Nucleic Acids Research, 2006, Vol. 34, Database issue ***D659–D663***
doi:10.1093/nar/gkj138

PRIDE: a public repository of protein and peptide identifications for the proteomics community

Philip Jones[1,*], Richard G. Côté[1], Lennart Martens[2], Antony F. Quinn[1], Chris F. Taylor[1], William Derache[1], Henning Hermjakob[1] and Rolf Apweiler[1]

[1]EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [2]Department of Medical Protein Research, Flanders Interuniversity Institute for Biotechnology, Faculty of Medicine and Health Sciences, Ghent University, Rommelaere Institute, Building D, A. Baertsoenkaai 3, B-9000 Ghent, Belgium

Received August 15, 2005; Revised and Accepted October 27, 2005

ABSTRACT

PRIDE, the 'PRoteomics IDEntifications database' (http://www.ebi.ac.uk/pride) is a database of protein and peptide identifications that have been described in the scientific literature. These identifications will typically be from specific species, tissues and sub-cellular locations, perhaps under specific disease conditions. Any post-translational modifications that have been identified on individual peptides can be described. These identifications may be annotated with supporting mass spectra. At the time of writing, PRIDE includes the full set of identifications as submitted by individual laboratories participating in the HUPO Plasma Proteome Project and a profile of the human platelet proteome submitted by the University of Ghent in Belgium. By late 2005 PRIDE is expected to contain the identifications and spectra generated by the HUPO Brain Proteome Project. Proteomics laboratories are encouraged to submit their identifications and spectra to PRIDE to support their manuscript submissions to proteomics journals. Data can be submitted in PRIDE XML format if identifications are included or mzData format if the submitter is depositing mass spectra without identifications. PRIDE is a web application, so submission, searching and data retrieval can all be performed using an internet browser. PRIDE can be searched by experiment accession number, protein accession number, literature reference and sample parameters including species, tissue, sub-cellular location and disease state. Data can be retrieved as machine-readable PRIDE or mzData XML (the latter for mass spectra without identifications), or as human-readable HTML.

INTRODUCTION

The vast quantity of data associated with a single proteomics experiment can become problematic at the point of publishing the results. Laboratories tend to publish their work in an appropriate journal with perhaps a PDF document listing the proteins described. If space allows, the individual peptide sequences may be included but there is little possibility of including details of the mass spectra in this format. This clearly creates difficulties while attempting to reproduce the work of a laboratory to confirm their results.

Fortunately, the community has recognized and is tackling this problem through the formation of groups concerned with the development of standards for the capture and sharing of proteomics data. One such group is the HUPO Proteomics Standards Initiative (PSI) (1) who are in the process of developing standards tackling several aspects of proteomics, including ontologies of proteomics related terms, XML schemata and minimal reporting guidelines.

The Proteomics Identifications database (PRIDE), previously described by Martens *et al*. (2) is a PSI compliant public repository for proteomics identifications to which any proteomics laboratory is welcome to submit data. It is envisaged, but not mandated, that any such submission would normally be in the context of the corresponding submission of a manuscript to a journal describing the identifications submitted to PRIDE. As such, PRIDE aims to become the proteomics equivalent of the ArrayExpress database (3) used to capture microarray experiment data in support of journal publications.

PRIDE is not alone in this endeavor. Several other publicly available databases exist for the purpose of capturing and disseminating proteomics data from mass spectrometry. Such databases include the Global Proteome Machine Database (gpmDB) (4), The Institute for Systems Biology's PeptideAtlas (5) and the University of Texas' Open Proteomics Database (opd) (6). Currently in progress is the development of a collaborative agreement to exchange data between

*To whom correspondence should be addressed. Tel: +44 1223 492610; Fax: +44 1223 494468; Email: pjones@ebi.ac.uk

these and other emerging proteomics data repositories, including PRIDE.

DATABASE DESCRIPTION

What is the scope of PRIDE?

PRIDE can store

(i) The title and description of the experiment, together with contact details of the submitter.
(ii) Literature references.
(iii) Protein identifications by accession number supported by a corresponding list of one or more peptide identifications.
(iv) For each peptide identified, the sequence and coordinates of the peptide within the protein that it provides evidence for. Optionally, a reference to any submitted mass spectra that form the evidence for the peptide identification.
(v) Any post-translational modifications (natural or artefactual) coordinated in relation to the specific peptide that they have been found upon.
(vi) A description of the sample under analysis, including but not limited to the species of origin, tissue, sub-cellular location (if appropriate), disease state and any other relevant annotation.
(vii) A description of the instrumentation used to perform the analysis, including mass spectrometer source, analysers and detector, instrument settings and software settings used in data processing to generate peak lists.
(viii) Processed peak lists supporting the identifications in PRIDE in the versatile PSI mzData format.

PRIDE version 2.0, the release of PRIDE available at the time of writing, makes use of HUPO PSI deliverables such as the mzData XML schema (7) for capturing the settings and output from mass spectrometry work flow, including items vi–viii listed earlier. At present, the PRIDE XML schema encompasses the mzData schema with additional elements to allow protein and peptide identifications and post-translational modifications to be captured. It is envisaged that the analysisXML XML schema will be incorporated into PRIDE following its first release as a finalized schema, expected by early 2006, replacing large parts of the custom schema currently present in PRIDE.

Datasets currently available in PRIDE

A significant dataset that is publicly available from PRIDE at the time of writing is the set of protein and peptide identifications from the individual laboratories involved in the HUPO Plasma Proteome Project (8). This project was in part responsible for the requirements statement that initiated the PRIDE project.

Another publicly available dataset in PRIDE is a profile of the human platelet proteome (9) submitted by the Department of Medical Protein Research, Ghent University. This department is also scheduled to contribute a substantial dataset identifying proteolytic cleavage by caspases in apoptotic Jurkat T-cells (10) as well as a large set of spectra used to evaluate spectrum quality filtering software (11).

A dataset of protein and peptide identifications describing the organelle proteome of the secretory pathway is currently held as private data in PRIDE but is expected to be publicly available following publication of the related manuscript. At present this dataset can only be viewed by prior permission of the submitters.

It is expected that by the end of 2005 PRIDE will also contain the protein and peptide identifications and related mass spectra from the HUPO Brain Proteome Project (12) as a publicly available dataset.

Submission and retrieval of data

Data can be both submitted to and retrieved from PRIDE through a web interface, using either the PRIDE XML schema, which embeds mzData as a sub-element to allow inclusion of details of the spectra, or using the mzData XML schema, in which case all identifications will be omitted.

Data can also be viewed as a human-readable HTML table illustrated in Figure 1.

Figure 2 illustrates the search page. Queries can include experiment identifier, protein accession or identifier, literature references and sample parameters, including species, tissue, sub-cellular location and disease. The search results include all the experiments that match the query, together with options of how the data should be presented.

Data security in PRIDE: PRIDE as a tool for journal review

Data submitted to PRIDE is marked as public or private. Private data can be shared through a collaborative mechanism that allows individuals to apply to join a collaboration, their application then being confirmed or rejected by the creator of the collaboration. As well as allowing collaborating laboratories to share their data, this mechanism can also be used to allow manuscript reviewers to access the corresponding PRIDE entry in a confidential manner on a neutral site.

Use of controlled vocabularies and ontologies in PRIDE

By extending the mechanism designed for the mzData XML schema, PRIDE makes extensive use of external controlled vocabularies and ontologies (hereafter 'CVs') to annotate entries. The use of CVs ensures that queries for particular terms will capture all of the relevant data without omission due to differences in terminology. As a spin-off of the PRIDE development program, a SOAP web service to allow external CVs to be queried in an intelligent manner has been developed at the EBI, initially for use by PRIDE (http://www.ebi.ac.uk/ontology-lookup/). This service allows queries to take advantage of the hierarchical nature of ontologies. For example, if a user requests all protein identifications found in *pancreas*, the relevant Medical Subject Headings (MeSH) term will be looked up in the ontology web service and PRIDE will be queried for entries relating to the MeSH term 'Pancreas' as well as all child terms, currently in this case including 'Islets of Langerhans', 'Pancreas, Exocrine' and 'Pancreatic Ducts'. This mechanism assists the user by retrieving all the relevant data without the need to have a detailed knowledge of the terms involved.

CVs and ontologies suggested for use in PRIDE include MeSH (13) for animal anatomy and disease states; Gene

PRIDE Experiment Collection Version 2.0

Experiment 1: COFRADIC methionine proteome of unstimulated human blood platelets

Experiment: (top ↑)	**Description:** COFRADIC methionine proteome of unstimulated human blood platelets **Short Title:** Platelets MetOx **Accession:** 1
References:	Martens, L., Van Damme, P., Van Damme, J., Staes, A., Timmerman, E., Ghesqui?re, B., Thomas, G.R., Vandekerckhove, J., Gevaert, K., Proteomics, in press
Protocol:	**Name:** methionine oxidation induces a chromatographic shift on a diagonal RP-HPLC system.

Identifications

Accession	Splice Isoform	Database	Score	Threshold	Search Engine	Additional Information (Source / Name / Value)
IPI00295313		IPI human 2.31	52.0	35.0	Mascot 2.0.03	
IPI00017340		IPI human 2.31	56.0	35.0	Mascot 2.0.03	
IPI00026128		IPI human 2.31	80.6667	27.3333	Mascot 2.0.03	
IPI00031169		IPI human 2.31	48.0	33.5	Mascot 2.0.03	
IPI00291262		IPI human 2.31	60.6667	41.0	Mascot 2.0.03	
IPI00395553		IPI human 2.31	48.0	35.0	Mascot 2.0.03	
IPI00328748		IPI human 2.31	43.0	37.0	Mascot 2.0.03	
IPI00027497		IPI human 2.31	48.0	35.3333	Mascot 2.0.03	

mzData

Information:	**Version:**	1.05
	Accession:	1
Sample:	**Name:**	unstimulated human platelets
	Description:	
	Additional:	Source: NEWT, Name: Homo Sapiens, Value:
		Source: MeSH, Name: blood_platelets, Value:
Source File:		
Contact:	**Name:**	Kris Gevaert
	Institution:	Ghent University, Dept. of Medical Protein Research
	Contact information:	kris.gevaert@UGent.be
Instrument:	**Name:**	Micromass UK Limited, Cheshire, UK Q-TOF I
	Source:	Source: PSI, Name: IonizationType, Value: ESI
		Source: User, Name: comment, Value: Fragmentation time per ms/ms spectrum was 8 sec
	Analyzer #1:	Source: PSI, Name: AnalyzerType, Value: Q-TOF
	Detector:	Source: User, Name: No Detector Componen..., Value:
	Additional Information:	Source: PSI, Name: Vendor, Value: Micromass UK Limited, Cheshire, UK
		Source: PSI, Name: Model, Value: Q-TOF I

Figure 1. An example of PRIDE data in tabulated HTML format.

Ontology (GO) (14) for sub-cellular location; NEWT (15) for taxonomy, which is a superset of the NCBI taxonomy (16); the mass spectrometry ontology being developed by the HUPO PSI; RESID for naturally occurring post-translational modifications (17) and UNIMOD for protein modifications encountered in mass spectrometry experiments (18).

A PRIDE CV has been created for cases where existing CVs do not include a term required to annotate data in PRIDE.

The use of CVs and ontologies will allow the annotation of certain specific experimental results such as peptide retention times for LC-MS experiments or protein quantitation information for quantitive or differential proteomics experiments.

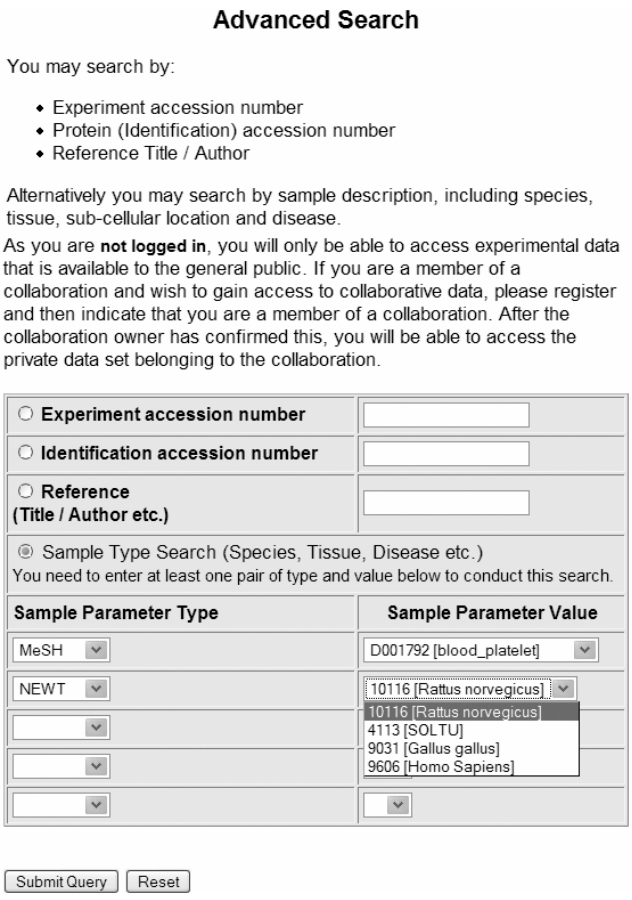

Figure 2. The PRIDE Advanced Search form.

Where the required CV terms do not exist already, PRIDE can accommodate these data elements through the use of user parameters.

PRIDE is an open-source software development project

Care has been taken throughout the development of PRIDE to ensure that all system components are open-source and freely available. PRIDE is written in Java and made available under the open-source Apache license. All the source code are freely available from the CVS repository (http://sourceforge.net/cvs/?group_id=122040). PRIDE uses the open-source Object-Relational Bridge (OJB) (http://db.apache.org/ojb/) API for database connectivity. As a consequence, PRIDE can easily be adapted to run on any SQL-based relational database management system. Configuration files exist for both Oracle (http://www.oracle.com) and MySQL (http://www.mysql.com/).

DISCUSSION

Here we consider possible applications of PRIDE from the perspective of the typical proteomics researcher. PRIDE offers the user several useful query opportunities including:

- Retrieving all proteomics experiments in which a particular protein of interest has been observed.
- Downloading proteome datasets of interest in a standard format for further local analysis.
- Retrieving the complete list of protein identifications (and the specific peptides found) for a given publication.
- Using the links provided via PRIDE to further explore the proteins identified.
- Comparison of one's own results with previous findings to quickly determine overlap as well as potentially novel findings.
- Planning of one's experiments: finding experimental protocols that have already been applied successfully to analyse your sample or even protein of interest.
- Re-analysing previously published results using your own techniques.
- Obtaining test sets for training or trying out novel algorithms (e.g. algorithms for protein or peptide identification).
- Retrieving typical base line proteomes for specific tissues and species.
- Allowing journal appointed reviewers to analyse the details of the identifications and potentially the supporting spectra as part of their review in a standardized manner.

FUTURE DEVELOPMENTS

The developers of PRIDE recognize that the system has room to evolve in several important aspects.

It is important for the future of PRIDE to keep apace with developments in the HUPO PSI. One important development of this initiative is the analysisXML XML schema, designed to hold details of protein and peptide identifications and

post-translational modifications, together with cross references to the relevant mzData entries describing spectra. It is intended that analysisXML will be fully supported by PRIDE for import and export, without modification or data loss, as soon as possible after the first stable release of the new analysisXML XML schema.

Submitters of identifications to PRIDE will naturally make use of their favored protein sequence database against which to search their spectra. Consequently, PRIDE will quickly fill with protein accessions and IDs from disparate protein sequence databases. An important short-term goal of the PRIDE project is to map all identifications to the UniProt database (19), including cross references to as many other protein databases as possible. This work will borrow heavily from the IntAct project (20), both in terms of code base and procedures for automatic and human curation.

A long-term goal of the PRIDE project is to provide an automated program of regular re-analysis of mass spectra deposited in PRIDE using the most up-to-date protein sequence databases and available open-source search algorithms such as X!Tandem (http://www.thegpm.org/TANDEM/) (21). The submitter's original identifications would continue to be available as described in the corresponding manuscript.

The EBI has developed a Distributed Annotation Server (DAS) (22) service for PRIDE (http://www.ebi.ac.uk/das-srv/pride/das/) using the BioJava Dazzle servlet (http://www.biojava.org/dazzle/). This service is publicly available and can be used to enable DAS clients such as Dasty (23), designed for visualizing protein sequence and annotation, to display identifying peptides for the protein specified in the DAS request.

ACKNOWLEDGEMENTS

PRIDE is supported through BBSRC iSPIDER and HUPO Plasma Proteome Project funding as well as a EU Marie Curie fellowship. L.M. is a research assistant of the Fund for Scientific Research, Flanders (Belgium) (F.W.O. Vlaanderen). L.M. would like to thank Prof Dr Kris Gevaert and Prof Dr Joël Vandekerckhove for their support. Funding to pay the Open Access publication charges for this article was provided by BBSRC iSPIDER and HUPO.

Conflict of interest statement. None declared.

REFERENCES

1. Orchard,S., Hermjakob,H. and Apweiler,R. (2003) The proteomics standards initiative. *Proteomics*, **3**, 1374–1376.
2. 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.
3. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E., Kapushesky,M. *et al*. (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **33**, D553–D555.
4. Craig,R., Cortens,J.P. and Beavis,R.C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res*., **3**, 1234–1242.
5. Deutsch,E.W., Eng,J.K., Zhang,H., King,N.L., Nesvizhskii,A.I., Lin,B., Lee,H., Yi,E.C., Ossola,R. and Aebersold,R. (2005) Human Plasma PeptideAtlas. *Proteomics*, **5**, 3497–3500.
6. 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.
7. Orchard,S., Hermjakob,H., Taylor,C.F., Potthast,F., Jones,P., Zhu,W., Julian,R.K.Jr and Apweiler,R. (2005) Second proteomics standards initiative spring workshop. *Expert Rev. Proteomics*, **2**, 287–289.
8. Omenn,G.S. (2004) The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics*, **4**, 1235–1240.
9. Martens,L., Van Damme,P., Van Damme,J., Staes,A., Timmerman,E., Ghesquiere,B., Thomas,G.R., Vandekerckhove,J. and Gevaert,K. (2005) The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics*, **5**, 3193–3204.
10. Van Damme,P., Martens,L., Van Damme,J., Hugelier,K., Staes,A., Vandekerckhove,J. and Gevaert,K. (2005) Caspase-specific and nonspecific *in vivo* protein processing during Fas-induced apoptosis. *Nat. Methods*, **2**, 771–777.
11. Flikka,K., Martens,L., Vandekerckhove,J., Gevaert,K. and Eidhammer,I. (2005) Improving the reliability and throughput of mass spectrometry based proteomics by spectrum quality filtering. *Proteomics*, in press.
12. Stephan,C., Reidegeld,K., Meyer,H.E. and Hamacher,M. (2005) HUPO Brain Proteome Project Pilot Studies: bioinformatics at work. *Proteomics*, **5**, 2716–2717.
13. Lipscomb,C.E. (2000) Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc*., **88**, 265–266.
14. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
15. Phan,I.Q., Pilbout,S.F., Fleischmann,W. and Bairoch,A. (2003) NEWT, a new taxonomy portal. *Nucleic Acids Res*., **31**, 3822–3823.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
17. Garavelli,J.S. (2004) The RESID Database of protein modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
18. Creasy,D.M. and Cottrell,J.S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536.
19. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
20. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
21. Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
22. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
23. Jones,P., Vinod,N., Down,T., Hackmann,A., Kahari,A., Kretschmann,E., Quinn,A., Wieser,D., Hermjakob,H. and Apweiler,R. (2005) Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, **21**, 3198–3199.