscript under review.

At the time of writing, PRIDE comprises 2703 public experiments out of a total of 3185 submitted experiments (85% of the total being public). All of the experiments in PRIDE are organized into projects to improve the accessibility of the data, via both the PRIDE browse page and the PRIDE BioMart.

The complete set of data in PRIDE covers 25 different species including several model animal and plant species as well as fungi, bacteria and viruses (Figure 1). Direct access to PRIDE data organized by species, tissue, sub-cellular location, disease state and project name can be obtained via the 'Browse Experiments' menu item. PRIDE makes use of the NEWT taxonomy (11) to annotate species and selected ontologies from the OBO Foundry (<http://obofoundry.org/>) that cover specific domains to annotate other aspects of the sample type. These include the BRENDA tissue ontology (12), the Cell Type ontology (13) and the Gene Ontology (14) to annotate sub-cellular location. Disease state is annotated using the Disease Ontology (<http://diseaseontology.sourceforge.net/#overview>).

As may be expected, the majority of identifications of proteins and peptides come from human samples, comprising 81% of the peptide identifications in PRIDE (48% of unique peptide identifications in PRIDE). Apart from species-specific studies, PRIDE also contains an interesting dataset describing the protein content of an environmental community sample, the *acid mine drainage* dataset described in Nature (15), in which a community genomic dataset was searched to generate

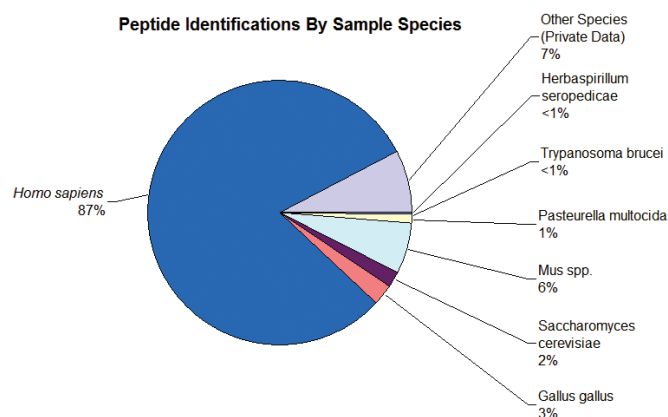


Figure 1. PRIDE includes data for 25 different species. This pie chart illustrates the representation of each species in terms of the total number of peptide identifications for each species.

strain-specific protein identifications for the resident acidophilic bacterial biofilm.

The inception of the PRIDE project was in part inspired, and indeed funded, by the HUPO Plasma Proteome Project (HPPP). Since the HPPP pilot phase has been completed, two other collaborative HUPO projects have also contributed significant datasets to PRIDE, including mass spectra. The European HUPO Brain Proteome Project (HBPP) (16) has submitted 5555 protein identifications to PRIDE, based upon 154 132 peptide identifications in both human and mouse brain tissue. The Beijing Proteome Research Center has submitted 32 421 separate protein identifications, based upon 299 869 peptide identifications in liver, as part of their contribution to the international HUPO Liver Proteome Project (HLPP) (17,18).

Part of the work of the PRIDE team has been to contribute to the development of data conversions into PRIDE XML format from commonly used proteomic analysis pipelines. A specific example of this is the large set of data supporting an analysis of the proteome of human cerebrospinal fluid (19). This dataset includes 890 identified proteins based upon 49 185 peptide identifications. The analysis for this dataset was performed using the Trans-Proteomic Pipeline (TPP) (20) originally developed by the Institute for Systems Biology. The output formats from the TPP (mzXML for mass spectra, pepXML for peptide identifications and protXML for protein identifications) were parsed and the publicly available PRIDE core API was used to convert this data into PRIDE XML format. This work has since served as the basis for importing other TPP-formatted data into PRIDE.

The large number and variety of bioinformatics data resources available from the EBI can be daunting, however using and linking these resources can add value to complex datasets. This has been achieved with the recent submission from the Cellzome research team, reporting the interaction of protein kinases with small inhibitory molecules (21). For this dataset, bi-directional links to the IntAct database of molecular interactions

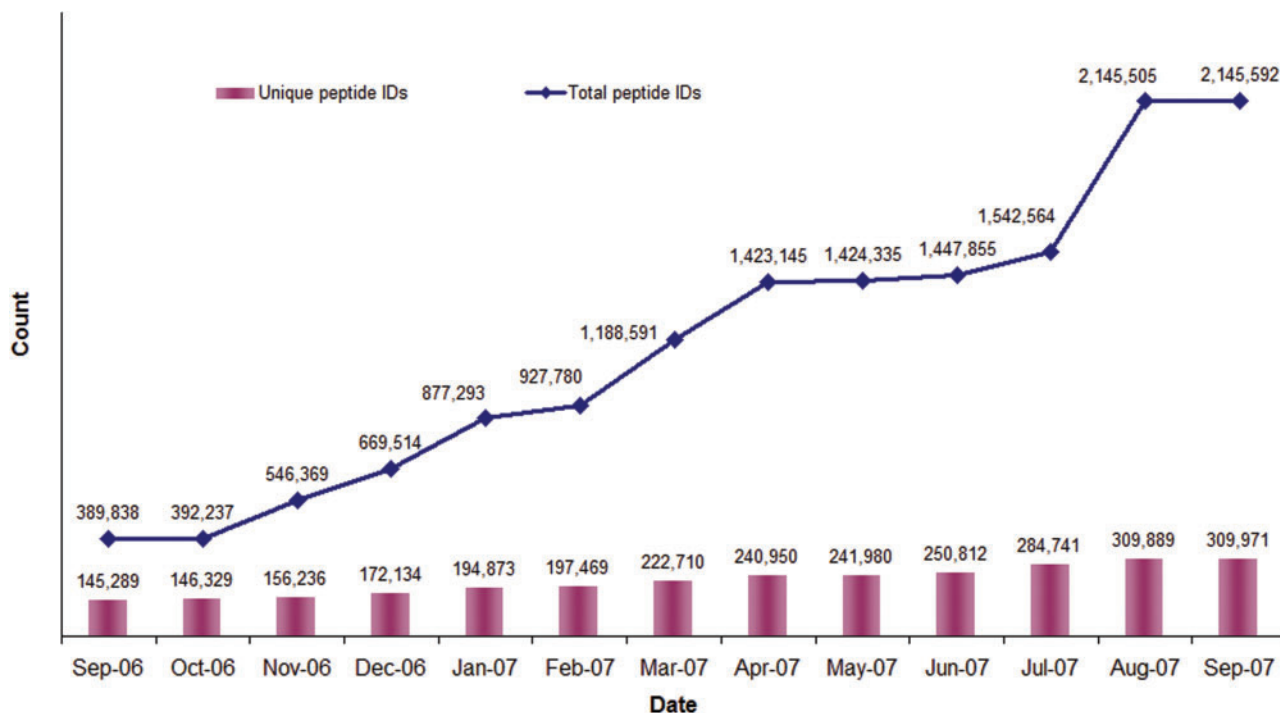


Figure 2. This graph illustrates the increasing redundancy of the peptide identifications submitted to PRIDE over the last year, as repeated identifications of the same peptides are performed. The total number of peptide identifications has increased 5.5-fold, however the number of unique peptide identifications in PRIDE has only doubled.

(<http://www.ebi.ac.uk/intact>) have been included from PRIDE (22). PRIDE also links to ChEBI, the EBI's database of chemical entities of biological interest (<http://www.ebi.ac.uk/chebi/>).

It is of note that as the quantity of data submitted to PRIDE grows, repeat identifications of the same unique peptide sequence are becoming increasingly frequent. Indeed, at the time of writing, each unique peptide identification in PRIDE is represented an average of 7 times, comprising repeat identifications both within individual experiments and across separate experiments (Figure 2).

Supporting data submission

The submission format for the PRIDE database is necessarily complex, reflecting the complexity of the domain and technologies used in proteomics. This complexity is compounded by the need to add value to datasets with thorough annotation. As described above and in Ref. (2), PRIDE makes use of several controlled vocabularies and ontologies to support this annotation in a uniform manner.

Unfortunately, this complexity makes creating a complete submission a difficult task, especially where access to programming support is limited. To mitigate this, several strategies have been employed. For laboratories with good programming support, a comprehensive Java Application Programming Interface (API) can be used to generate PRIDE XML and mzData XML.

For laboratories with more limited bioinformatics resources, two avenues are available. The PRIDE team

has expanded to include a data curator who provides direct support to submitting laboratories. This support may be limited to checking XML files that the laboratory has produced, or for a limited number of cases, it is possible for the PRIDE curator to provide direct programming support for the generation of PRIDE XML.

Finally, the PRIDE team has developed an interactive tool that runs in Microsoft Excel. The Proteome Harvest PRIDE Submission Spreadsheet (<http://www.ebi.ac.uk/pride/proteomeharvest>) is an Excel workbook that breaks down the complexity of a complete submission into several relatively simple spreadsheets. The 'Peptides' sheet is illustrated in Figure 3. The workbook makes use of embedded Visual Basic for Applications (VBA) to assist the user in generating a PRIDE XML file directly from the data that they have entered into the spreadsheet. To assist with the problematic step of annotating various parts of the data with appropriate ontology or controlled vocabulary terms, the workbook includes a form giving direct access to the Ontology Lookup Service (OLS) (23).

Collaborative development of PRIDE submission tools

A welcome development in the evolution of PRIDE has been the increasing involvement of collaborating groups outside the EBI. A team led by Simon Hubbard from the Faculty of Life Sciences, University of Manchester, has developed a mechanism to allow quantitative proteomic data to be encoded in PRIDE XML and mzData XML (24) by using cross-referenced controlled vocabulary terms to describe the samples and their relative quantities,

To add details of a peptide, you first need to enter detail of the protein that it provides evidence for on the "Proteins" worksheet and then click on the 'Add Peptide' button on column A of sheet "Proteins". Rows will then automatically be added to this sheet as required. If you wish to add details of protein modifications to an individual peptide on this sheet, click on the "Add PTM" button in column A. This will create a record on the "PTMs" sheet that you can populate further.

NOTE: If you wish PTMs to be automatically created from sequence annotations, you first need to define the meaning of the annotations on the worksheet "PTM Codes"

	↓ Peptide Unique ID	↓ Protein Unique ID	↓ Protein Accession	Sequence (Please do not include flanking sequences here)	↓ Start	↓ End	↓ Spectrum Reference
5	Add PTM	1	1 Q12341	EWTRWEFSDFVXCV	19	32	1001
6	Add PTM	2	1 Q12341	SWDFWEFRSDFVDSD	24	38	1000
7	Add PTM	3	1 Q12341	SSERSDFRSDCDV	29	41	999
8	Add PTM	4	1 Q12341	SDCVDFSERFSDFSD	34	49	998
9	Add PTM	5	1 Q12341	WERSDFSD	39	46	997
10	Add PTM	6	2 Q12342	SDFWEFREFSDSDSDFSFS	44	62	996
11	Add PTM	7	2 Q12342	EWRSDFSDFVSDFSDF	49	65	995
12	Add PTM	8	2 Q12342	VERSDFSVSDVSDVDFVDFVWEF	54	77	994
13	Add PTM	9	2 Q12342	SWERDSDFWERE	59	70	993
14	Add PTM	10	2 Q12342	SDFVWEFWEGVWEF	64	77	992
15	Add PTM	11	2 Q12342	WCFRTRGTRVCWVGWEWCF	69	87	991
16	Add PTM	12	3 Q12343	TRBGEVVRGEVFTGERE	74	90	990
17	Add PTM	13	3 Q12343	ERCFERGEVGEGERGE	79	96	989
18	Add PTM	14	4 Q12344	EWTRWEFSDFVXCV	84	97	988
19	Add PTM	15	4 Q12344	SWDFWEFRSDFVDSD	89	103	987
20	Add PTM	16	4 Q12344	SSERSDFRSDCDV	94	106	986
21	Add PTM	17	5 Q12345	SDCVDFSERFSDFSD	99	114	985
22	Add PTM	18	5 Q12345	WERSDFSD	104	111	984
23	Add PTM	19	5 Q12345	SDFWEFREFSDSDSDFSFS	109	127	983

Figure 3. A single sheet from the ProteomeHarvest PRIDE Submission Spreadsheet—Peptide Identification Data Entry.

illustrating the extensibility of the PRIDE XML format. This mechanism has been successfully demonstrated for iTRAQ labeling, but has the scope to encompass other quantitative techniques. The same team has developed 'PrideWizard, a tool that parses mass spectrometry data and Mascot search engine output, converting the collated data to PRIDE XML. This tool is available from <http://www.mcisb.org/software/PrideWizard>.

IMPROVING QUERY ACCESS TO PRIDE

Providing query and visualization of the large and complex datasets describing complete proteomics experiments is challenging and problematic. To attempt to meet this challenge, several new facilities have been added to the PRIDE user interface over the last two years and PRIDE has also benefited from the re-engineering of the EBI website, including the new EB-eye search engine that incorporates indexed data from almost every resource at the EBI, including PRIDE.

Resolving the database accession problem

The PRIDE database is populated by submissions of proteomic data from a wide variety of laboratories around the world, each of which selects a protein sequence or genomic sequence database against which searches are performed. The criteria for selecting a database varies with the species and the group concerned, with the consequence that the identifications in PRIDE do not fit under a single sequence accession system. This has proved problematic in that searching PRIDE with an accession from one sequence database will not return results annotated with a different database, even though they may identify the same protein.

The PRIDE team has developed the Protein Identifier Cross-Reference Service (PICR) (25) (<http://www.ebi.ac.uk/Tools/picr/>) which is able to map protein sequence identifiers from over 60 different databases via UniParc (the UniProt Archive) (26). These cross-references are now being included in the PRIDE database, which will enable

EBI > Databases > Proteomics > PRIDE > PRIDE BioMart

PRIDE

Dataset
PRIDE

Filters
Filter by Species : Homo sapiens (Human)
Filter by Tissue : JURKAT cell

Attributes
Experiment Title
Submitted Protein Accession
Peptide Sequence
Start Coordinate
End Coordinate

Export all results to Unique results only

Email notification to

View rows as Unique results only

Experiment Title	Submitted Protein Accession	Peptide Sequence	Start Coordinate	End Coordinate
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	15079369	QTEDVAEVLNR	214	224
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	Q15526	MAAVALQLGLR	1	12
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	Q9Y6A2	QPEIVAR	321	327
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	Q9Y6A2	QPEIVAR	321	327
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	21389381	ATGTPESQAR	2	11
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	17149828	QHPFYFKAAR	550	559
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	17149828	QHPFYFKAAR	550	559
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	O00273	QSCATVQR	163	170
Differential COFRADIC N-terminal proteomic analysis of Fas-induced apoptotic protein cleavage in Jurkat cells (2/4)	21359982	AEEQVNR	2	8

Figure 4. The PRIDE BioMart: results summary view for a simple query. The filter and display attributes can be seen on the left.

users to successfully query PRIDE with their favored accession system. The accession mapping task is performed for new data within 24 h and is refreshed for the entire PRIDE database every week. A useful side-effect is that the latest active accession is available for all submitted identifications, irrespective of the time that has passed since submission. The submitted accession is maintained in the PRIDE database. The PRIDE team is now working on including this mapping data in all PRIDE query and reporting mechanisms.

The PRIDE BioMart query interface

The PRIDE database now includes a BioMart interface (27) that offers several advantages. This query interface can be accessed from the menu item 'PRIDE BioMart' on the left of the PRIDE home page or directly at (<http://www.ebi.ac.uk/pride/prideMart.do>).

The PRIDE BioMart provides access to public PRIDE data from a query-optimized data warehouse that is synchronized with the main PRIDE database at regular intervals. The BioMart interface allows simple or complex queries to be built. The user has control over how the data is filtered, to restrict which records are included, and is able to select the attributes, equivalent to columns in a spreadsheet, that are included in the results. This avoids the need to search through a large table of results, much of which may be irrelevant, allowing the user to focus specifically on the information that is important to them.

The user can specify how the results are formatted; choosing from an HTML table displayed in a browser, a comma or tab-separated values file or a Microsoft Excel spreadsheet. The results file can be compressed to speed up data retrieval. The latest version of BioMart allows asynchronous data access in which the user specifies an

email address to which a link to large result sets can be sent.

The PRIDE BioMart provides programmatic web service access to public PRIDE data with the same flexibility as the web form.

The PRIDE BioMart is illustrated with the screen shot in Figure 4, showing a summary of the results for a customized query.

DISCUSSION

The focus of the PRIDE team at the EBI over the last two years has been on improving the ability of proteomics scientists to submit their data to PRIDE and to query PRIDE in more powerful and flexible ways. Progress has been made in both respects with the development of the Proteome Harvest PRIDE Submission Spreadsheet, the support provided by the PRIDE data curator, and the development of new user interface elements such as the PRIDE BioMart. Usage of the BioMart represents ~50% of the data volume downloaded from PRIDE. As BioMart queries return much more compact and customized result sets, this corresponds to the majority of queries to PRIDE now being made via the BioMart interface.

The Proteome Harvest PRIDE Submission Spreadsheet has been used extensively over the last year, contributing 17 PRIDE experiments including the Acid Mine Drainage environmental dataset described above.

It is recognized however that there is still work to be done. The PRIDE team continues to follow closely the development of the HUPO PSI data exchange formats (<http://psidev.info>). The goal of keeping PRIDE compatible with these standards will pay dividends in supporting data submission to PRIDE. The PRIDE query

interface still requires further development, including the incorporation of the Dasty2 DAS client (<http://www.ebi.ac.uk/dasty>) to allow graphical visualization of peptide identifications, together with improvements to the main PRIDE query interface and the PRIDE BioMart. An interesting consequence of the incorporation of PICR protein cross-references is that it will be possible to link the PRIDE BioMart directly to other BioMarts, allowing federated queries across these resources. We intend to set up links to the BioMart services offered by Ensembl (28) and Reactome (29) over the next few months, potentially with similar links in the reverse direction.

ACKNOWLEDGEMENTS

The PRIDE team would like to thank all data submitters for their contributions. The authors would also like to thank Dr Rolf Apweiler for his support and Dr Matthieu Visser for his valuable suggestions. S.Y.C. was supported by a grant from the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (A030003 to Y.-K. Paik). Biotechnology and Biological Sciences Research Council (BBS/B/17239, BB/E00573X/1). Funding to pay the Open Access publication charges for the article was provided by the European Union, "ProDaC" grant number LSHG-CT-2006-036814.

Conflict of interest statement. None declared.

REFERENCES

- Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) Pride: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
- Jones,P., Côté,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.
- Mead,J.A., Shadforth,I.P. and Bessant,C. (2007) Public proteomic ms repositories and pipelines: available tools and biological applications. *Proteomics*, **7**, 2769–2786.
- Editorial (2007) Democratizing proteomics data. *Nat. Biotechnol.*, **25**, 262.
- Craig,R., Cortens,J.C., Fenyo,D. and Beavis,R.C. (2006) Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.*, **5**, 1843–1849.
- Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. *et al.* (2006) The peptideatlas project. *Nucleic Acids Res.*, **34**, D655–D658.
- Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K.B. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
- Prince,J.T., Carlson,M.W., Wang,R., Lu,P. and Marcotte,E.M. (2004) The need for a public proteomics repository. *Nat. Biotechnol.*, **22**, 471–472.
- Hermjakob,H. and Apweiler,R. (2006) The proteomics identifications database (pride) and the proteomexchange consortium: making proteomics data accessible. *Expert Rev. Proteomics*, **3**, 1–3.
- Editor (2007) Mind the technology gap. *Nat. Methods*, **4**, 765–765.
- Phan,I.Q.H., Pilbout,S.F., Fleischmann,W. and Bairoch,A. (2003) Newt, a new taxonomy portal. *Nucleic Acids Res.*, **31**, 3822–3823.
- Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.
- Bard,J., Rhee,S.Y. and Ashburner,M. (2005) An ontology for cell types. *Genome Biol.*, **6**, R21.
- Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherrington,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
- Lo,I., Denef,V.J., Verberkmoes,N.C., Shah,M.B., Goltsman,D., DiBartolo,G., Tyson,G.W., Allen,E.E., Ram,R.J. *et al.* (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature*, **446**, 537–541.
- Hamacher,M., Apweiler,R., Arnold,G., Becker,A., Blüggel,M., Carrette,O., Colvis,C., Dunn,M.J., Fröhlich,T. *et al.* (2006) Hupo brain proteome project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. *Proteomics*, **6**, 4890–4898.
- He,F. (2005) Human liver proteome project: plan, progress, and perspectives. *Mol. Cell. Proteomics*, **4**, 1841–1848.
- Zheng,J., Gao,X., Beretta,L. and He,F. (2006) The human liver proteome project (hlpp) workshop during the 4th hupo world congress. *Proteomics*, **6**, 1716–1718.
- Pan,S., Zhu,D., Quinn,J.F., Peskind,E.R., Montine,T.J., Lin,B., Goodlett,D.R., Taylor,G., Eng,J. *et al.* (2007) A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. *Proteomics*, **7**, 469–473.
- Keller,A., Eng,J., Zhang,N., Li,X. and Aebersold,R. (2005) A uniform proteomics ms/ms analysis platform utilizing open xml file formats. *Mol. Syst. Biol.*, **1**, 2005.0017.
- Bantscheff,M., Eberhard,D., Abraham,Y., Bastuck,S., Boesche,M., Hobson,S., Mathieson,T., Perrin,J., Raida,M. *et al.* (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical abl kinase inhibitors. *Nat. Biotechnol.*, **25**, 1035–1044.
- Kerrien,S., Alam-Farouque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A. *et al.* (2007) Intact – open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
- Côté,R.G., Jones,P., Apweiler,R. and Hermjakob,H. (2006) The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.
- Siepen,J.A., Swainston,N., Jones,A.R., Hart,S.R., Hermjakob,H., Jones,P. and Hubbard,S.J. (2007) An informatic pipeline for the data capture and submission of quantitative proteomic data using itraq. *Proteome Sci.*, **5**, 4.
- Côté,R.G., Jones,P., Martens,L., Kerrien,S., Reisinger,F., Lin,Q., Leinonen,R., Apweiler,R. and Hermjakob,H. (2007) The protein identifier cross-reference (picr) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, **8**, 401.
- Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2004) Uniprot archive. *Bioinformatics*, **20**, 3236–3237.
- Durincq,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
- Hubbard,T.J.P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
- Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.