

The UCSC Proteome Browser

Fan Hsu*, Tom H. Pringle², Robert M. Kuhn, Donna Karolchik, Mark Diekhans,
David Haussler¹ and W. James Kent

Center for Biomolecular Science and Engineering, School of Engineering and ¹Howard Hughes Medical Institute, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA and ²Sperling Biomedical Foundation, Eugene, OR 97405, USA

Received August 13, 2004; Revised and Accepted October 14, 2004

ABSTRACT

The University of California Santa Cruz (UCSC) Proteome Browser provides a wealth of protein information presented in graphical images and with links to other protein-related Internet sites. The Proteome Browser is tightly integrated with the UCSC Genome Browser. For the first time, Genome Browser users have both the genome and proteome worlds at their fingertips simultaneously. The Proteome Browser displays tracks of protein and genomic sequences, exon structure, polarity, hydrophobicity, locations of cysteine and glycosylation potential, Superfamily domains and amino acids that deviate from normal abundance. Histograms show genome-wide distribution of protein properties, including isoelectric point, molecular weight, number of exons, InterPro domains and cysteine locations, together with specific property values of the selected protein. The Proteome Browser also provides links to gene annotations in the Genome Browser, the Known Genes details page and the Gene Sorter; domain information from Superfamily, InterPro and Pfam; three-dimensional structures at the Protein Data Bank and ModBase; and pathway data at KEGG, BioCarta/CGAP and BioCyc. As of August 2004, the Proteome Browser is available for human, mouse and rat proteomes. The browser may be accessed from any Known Genes details page of the Genome Browser at <http://genome.ucsc.edu>. A user's guide is also available on this website.

INTRODUCTION

The UCSC Genome Browser (1,2), which was developed in conjunction with the assembly and publication of the first Human Draft Genome (3), has become a popular website

for genomic researchers around the world, serving more than 5000 users each day and fulfilling data download requests of more than 50 GB per day. To support an ever-increasing number of sequenced genomes and their associated annotations, the Genome Browser has been extended with additional tools, such as the Table Browser (4) and Gene Sorter. Recently, we added another tool, the Proteome Browser, to complement the Genome Browser in covering the proteome world. By pairing these two browsers, researchers can gain quick, integrated access to genome and proteome data simultaneously.

The Proteome Browser provides a wealth of protein information presented in the form of graphical images of tracks (Figure 1) and histograms (Figure 2) and links to other Internet sites (Figure 3). The browser currently offers data for two human assemblies (41 486 proteins for the July 2003 release and 41 890 proteins for the May 2004 release), two mouse assemblies (36 591 proteins for the October 2003 release and 36 741 proteins for the May 2004 release) and one rat assembly (7415 proteins for the June 2003 release).

Version 1.0 of the Proteome Browser is tightly coupled with the Genome Browser. To access the Proteome Browser, a user should first open the Genome Browser at <http://genome.ucsc.edu>, enter a protein ID or a gene ID to display the Genome Browser annotation tracks page and then click a gene on the Known Genes track to display a page with detailed information about the gene. The 'Proteome Browser' link in the Quick Links section on this page will start the Proteome Browser.

DATABASE ORGANIZATION

The architecture of the Proteome Browser database is similar to that of the Genome Browser database, with the addition of a central protein data store that supports multiple genomes. The central protein store is implemented in two MySQL databases: swissProt and proteins. The swissProt database consists of 29 tables that store the parsed data from Swiss-Prot (5), TrEMBL and TrEMBL-NEW. The proteins database contains 15 tables, most of which store cross-references among

*To whom correspondence should be addressed. Tel: +1 831 459 5692; Fax: +1 831 459 1809; Email: fanhsu@soe.ucsc.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

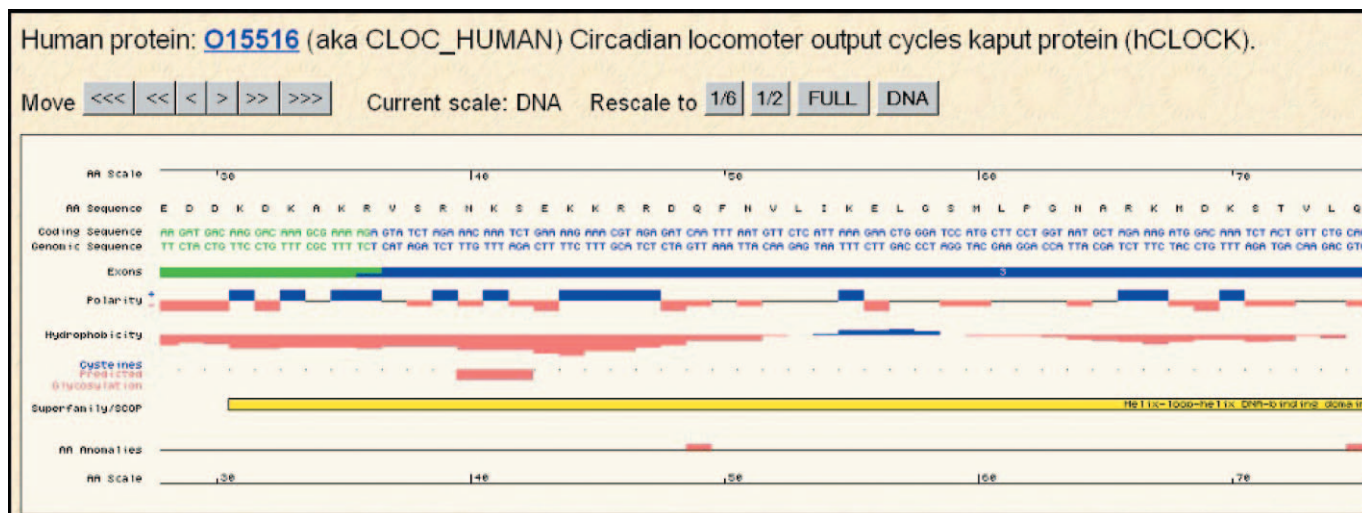


Figure 1. This Proteome Browser tracks display sample shows the tracks of human protein CLOC_HUMAN in DNA mode that display both the protein and genomic sequences. As the corresponding mRNA is a negative strand, the DNA sequence and its complementary coding sequence are both shown. On the exon track, a part of the 2nd and 3rd exons is shown. The start of a Superfamily domain, Helix-loop-helix DNA-binding domain, is shown by the yellow bar. In this particular protein, the amino acid glutamine (Q) is flagged to indicate that it is present in a higher-than-average percentage.

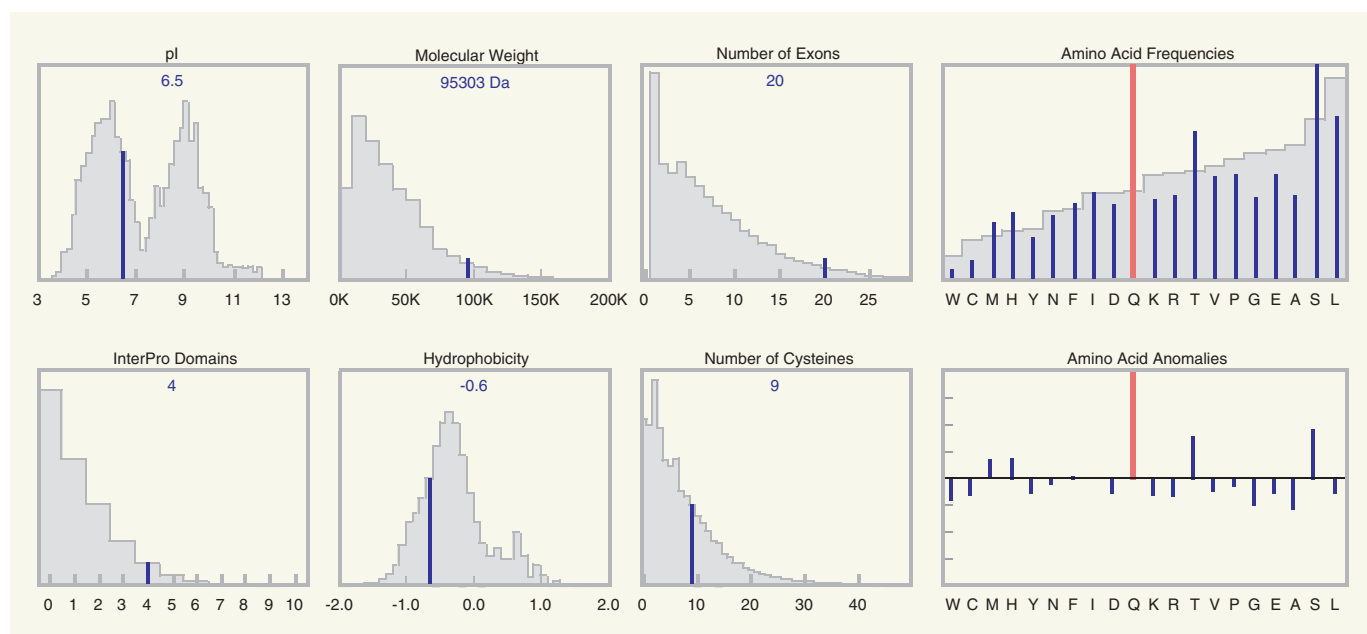


Figure 2. The Proteome Browser property histograms sample shows that this protein has high molecular weight and a large number of exons. It has four InterPro domains and an unusually high concentration of glutamine.

Swiss-Prot, UCSC Known Genes, PDB, Ensembl, HUGO, Superfamily, Pfam, InterPro, etc. The two protein databases are updated once every two months. In addition, there are 33 tables in each genome assembly database that enable the Proteome Browser. These tables store assembly-specific data, e.g. protein property distribution data calculated from the protein dataset of the specific genome.

The Proteome Browser is a CGI-based web application written in C that uses the MySQL database. The entire source code of the Genome Browser and Proteome Browser is available at <http://genome.ucsc.edu/admin/jksrc.zip>. The source

code for the Proteome Browser's CGI executable, pbTracks, can be found in the kent/src/hg/protein/pbTracks subdirectory.

DOWNLOADING DATA

All the central protein databases and genomic databases used by the Proteome Browser and the Genome Browser can be downloaded through specific links at <http://hgdownload.cse.ucsc.edu/downloads.html>.

UCSC links:

- Genome Browser - [AB002332](#)
- Gene Details Page - [AB002332](#)
- Gene Sorter - [AB002332](#)


InterPro Domains: [Graphical view of domain structure](#)

- [IPR001092](#) - HLH_basic
- [IPR001067](#) - Nuc_translocat
- [IPR001610](#) - PAC
- [IPR000014](#) - PAS

Pfam Domains:

- [PF00010](#) - Helix-loop-helix DNA-binding domain
- [PF00785](#) - PAC motif
- [PF00989](#) - PAS domain

ModBase Predicted Comparative 3D Structure on [O15516](#)



Front Top Side

The pictures above may be empty if there is no ModBase structure for the protein. The ModBase structure frequently covers just a fragment of the protein. You may be asked to log onto ModBase the first time you click on the pictures. It is simplest after logging in to just click on the picture again to get to the specific info on that model.

Pathways:

- BioCarta - [circadianPathway](#) - Circadian Rhythms
- KEGG - [hsa04710](#) - Circadian rhythm - Homo sapiens

Figure 3. This example shows the web links to UCSC Genome Browser, Known Gene details page and Gene Sorter; InterPro and Pfam domains; predicted 3D structure by ModBase; and pathways from BioCarta and KEGG.

PROTEIN BROWSER TRACKS

The protein tracks image displays a set of aligned tracks showing information about the selected protein's amino acid sequence and distribution anomalies, the corresponding DNA sequence and exon composition, and protein traits such as hydrophobicity, glycosylation potential, polarity, cysteine content and Superfamily domain composition. Figure 1 shows an example of the Proteome Browser tracks display for a human protein, CLOC_HUMAN.

The navigation buttons permit scrolling left or right and resizing of the tracks image. The image retains its scrolled settings when it is rescaled. In the default FULL image mode, the browser displays the amino acid sequence of the protein. If the DNA display mode is selected, the browser also shows the genomic sequence as a series of 3-nucleotide codons.

The Proteome Browser offers an option to automatically generate either Postscript or PDF format files of the images to support the publication and presentation needs of users.

Genomic sequence and complement. This track shows the corresponding genomic sequence of the selected protein. If the corresponding mRNA matches the forward (+) strand of

the genomic sequence as displayed in the Genome Browser, the DNA sequence is displayed as a single row of nucleotides with each codon directly under the corresponding amino acid. When the mRNA matches the reverse (-) strand of the genomic sequence, the track shows both the positive and negative strands of the DNA to allow users to more easily compare the amino acid and the nucleotide sequence.

The protein-to-genome alignments are derived from the Known Genes dataset of the UCSC Genome Browser. This dataset was produced by selecting GenBank mRNA sequences associated with proteins in the Swiss-Prot, TrEMBL and TrEMBL-NEW databases. These mRNAs were aligned to the base genome using BLAT (6). A valid mRNA alignment was required to have 40% of its sequence aligned with at least 97% sequence identity. From the multiple mRNAs associated with a protein, the mRNA with the highest score (based on sequence length, closeness of its translation to the protein and release date) was designated as the representative mRNA of a Known Gene. A small number of RefSeq entries that did not have supporting mRNA evidence were also added to the Known Genes dataset.

Each amino acid of a protein in the Known Gene dataset was then mapped onto three genomic base positions using the representative mRNA. First, the protein was aligned to the mRNA using TBLASTN (7). Partially aligned or unaligned amino acids were ignored. The protein-to-mRNA alignment was then mapped onto the genome using BLAT. This approach compensated for cases where the representative mRNA did not exactly code for the protein. It also supported the collection of data for spliced codons. The alignment results were stored in the kgProtMap table.

Exons. This track shows the correspondence of the gene's exons to the protein's amino acid sequence. Based on the exon structure data from the kgProtMap table, exons are depicted by alternating blue and green bands. When the codon of an amino acid is split across two exons, the exon boundaries are stair-stepped (to show intron phase). Clicking an exon on this track will display a UCSC Genome Browser page focusing on the selected exon.

A thin black line immediately above this track indicates the viewing display range of the Genome Browser. The user can zone in to a specific exon of a Known Gene in the Genome Browser, click the exon, and then select the Proteome Browser to see that particular exon in the protein.

Polarity. This track shows the polarity of each amino acid in the protein sequence. Amino acids tending toward negative charge are represented by red blocks below the centerline; amino acids tending to positive charge are represented in blue above the line.

Hydrophobicity. This track shows the distribution of hydrophobic residues across the selected protein, according to the Kyte–Doolittle (8) scale, using a sliding six-amino-acid window.

Cysteines. This track shows the locations of cysteines as blue blocks along the peptide chain.

Glycosylation. This track shows predicted glycosylation sites within the protein, represented by red blocks displayed below the centerline. Potential *N*-glycosylation sites are predicted by a tripeptide motif, NxT or NxS, where x is not proline.

Superfamily/SCOP. This track shows the positions and names of predicted Superfamily (9) domains within the protein as determined by hidden Markov model comparison to all known three-dimensional (3D) structures.

Amino acid anomalies. This track marks amino acids in the protein sequence that differ significantly in abundance from the average genome-wide occurrence frequency.

PROTEIN PROPERTY HISTOGRAMS

The second section of the Proteome Browser web page displays histograms depicting genome-wide distribution of protein properties such as isoelectric point, molecular weight, number of exons, InterPro domains and cysteines, plotted together with the specific value for the selected protein. Figure 2 shows an example of a Proteome Browser histograms display.

Isoelectric point (pI). The pI (isoelectric point) histogram shows the pI of the selected protein relative to a

genome-wide statistical distribution of pI of all proteins from Swiss-Prot and TrEMBL. The distribution data were generated by the same algorithm used by the Swiss-Prot pI calculation tool (10).

Molecular weight. This histogram shows the selected protein's molecular weight relative to a genome-wide statistical distribution of molecular weights of all proteins from Swiss-Prot and TrEMBL.

Number of exons. This histogram displays a genome-wide statistical distribution based on the number of coding exons in the kgProtMap table.

Amino acid frequencies. This histogram shows the abundance of amino acids within the selected protein relative to their average occurrence genome-wide. Any amino acid with an abundance that falls outside the 2.5–97.5% genome-wide distribution range is represented by a red line. The amino acid data were obtained from proteins in the Known Gene table that were found in Swiss-Prot (TrEMBL and TrEMBL-NEW excluded). Standard deviations were derived from compositions of the genome-wide set of proteins.

InterPro domains. This histogram shows the number of predicted InterPro domains contained in the selected protein relative to the number of domains found in proteins genome-wide.

Hydrophobicity. This histogram shows the mean hydrophobicity of the selected protein relative to a genome-wide distribution of all proteins in the Swiss-Prot database.

Number of cysteines. This histogram compares the selected protein's cysteine count to a genome-wide statistical distribution of cysteine abundance in proteins found in the Known Genes dataset for the specific genome assembly.

Amino acid anomalies. This histogram displays the extent to which certain amino acids within the protein sequence differ in abundance from their average occurrence genome-wide. Amino acids present in excessive or deficient amounts are represented by a red vertical bar extending above or below the middle line. The height of the bar indicates the degree of departure of the amino acid frequency from the expected value genome-wide.

The amino acid data in these tables were obtained from Swiss-Prot and TrEMBL. Amino acids are flagged as anomalous if their occurrence in a protein falls outside the 2.5–97.5% genome-wide distribution range. Proteins with <100 amino acids have been discarded from the sample to exclude variances attributable to short length.

LINKS TO OTHER PROTEIN RESOURCES

The final section of the Proteome Browser web page provides links to a variety of related protein resources on the Internet. Figure 3 shows an example of links to various protein resources available on the Web.

UCSC links. Through these links, the user can access related gene annotation information available in the UCSC Genome Browser tool set, including the Genome Browser annotation tracks display, the associated Known Genes details page and data on related genes in the Gene Sorter.

Domain information links. The browser provides three primary sources of detailed domain information from Superfamily/SCOP, InterPro and Pfam.

3D structure links. The browser displays small stamp-sized 3D-structure images from PDB (11) and ModBase (12) when corresponding structures are available.

Pathways links. The browser provides links to three major pathway databases, KEGG (Kyoto Encyclopedia of Genes and Genomes), BioCarta/CGAP (human and mouse) and BioCyc (human only).

FASTA format sequence. For users' convenience, the protein sequence is displayed in the FASTA format. This text may be copied to other bioinformatics tools using cut and paste.

FUTURE DIRECTIONS

The UCSC Proteome Browser will be extended to cover additional genomes in the future. We plan to add support for stand-alone queries on all Swiss-Prot and TrEMBL proteins including those not associated with the genomes currently covered by the Proteome Browser. This new functionality will be available at <http://genome.ucsc.edu/cgi-bin/pbGateway> before the end of 2004. We would also like to add protein-protein interaction and other types of protein-related data and web links in the future.

CONTACTING US

The mailing list genome@cse.ucsc.edu provides a forum for announcements of new releases and features questions, and discussion about the UCSC Proteome Browser, Genome Browser, Table Browser, Gene Sorter and databases. Users may subscribe to this list at <http://www.cse.ucsc.edu/mailman/listinfo/genome>. To report problems while accessing the website, servers or mirror sites, or for correspondence inappropriate for the public forum, send email to genome-www@cse.ucsc.edu.

ACKNOWLEDGEMENTS

We would like to thank Swiss-Prot for sharing their high-quality protein data and the isoelectric point calculation

algorithm. We would also like to thank the many collaborators who have contributed sequence and annotation data to our project, as well as our users for their feedback and support. The UCSC Proteome Browser project is funded by National Human Genome Research Institute (NHGRI) and the Howard Hughes Medical Institute (HHMI).

REFERENCES

1. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
2. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
3. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32** (Database issue), D493–D496.
5. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Res.*, **31**, 365–370.
6. Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
7. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Kyte, J. and Doolittle, R. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
9. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
10. Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.-C., Williams, K.L., Appel, R.D. and Hochstrasser, D.F. (1998) Protein identification and analysis tools in the Expasy server. In Link, A.J. (ed.), *2-D Proteome Analysis Protocols*. Humana Press, NJ, pp. 531–552.
11. 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.
12. Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., Webb, B., Greenblatt, D., Huang, C.C., Ferrin, T.E. and Sali, A. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32** (Database issue), D217–D222.