Nucleic Acids Research, 2006, Vol. 34, Database issue ***D411–D414***
doi:10.1093/nar/gkj141

Human protein reference database—2006 update

Gopa R. Mishra[1], M. Suresh[1], K. Kumaran[1], N. Kannabiran[1], Shubha Suresh[1], P. Bala[1], K. Shivakumar[1], N. Anuradha[1], Raghunath Reddy[1], T. Madhan Raghavan[1], Shalini Menon[1], G. Hanumanthu[1], Malvika Gupta[1], Sapna Upendran[1], Shweta Gupta[1], M. Mahesh[1], Bincy Jacob[1], Pinky Mathew[1], Pritam Chatterjee[1], K. S. Arun[1], Salil Sharma[1], K. N. Chandrika [1], Nandan Deshpande[1], Kshitish Palvankar[1], R. Raghavnath[1], R. Krishnakanth[1], Hiren Karathia[1], B. Rekha[1], Rashmi Nayak[1], G. Vishnupriya[1], H. G. Mohan Kumar[1], M. Nagini[1], G. S. Sameer Kumar[1], Rojan Jose[1], P. Deepthi[1], S. Sujatha Mohan[1], T. K. B. Gandhi[1], H. C. Harsha[1], Krishna S. Deshpande[1], Malabika Sarker[1], T. S. Keshava Prasad[1] and Akhilesh Pandey[2,3,4,*]

[1]Institute of Bioinformatics, International Tech Park, Bangalore 560 066, India, [2]McKusick-Nathans Institute of Genetic Medicine, [3]Department of Biological Chemistry and [4]Department of Oncology, Johns Hopkins University, Baltimore, MD 21205, USA

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

ABSTRACT

Human Protein Reference Database (HPRD) (http://www.hprd.org) was developed to serve as a comprehensive collection of protein features, post-translational modifications (PTMs) and protein–protein interactions. Since the original report, this database has increased to >20 000 proteins entries and has become the largest database for literature-derived protein–protein interactions (>30 000) and PTMs (>8000) for human proteins. We have also introduced several new features in HPRD including: (i) protein isoforms, (ii) enhanced search options, (iii) linking of pathway annotations and (iv) integration of a novel browser, GenProt Viewer (http://www.genprot.org), developed by us that allows integration of genomic and proteomic information. With the continued support and active participation by the biomedical community, we expect HPRD to become a unique source of curated information for the human proteome and spur biomedical discoveries based on integration of genomic, transcriptomic and proteomic data.

INTRODUCTION

The Human Protein Reference Database (HPRD) is a protein information resource that provides extensive information pertaining to human proteins including domain architecture, protein functions, protein–protein interactions, post-translational modifications (PTMs), enzyme–substrate relationships, subcellular localization, tissue expression and disease association of genes (1–3). In order to make HPRD a more comprehensive resource, we have greatly expanded the number of protein entries, protein–protein interactions and PTMs. We have also incorporated additional query (e.g. BLAST) and browse options and provided explanatory pages for motifs found in the proteins cataloged in HPRD. Some of the new features include protein isoforms, links to signal transduction pathways and integration of GenProt Viewer, a novel browser that we have recently developed. HPRD currently contains over 20 000 protein entries including 1587 protein isoforms and has grown significantly in size over the last 3 years (Figure 1a).

Cataloging protein–protein interactions

A crucial aspect of any proteomic analysis is the elucidation of interacting proteins—the interactome (4). HPRD currently has 33 710 unique protein–protein interactions. The experimental evidence for the interactions is derived from *in vivo*

*To whom correspondence should be addressed at McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, 733 N. Broadway, BRB Room 569, Baltimore, MD 21205, USA. Tel: +1 410 502 6662; Fax: +1 410 502 7544; Email: pandey@jhmi.edu

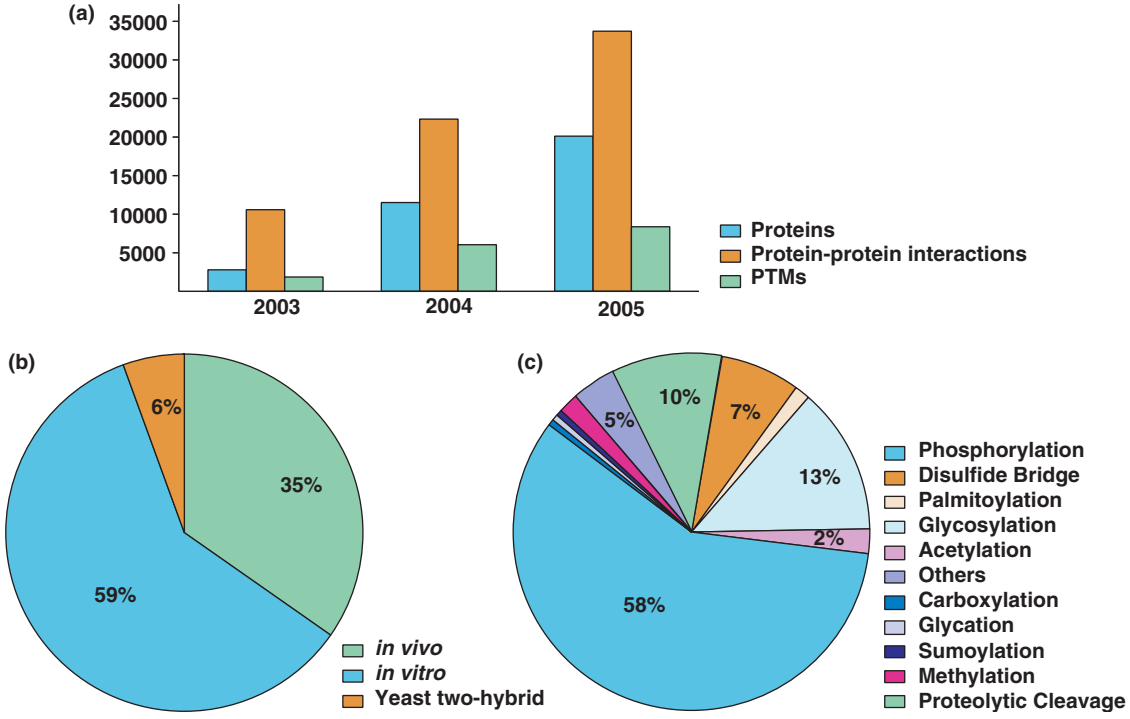

Figure 1. Statistics pertaining to HPRD growth, experimental types for protein–protein interactions and a breakdown of PTMs. (**a**) Growth of HPRD over the last 3 years with respect to protein entries, protein–protein interactions and PTMs. (**b**) Distribution of protein–protein interactions in HPRD based on the type of the experimental method. (**c**) Distribution of various types of PTMs in HPRD. The percentage of the respective PTM is indicated only when it is greater than or equal to 2.

experiments for 19 175 interactions, *in vitro* for 11 114 interactions and yeast two-hybrid for 1813 interactions. Figure 1b shows the distribution of the protein–protein interactions annotated in HPRD. Table 1 shows the overall statistics as of September 15, 2005.

Enrichment of PTMs, subcellular localization and tissue expression data in HPRD

PTMs can alter both structure and function of proteins. In recent years, several large-scale studies have been carried out to characterize PTMs using proteomic methods. For instance, 2002 phosphorylation sites were identified using mass spectrometry from HeLa cell nuclear extract in a single experiment (5). A total of 5011 phosphorylation events and 1132 glycosylation events are among the 8409 recorded PTMs in HPRD (Figure 1c). Updated annotations involving subcellular localization include 489 nucleolar proteins (6,7) and 270 secreted proteins (8). Similarly, tissue expression data have been added to a number of entries including those encoded by KIAA cDNAs (9).

Novel features added since initial release of HPRD

Protein isoforms. One of the important additions to HPRD is the inclusion of protein isoforms. Criteria for inclusion as an isoform include only those RefSeq database (10) entries with different CDS (coding sequence) for the same gene. Thus, only those alternate splice forms are considered in which the splicing involves the coding region and not the 5′ or 3′ untranslated regions. All annotations are displayed for all isoforms by default except when isoform-specific data regarding subcellular localization, PTMs, domain architecture or tissue expression are available. Mainly due to lack of data, isoform-specific annotations for protein–protein interactions, substrates and disease involvement are not provided currently but are common to all isoforms.

Table 1. The total numbers of entries in the various fields in HPRD are shown

Feature	Numbers
Total protein entries	20 097
Isoforms included in the protein entries	1587
Total number of protein interactions	33 710
Total number of PTMs	8409
Total number of domains and motifs	478
Total number of enzyme–substrate relationships	3343

Enhanced search options. HPRD can be queried through gene symbols or a variety of database accession numbers such as RefSeq (10), OMIM (11), Swiss-Prot (12), HPRD and Entrez Gene (13). A multiple search option is included in the updated query system that allows the database to be queried by simultaneously specifying several different parameters. Because accession numbers, gene symbols or protein names might still not yield the protein being searched for, we have now also included a BLAST option as a search tool.

Links to pathways. In order to visualize and identify the potential function of a protein in the context of a large signaling network and its interaction partners, we have curated a number of pathways. These pathways have been integrated through the

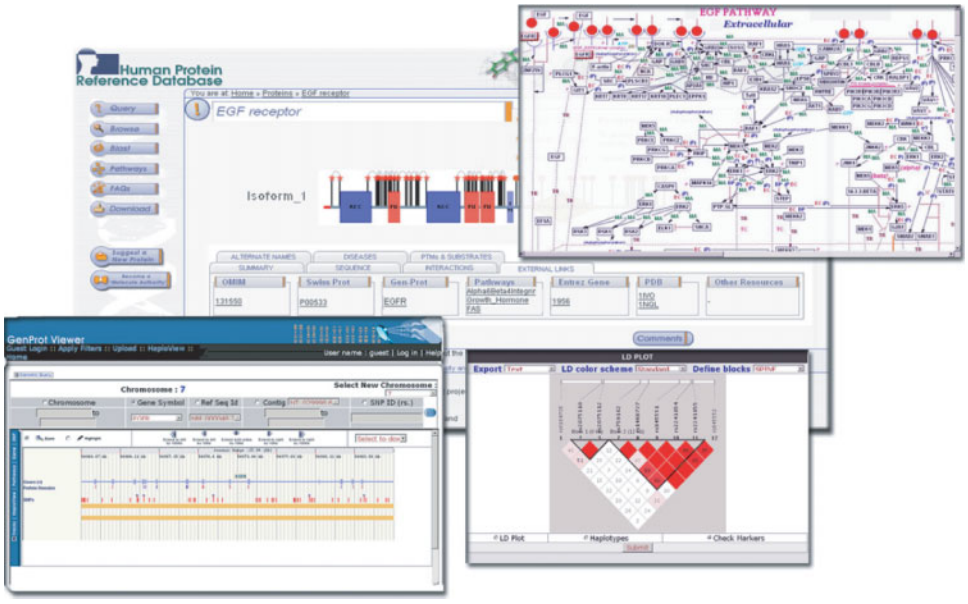

Figure 2. A screenshot of the molecule page of EGF receptor in HPRD is shown. The molecule page shows a graphical representation of the protein with its protein domains as polygons and sites of phosphorylation as vertical straight lines with red circles at the ends and disulfide bonds represented as gray lines. A link to the GenProt Viewer showing a number of SNPs and exons is shown in the left lower corner. Haploview that shows population haplotype patterns is shown on the lower right corner. The popup on the top right shows the EGF signaling pathway.

'Pathways' tab. The pathway data are diagrammatically represented using GenMAPP (Gene Microarray Pathway Profiler) (14), a computer application designed for viewing and analyzing the data in the context of biological pathways (Figure 2). These pathways are a collection of literature-derived information usually downstream of ligand-receptor interactions. In addition to information about protein–protein interactions, pathways include reactions involving PTMs, shuttling of proteins between subcellular compartments, activation or inhibition of enzymatic activity and up or down regulation of mRNAs.

Integration of GenProt Viewer. HPRD provides annotations by mapping the protein onto the genome alongwith transcript and SNP information for each protein from the molecule page, through GenProt Viewer (Figure 2). The GenProt Viewer, a browser (http://www.genprot.org) developed by our group, provides an integrated genomic, transcriptomic and proteomic view of the human genome. Genomic annotations that have been addressed here include the mapping of single nucleotide polymorphisms and homology blocks of the human genome when compared against that of mouse. Transcriptomic data include RefSeq annotations and the categorization of protein-coding transcripts into untranslated regions and open reading frame. It also integrates 'Haploview' for investigating population haplotype patterns (15). Experimentally derived peptide sequences obtained by mass spectrometry that have been deposited in PeptideAtlas (http://www.peptideatlas.org) (16) and PRIDE (17) repositories are also mapped onto the genomic sequence in GenProt Viewer. The peptides are linked to the sequence pages in these two repositories. Finally, the BLAST option allows users to query the genome using protein or nucleotide sequences.

Downloading HPRD data

HPRD data are available for download in XML as well as tab delimited file formats. Regular updates of full release of all the data in a compressed format is available using the 'Download' tab (http://www.hprd.org/FAQ?selectedtab=DOWNLOAD+REQUESTS). Interaction datasets in PSI-MI (18) format are provided as individual files for each protein as well as a single combined file for the entire dataset. The PSI-MI is an evolving data format which was originally released as level 1.0. We are currently formatting the protein–protein interaction data in HPRD to stay compliant with the latest version of this specification (PSI-MI level 2.5).

Future plans

We wish to develop a Protein Distributed Annotation System, which will enable laboratories throughout the world to annotate valuable proteomic information including PTMs, tissue expression, protein–protein interactions and enzyme–substrate relationships in the context of HPRD data. We hope to link any data obtained by mass spectrometry directly to such annotations in HPRD. We are also in the process of integrating transcriptomic data into HPRD, which will allow gene expression patterns to be visualized in normal and diseased states. Based on user input over the last 3 years, we also hope to include a list of genes regulated by the major transcription factors.

CONCLUSIONS

Our strategy of involving the biomedical community in providing feedback for individual entries using the 'Comments' button and designating interested researchers as 'Molecule

Authority' listed under the 'Credits' tab is already successful. To make the best use of HPRD and to understand the annotation procedure and philosophy, we strongly encourage all users to visit the 'FAQs' page (http://www.hprd.org/FAQ). We hope that this community involvement will continue to intensify over the coming years in our effort to make HPRD a knowledgebase of human proteins that will assist in biomedical discoveries by serving as a complete resource of genomic, transcriptomic and proteomic information and in providing an integrated view of sequence, function and protein networks in health and disease.

ACKNOWLEDGEMENTS

The HPRD was developed with funding from the National Institutes of Health and the Institute of Bioinformatics. Funding to pay the Open Access publication charges for this article was provided by a grant from the NIH (RR020839).

Conflict of interest statement. A.P. serves as Chief Scientific Advisor to the Institute of Bioinformatics. A.P. is entitled to a share of licensing fees paid to the Johns Hopkins University by commercial entities for use of the database. The terms of these arrangements are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

REFERENCES

1. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*., **32**, D497–D501.
2. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.
3. Navarro,J.D., Talreja,N., Peri,S., Vrushabendra,B.M., Rashmi,B.P., Padma,N., Surendranath,V., Jonnalagadda,C.K., Kousthub,P.S., Deshpande,N. *et al*. (2004) BioBuilder as a database development and functional annotation platform for proteins. *BMC Bioinformatics*, **5**, 43.
4. Walhout,A.J. and Vidal,M. (2001) Protein interaction maps for model organisms. *Nature Rev. Mol. Cell Biol*., **2**, 55–62.
5. Beausoleil,S.A., Jedrychowski,M., Schwartz,D., Elias,J.E., Villen,J., Li,J., Cohn,M.A., Cantley,L.C. and Gygi,S.P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. USA*, **101**, 12130–12135.
6. Andersen,J.S., Lyon,C.E., Fox,A.H., Leung,A.K., Lam,Y.W., Steen,H., Mann,M. and Lamond,A.I. (2002) Directed proteomic analysis of the human nucleolus. *Curr. Biol*., **12**, 1–11.
7. 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.
8. Zhang,Z. and Henzel,W.J. (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci*., **13**, 2819–2824.
9. Kikuno,R., Nagase,T., Nakayama,M., Koga,H., Okazaki,N., Nakajima,D. and Ohara,O. (2004) HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE. *Nucleic Acids Res*., **32**, D502–D504.
10. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
11. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.
12. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
13. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
14. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet*., **31**, 19–20.
15. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
16. 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*, **13**, 3497–3500.
17. 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*, **13**, 3537–3547.
18. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **22**, 177–183.