Nucleic Acids Research, 2002, Vol. 30, No. 1 137–141

Annotating the human proteome: the Human Proteome Survey Database (HumanPSD™) and an in-depth target database for G protein-coupled receptors (GPCR-PD™) from Incyte Genomics

Peter E. Hodges*, Pauline M. Carrico, Jennifer D. Hogan, Kathy E. O'Neill, J. J. Owen, Mary Mangan, Brian P. Davis, Joan E. Brooks and James I. Garrels

Incyte Genomics, Proteome Division, 100 Cummings Center, Suite 435M, Beverly, MA 01915, USA

Received October 17, 2001; Revised and Accepted November 26, 2001

ABSTRACT

The Proteome Division of Incyte Genomics has released new volumes to the BioKnowledge® Library to add human, mouse and rat protein information to its rich collection of model organism Proteome Databases. The Human Proteome Survey Database (HumanPSD™) compiles the fundamental properties of more than 25 000 characterized mammalian proteins. HumanPSD™ includes clear, concise and current protein descriptions (Title Lines), the protein sequence, calculated physical properties, precomputed BLAST alignments, controlled-vocabulary protein properties and Gene Ontology™ terms, and a list of published references. Each report also contains expression data, Pfam domain information and an associated Mouse Mutant Phenotype section describing behavioral, physiological and cellular phenotypes for over 1500 mouse mutant phenotypes. GPCR-PD™ contains more than 3200 Protein Reports from the three mammalian species for G protein-coupled receptors, their protein ligands, associated G-proteins and their downstream signaling proteins. In addition to the features described above, each GPCR-PD Protein Report displays annotations of experimental findings from over 10 000 publications. These databases provide important new volumes of Proteome's BioKnowledge Library (http://www.incyte.com), integrating protein information from model organisms with the human proteome.

INTRODUCTION

On June 26, 2000, the completion of the first draft of the human genome was announced, marking the inauguration of the post-genomic era (1,2). Already, scientific focus has shifted from the monumental task of sequencing the genome to the even more daunting task of predicting the genes from the genome sequence, predicting the transcribed and spliced messenger RNAs, identifying the proteins encoded by these transcripts, predicting their functions and experimentally testing these predictions. Understanding the human proteome will require the integration of three levels of experimental analysis: (i) increasingly sophisticated predictions from the genomic sequence, (ii) functional genomic experiments that assay the biological characteristics of thousands of genes, proteins or mutant strains in a single parallel analysis, and (iii) the contributions of thousands of smaller, but higher resolution, experimental projects published in hundreds of scientific journals each year.

The Proteome Division of Incyte Genomics is uniquely positioned to curate the human proteome. Since 1995, Proteome has produced databases capturing the cumulative knowledge of genetic, biochemical and functional properties for the complete proteomes of experimentally important model organisms and microbial pathogens, which now include YPD™ for *Saccharomyces cerevisiae*, MycoPathPD™ for 17 different fungal pathogens, including *Candida albicans*, PombePD™ for *Schizosaccharomyces pombe* and WormPD™ for *Caenorhabditis elegans* (3). Each database represents a volume in Proteome's BioKnowledge® Library, an interconnected resource of protein information for academic, biotechnology and pharmaceutical scientists. Experimental results are extracted by highly trained scientific curators from the published scientific literature, bringing the experimental results from a large collection of scientific papers together on a single page. This information is presented in a user-friendly, searchable format consistent across all species. As evidenced by the Proteome model organism databases, the BioKnowledge Library serves as an excellent platform for the display of large-scale experimental data sets, allowing functional genomic information to be seen within the context provided by all of the other experimental findings about those proteins. The growth of the BioKnowledge Library is supported by the bioinformatic expertise to collect and manage huge quantities of bibliographic and sequence data, and augmented by BioKnowledge Transfer, a process for prediction of protein functions and properties based on comparative genomics.

The BioKnowledge Library has expanded to include human, mouse and rat protein information in the form of the Human Proteome Survey Database (HumanPSD™) and the G Protein-Coupled Receptor Proteome Database (GPCR-PD™). The

*To whom correspondence should be addressed. Tel: +1 978 922 1643; Fax: +1 978 922 3971; Email: info@incyte.com

Table 1. HumanPSD™ statistics (as of September 16, 2001)

	Total	Human	Mouse	Rat
Protein Reports	24 573	13 900	6988	3685
Protein Properties	88 642	44 422	28 865	15 535
Proteins with at least one assigned property	24 359	13 805	6881	3673
Proteins characterized by experimental evidence	13 532	6448	4406	2478

addition of these databases has supported the coming surge of proteomic and functional genomic research driven by the human genome sequence. HumanPSD and GPCR-PD, in addition to the other volumes of the database, have undergone massive growth in both content and functionality due to increased resources and expertise provided by the acquisition of Proteome by Incyte Genomics in December 2000. HumanPSD provides a curated and up-to-date description for each protein (the Proteome 'Title Line') and selected protein properties. In addition, the Proteome Division of Incyte Genomics has released the GPCR-PD™, our first targeted in-depth Proteome Database for mammals. GPCR-PD focuses on the most medically relevant and best-characterized GPCR families and their associated downstream signaling molecules.

ORGANIZATION OF THE BIOKNOWLEDGE LIBRARY

HumanPSD and GPCR-PD follow the easily navigatable format developed in the other volumes of the BioKnowledge Library, YPD, MycoPathPD, PombePD and WormPD. (See Supplementary Material for example Protein Reports from HumanPSD and GPCR-PD.) Information is presented in web-page format, deployed via the subscriber's Intranet or accessed from Incyte Genomics servers via the Internet (http://www.incyte.com). Subscriptions are available to both academic and corporate organizations via the Internet. Corporate subscribers may access the relational form of the database, which can be integrated with a wide range of existing in-house resources, and can be mined using standard database query tools. Trained Incyte curators (all PhD scientists with research experience) collect information about each protein from multiple published references (often hundreds of papers). In addition to published data, we combine properties predicted by analysis of the protein sequence or predicted by sequence similarity to related proteins from other species. The collected information is displayed in a 'one protein, one page' format, the Protein Report. Access to this information is optimized by the presentation format, providing:

1. Clear identification of where you are (which species and which protein).
2. Clear paths of navigation.
3. Stratified presentation of important information first.
4. Tabular presentation of controlled-vocabulary protein classification.
5. Tabular summaries of BLAST similarities for all model organisms.
6. Auxiliary information, details and references accessible in pop-up windows.
7. Extensive links between Protein Reports, within a species and between species.

Entry points into the database are provided via search forms and allow users to perform searches by content including gene or protein names or synonyms, or full text searches. Users can also search the databases using controlled-vocabulary properties to search by protein properties, Gene Ontology™ (GO) categories (4), tissues, cell or tumor expression, or by sequence similarity to a user-provided sequence, or navigate via hypertext links from other Protein Report (based on sequence similarity, physical interaction or any functional relationship curated from the literature). This rich connectivity is a hallmark of Proteome Databases and allows the user to draw on experimental results from a range of model organisms. Incyte strives to provide complete and accurate attribution for each piece of information. Data are designated as being predicted or experimentally validated, and either directly associated with a reference or clearly designated as a prediction made by Incyte scientists. Users can search many properties and GO categories on the basis of this experimental validation, selecting or excluding predictions. Because Incyte curators are reviewing the current scientific literature as it is published, the content of the BioKnowledge Library is continually reviewed and updated. An updated version of each database is released every week.

FEATURES OF HumanPSD

HumanPSD was founded on an initial collection of 14 000 non-redundant sequences of characterized human, mouse and rat proteins, aided greatly by NCBI's curated sequence database LocusLink (5,6). The HumanPSD database is expanding with additions to LocusLink and from Incyte's curation of newly identified proteins from the recent biological literature, and now contains over 25 000 protein reports (Table 1). Protein Reports contain a Title Line, selected protein properties, GO categories, expression data, protein sequence and calculated physical properties, BLAST analysis for related proteins, domain information, and external links to other databases. Mouse protein reports also contain Mouse Mutant Phenotype information describing behavioral, physiological and cellular phenotypes for over 1500 mouse mutant phenotypes. The Title Lines, protein properties, GO categories, expression data and Mouse Mutant Phenotype sections have been curated from a set of more than 300 000 references covering 320 scientific journals.

The Title Line provides a succinct and up-to-date description of the protein and its function, with additional important attributes that may include subcellular localization, family membership, domain structure, similarity to characterized proteins in other species and disease involvement. All references link directly to their PubMed abstracts, and additional links are provided, where available, to NCBI's GenBank, LocusLink,

UniGene, the Protein Information Resource (PIR-International) (7) and the Online Mendelian Inheritance in Man (OMIM).

Each protein report also contains Incyte's controlled-vocabulary protein properties, which include the following categories: biochemical function, cellular role, organismal role, subcellular localization and molecular environment. Each property entered contains an evidence code, distinguishing those properties supported by experimental evidence from those predicted by similarity or by analogy to other proteins. GO classifications for every HumanPSD protein have also been assigned to proteins within three major categories: biological processes, molecular functions or cellular components. Similar to Incyte's classification schema, GO allows classification of proteins from all organisms to the same categories, promoting cross-species comparisons. However, GO's hierarchical classification allows finer definition, progressive grouping of proteins into broader or narrower categories by moving up or down the hierarchy, and capture of partial classification information for imprecisely characterized proteins by assigning a higher hierarchical category. Both Incyte properties and GOs co-exist on the protein reports and both types of classification are presented.

In order to allow the fullest prediction of meaningful biological functions for uncharacterized proteins, while reducing the chance of 'annotation catastrophe' (the propagation of inaccurate or unsubstantiated predictions), Incyte has developed a proprietary technology termed BioKnowledge Transfer (BKT). This annotation pipeline allows automatic functional predictions to be evaluated by scientific curators in a high-throughput, decision-supporting platform. Uncharacterized proteins are analyzed by sequence alignment (using BLAST with Smith–Waterman refinement) and by family or domain identification [using Pfam models (8) detected with hidden Markov models (HMMs)]. Scientific editors have created thresholds for transfer of specific protein functional predictions from these alignments, and have created an extensive database of experimentally determined properties associated with protein domains. The algorithmic predictions are presented to a scientific curator, with the tools to view the supporting evidence or further analyze the foundation for the predictions, and allowing the curator to give the final confirmation, modification or rejection of every prediction. The key features ensuring the success of BKT are:

1. The extensive resource of literature-supported protein properties collected in the BioKnowledge Library.
2. The transfer of only those properties with direct, documented experimental evidence.
3. The review of every prediction by a scientific curator with specific training in the analysis procedure.

Any predictions made through BKT are specifically indicated in the evidence code presented in the database, and can be included or excluded from database searches.

HumanPSD reports organ, tissue type, cell type and tumor location for expression of proteins and genes in a tabular format using controlled vocabulary. It also notes when proteins or transcripts are determined to be absent, when such information is reported in the literature. The experimental technique used to determine expression is included and all data is referenced to the publication source.

The Mouse Mutant Phenotype section describes gross behavioral, physiological and cellular phenotypes for over 1500 mouse mutant phenotypes (knockouts) represented in HumanPSD. The information is divided into four main categories: Viability, Anatomical Phenotype, Physiological Phenotype and Disease Model. Also included in HumanPSD is domain information derived from profile HMMs using the Pfam database (8). This content allows searching on one or more protein domain and family in conjunction with searches on other categories. On each protein report, the domains are listed with E-value and amino acid location in the protein sequence. HumanPSD links to a number of external databases including SWISS-PROT (9) and the Protein Data Bank (PDB) (10) for 3D protein structure information on over 13 000 Protein Reports. Protein Reports show pairwise BLAST alignment of HumanPSD proteins to entries in PDB.

In October 2001 a new field containing information about Alternative Forms on more than 3500 Protein Reports in HumanPSD was added to the Protein Reports. This category indicates the identification of experimental literature indicating alternative isoforms of the protein. Each observation is linked to PubMed so the original article can easily be found.

In addition to the extensive information available in the web format described above, two more features are provided for the in-house installed subscriptions of HumanPSD and GPCR-PD. First, subscribers can now navigate between the HumanPSD and GPCR-PD Protein Reports and Incyte's LifeSeq® Gold, the world's most comprehensive database of human gene information. Secondly, to facilitate the annotation of microarrays and other large-scale data sets, the relational table form of HumanPSD contains a GenBank Accession Translation Table. This table maps nearly 2 million GenBank sequences to HumanPSD protein identifiers based on DNA sequence similarity, greatly simplifying the task of associating annotation from HumanPSD to DNA microarray results.

Incyte strives to include the best and most complete sequence information within the BioKnowledge Library. Thus, as sources of protein sequence predictions grow, the Proteome Databases will continue to expand their repository of sequences. HumanPSD will continue to expand by adding modules of new content and new types of information that complement the existing content, which may include expanded information on protein and gene expression, genetic variation and protein networks (including modifications, participation in pathways, interaction with other proteins and membership in protein complexes). An important role for the BioKnowledge Library is integration with other data sources, some included in the BioKnowledge Library environment and some represented as links to data outside the BioKnowledge Library. Another important role for the BioKnowledge Library is to provide a standard for human protein research, and to that end, Incyte Genomics provided Title Lines and GO annotations for an initial set of 10 000 human proteins to the LocusLink database of NCBI in December 2000. These data are available to view and query in the Public HumanPSD at http://www.incyte.com.

In comparison to other available protein annotation resources, the most striking advantages of using Incyte's Proteome Databases as a knowledge resource supporting biomedical research are:

1. Its foundation in published experimental results.
2. The broad range of data brought together in one searchable environment.
3. The quality of the data brought by human review.

4. Ease of access to data, both in well-designed web page layouts and intuitively designed relation data tables.

The most distinct value of Incyte's Proteome Databases derives from the combination of algorithmic knowledge extraction and human evaluation. HumanPSD contains over 88 000 protein properties and 270 000 summaries of experimental results. Experimental results derive from a scientist's reading and understanding the scientific literature, and protein properties are extracted from the literature as the conclusions drawn by a trained scientist, in evaluation of evidence presented in high-throughput decision-supporting tools. Public databases are forced to rely on either low-throughput manual curation, unverified author submissions or unevaluated high-throughput algorithmic output. Manually curated databases (for example, Mendelian Inheritence in Man, or the manual curation at SWISS-PROT, LocusLink, the Mouse Genome Database or the Rat Genome Database) are typically slow in growth, cover a narrower range of published literature and often do not capture data in a computer-readable format or tie each observation to direct primary literature references. Automatically generated databases of (predicted and verified) proteins and their properties (for example, TrEMBL) cannot apply human evaluation to high-throughput predictions, and can make it difficult for the user to distinguish predictions from experimentally validated properties.

This dichotomy between automatic function prediction and manual curation, and Incyte's ability to bridge the two, is perhaps best exemplified by the assignment of GO annotations to human proteins (as summarized at http://www.geneontology.org/; update November 8, 2001]). Compugen Inc. has performed a fully automatic assignment of GO terms based on extraction of protein descriptions from database submissions, transference of properties based on protein homologies, and *de novo* subcellular location prediction from protein sequence information. They have applied 331 672 GO terms to 114 606 predicted human proteins, but do not provide any references to published experimental documentation. The European Bioinformatics Institute's GO annotation project has assigned GO properties based on algorithmic transfer of protein properties from database submissions, predictions based on protein family membership or domain matches, aggregation of GO assignments from public domain sources (Incyte's Public HumanPSD) and manual curation. They assign 66 997 GO terms to 16 492 human proteins, of which only 25 279 GO terms (for 8589 human proteins) are supported by literature references (9898 citations). The overwhelming majority of these documented assignments derive from Incyte's Public HumanPSD (which was released into the public domain in December 2000 and has not been updated), providing 26 767 GO terms associated with 8156 human proteins, including 23 125 GO terms associated with direct literature reference to 8766 references. The current subscription version of HumanPSD covers mouse and rat proteins as well as human, and has been expanded and updated by Quality Control and new curation. Today, HumanPSD contains the largest collection of experimentally documented protein properties for human, mouse and rat proteins, including 61 082 GO terms describing 10 903 human proteins, 39 350 GO terms describing 6938 mouse proteins and 21 207 GO terms describing 3892 rat proteins.

Table 2. GPCR-PD™ statistics (as of September 16, 2001)

	Total	Human	Mouse	Rat
Protein Reports	3348	1407	1042	899
GPCR Proteins	702	324	235	142
Associated proteins	2646	1109	820	764

FEATURES OF GPCR-PD™

GPCR-PD is the product of curation from selected literature on human, mouse and rat GPCRs and the proteins that affect, or are affected by, GPCR signaling. Curated information is presented both as controlled-vocabulary properties and free-text annotation of experimental results. Specifically, GPCR-PD covers 702 human, mouse and rat GPCRs identified from LocusLink, UniGene or reported in the recent literature (Table 2). Also included are 2910 Protein Reports for the protein ligands for the receptors, G proteins and other signaling partners, downstream targets, and upstream regulators. Because the primary focus of the literature selection for GPCR-PD is the receptors and their G proteins, these proteins are covered in greatest depth throughout the database. GPCR-PD is currently available by subscription only.

GPCR-PD utilizes the same web-based format as HumanPSD and contains all of the features described in HumanPSD. It is fully interlinked with HumanPSD and all of Incyte's model organism databases. The elements of the GPCR-PD Protein Report are very similar to those of the model organism databases. Curated information is presented as a Title Line, as controlled-vocabulary properties and GO categories, expression data, and as free-text annotations organized by topic. The properties and annotation topics are the same as those used in the model organism databases with a few additions: annotations are sorted into new topic headings of 'Regulates' (for the intracellular processes regulated by that protein), 'Ligand' (natural ligands, agonists and antagonists), and 'Variations' (genetic variation, polymorphisms, SNPs and disease-causing mutations). Each annotation also captures corresponding experimental detail (tissue or cell line used, experimental methodology, etc.) that the user can choose to display or hide.

The core of Incyte's curation process involves scientists reading published papers in their entirety, and GPCR-PD currently contains data extracted from over 10 000 references. GPCR-PD relies on Incyte's processes of bibliographic analysis and reference management. It was founded on a set of references spanning the last decade of research from a selected set of high-impact journals. Incyte is continuing to expand our journal coverage and developing more sophisticated and detailed queries to identify articles relevant to protein properties. The selected references focus on signaling by GPCRs, the roles they play in normal or disease physiology, their expression patterns and the regulation of both their expression and activity. Recently published articles (<1 month old) are the primary focus of our update of GPCR-PD, with an additional effort to expand coverage of the older literature.

ACCESS TO THE BIOKNOWLEDGE LIBRARY

Subscriptions to the BioKnowledge Library may be arranged by contacting Proteome Division of Incyte Genomics (info@incyte.com). Institute-wide academic subscriptions to the mammalian databases are now available as a single web-based product that includes HumanPSD and GPCR-PD. Information regarding access, pricing or a free Internet trial may be arranged by contacting Proteome Division of Incyte Genomics (info@incyte.com). Academic use of the model organism databases (YPD, MycoPathPD, WormPD and PombePD) is available at no charge on the Incyte–Proteome web site (http://www.proteome.com, and after January 2002 at http://www.incyte.com). For additional data tables and collaborative use of data from the BioKnowledge Library, contact info@incyte.com.

CONTACTING HumanPSD AND GPCR-PD

We appreciate feedback from our users concerning new data submission, additions, clarifications and corrections. Personal communications will be cited as such. Any correspondence should be directed to info@incyte.com.

CITING HumanPSD AND GPCR-PD

Authors wishing to make use of the information provided by HumanPSD or GPCR-PD should cite this article as a general reference for access to and content of these databases, and contact Incyte (info@incyte.com) regarding guidelines for publication of data or use in a web presentation.

SUPPLEMENTARY MATERIAL

The following material is available as Supplementary Material at NAR Online: a sample HumanPSD Protein Report; a sample GPCR-PD Protein Report; Functional Characterization of the Known Human Proteome.

ACKNOWLEDGEMENTS

We would like to express our appreciation of the hard work and dedication of our scientific curators, the foundation of our databases. We would also like to thank the diligence and expertise of the editors of these databases. We appreciate the helpful comments from academic advisors, collaborators and the feedback from our academic users and corporate subscribers.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Costanzo,M.C., Hogan,J.D., Cusick,M.E., Davis,B.P., Fancher,A.M., Hodges,P.E., Kondu,P., Lengieza,C., Lew-Smith,J.E., Lingner,C. *et al.* (2000) The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.*, **28**, 73–76.
4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
5. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
6. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.
7. Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.W., Orcutt,B.C., Pfeiffer,F., Tsugita,A. *et al.* (2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32. Updated article in this issue: *Nucleic Acids Res.*, (2002), **30**, 35–37.
8. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
9. Bairoch,A. and Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
10. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.