



HINT: a database of annotated protein-protein interactions and their homologs

Ashwini Patil¹ and Haruki Nakamura¹

¹*Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan*

Received 26 November, 2004; accepted 19 January, 2005

Despite the abundance of protein-protein interaction databases currently available online, a source that identifies and lists similar interactions in different species is lacking. The Homologous Interactions (HINT) database is such a collection of protein-protein interactions and their homologs in one or more species. The interactions and their homologs are annotated with Eukaryotic Cluster of Orthologous Groups (KOG) IDs, InterPro domains, Gene Ontology (GO) terminology and Protein Data Bank (PDB) structures. HINT is available as an interactive Web server at <http://helix.protein.osaka-u.ac.jp/hint/>.

Key words: homologous interactions, interaction families

Protein-protein interactions in various organisms are increasingly becoming the focus of study in the identification of cellular functions of proteins. For any given interaction, it is of significant interest to find similar interactions in different species. Such a comparative study helps in the transfer of annotations between interactions from better annotated species to poorly annotated ones. It also aids the identification of likely true interactions from error-prone high-throughput datasets since, intuitively, an interaction found in more than one species is likely to be universal. It has been recently estimated that the total number of interaction types is limited to about 10,000¹. Grouping similar interactions on the basis of sequence homology would help in their classification in different distinct interaction types or families.

Corresponding author: Haruki Nakamura, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan.
e-mail: harukin@protein.osaka-u.ac.jp

There are a number of protein-protein interaction databases available online that give information about experimentally determined interactions. Some of these are the Database of Interacting Proteins (DIP)², IntAct³ and Biomolecular Interaction Network Database (BIND)⁴. Although these databases provide considerable information about the interaction of interest, they do not provide any information about interactions similar or homologous to it.

With these goals in mind, we present here HINT, a database of homologous interactions with various annotations for the interacting proteins. HINT is available online at <http://helix.protein.osaka-u.ac.jp/hint/>.

Methods

Two interactions are considered homologous if the interacting proteins of one interaction are homologous to the interacting proteins for the other interaction (Fig. 1). Homologous interactions include, but are not limited to, orthologous interactions (similar interactions found in different species) and paralogous interactions (similar interactions in the same species).

We use protein-protein interaction data for different model organisms from DIP (July 2004 version) and IntAct (September 2004 version). For each interaction, the sequence homologs of the interacting proteins are determined using PSIBLAST⁵ with 5 iterations and an E-value cutoff of 10⁻⁸. If an interaction is found that involves any of the homologs of the interacting proteins, then it is deemed homologous to the interaction under consideration. Figure 1 illustrates this concept. We thus generate groups of homologous interactions that have been determined by small-scale or high-throughput experiments. We determine if any two interactions are orthologous or paralogous by assigning the

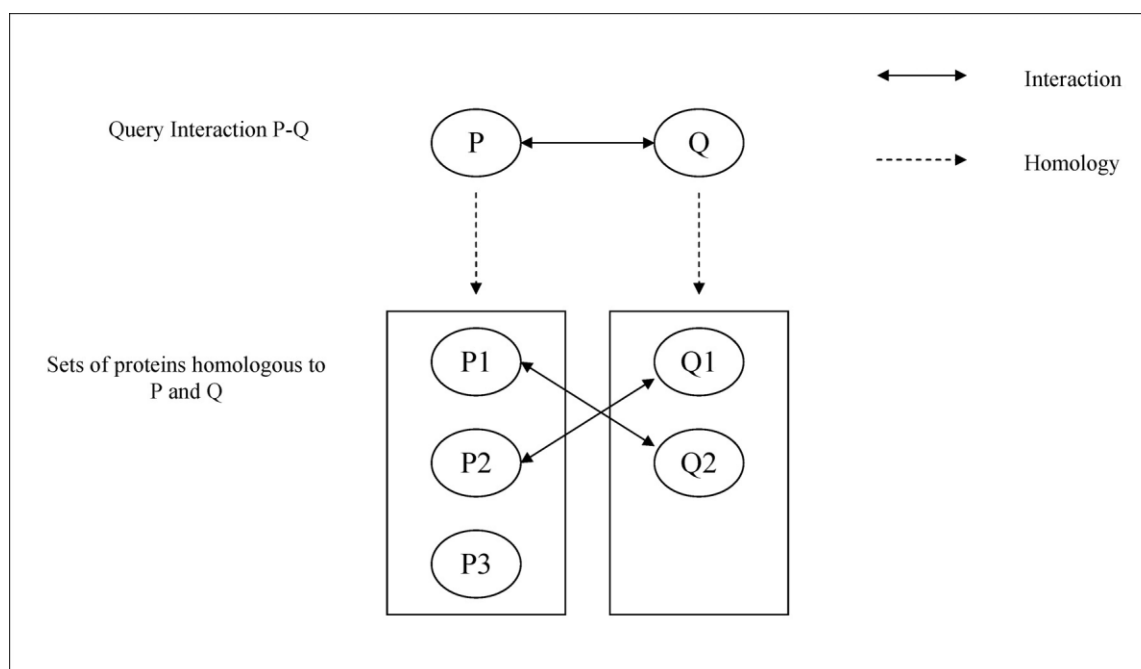


Figure 1 Homologous interactions — Proteins P1, P2, P3 are the sequence homologs of protein P. Similarly, proteins Q1, Q2 are the sequence homologs of protein Q. Interactions P1–Q2 and P2–Q1 are homologous to the interaction P–Q.

interacting proteins to clusters from the Eukaryotic Cluster of Orthologous Groups (KOG) database⁶. The interacting proteins for each interaction are also annotated with domain definitions from InterPro⁷, Gene Ontology (GO) terms⁸ and Protein Data Bank (PDB) structures, where available.

The interactions from DIP and IntAct were parsed from XML files in Proteomics Standards Initiative — Molecular Interaction (PSI-MI) XML format⁹. This allows for easy extension of the database by incorporating protein-protein interaction data from various other databases using this format. HINT is implemented as a relational database hosted on a PostgreSQL server and can be accessed over the Internet through an HTML web interface.

Results

11103 of the 45840 interactions (24%) have one or more homologs in HINT. Table 1 shows the distribution of the homologs across species. The web interface can be used to search interactions using various identifiers such as SwissProt Accession numbers, PIR IDs, GenBank Accession numbers, RefSeq IDs or descriptions of the interacting proteins. An interaction of interest can be selected from the results of the search to obtain detailed information about it. Figure 2 shows a snapshot of the Interactions web page. The homologs of the interaction selected are shown in graphical form as well as tabular form and sorted according to the score of the protein hits, with the best hits shown first. The graphical form helps the user to visualize the regions and domains that are common among the proteins of the

selected interaction and those of the homologous interactions. The tabular form gives details about the E-values and the percent identity given by PSIBlast. Further details about the usage of the web interface are provided in the form of an online Help document.

Discussion

HINT is a database of homologous protein-protein interactions that can be used by researchers to determine the detailed information about similar interactions in different species. It provides a graphical view of the homologous interactions and the various annotations of the interacting proteins and can be accessed over the Internet at <http://helix.protein.osaka-u.ac.jp/hint/>. For a given interaction, HINT is

Table 1 Species distribution of homologous protein-protein interactions in HINT

Organism	Two letter code	Interactions	Homologous Interactions
<i>D. melanogaster</i>	Dm	20581	3521
<i>S. cerevisiae</i>	Sc	14178	3879
<i>C. elegans</i>	Ce	4553	1165
<i>H. sapiens</i>	Hs	3933	2126
<i>H. pylori</i>	Hp	1409	128
<i>E. coli</i>	Ec	554	150
<i>M. musculus</i>	Mm	483	292
<i>A. thaliana</i>	At	76	65
<i>R. norvegicus</i>	Rn	67	43
<i>S. pombe</i>	Sp	6	4
Total		45840	11103

HINT

Homologous Interactions Database

Interaction 11596 [Help](#)
Species: *Drosophila melanogaster* **Method:** classical two hybrid **Pubmed ID:** [14605208](#) **Source DB:** [IntAct](#) **Confidence:** high

Interacting Proteins	External References	Structure (s)	COG	InterPro	GO
CG9952-PA S416	7291460 Q9W214		KOG1947 Leucine rich repeat proteins, some proteins contain F-box	IPR001611 - Leucine-rich repeat IPR001810 - Cyclin-like F-box IPR007089 - Leucine-rich repeat, cysteine-containing	
EG:115C2.4 protein 1803	17861930 O77430		KOG1724 SCF ubiquitin ligase, Skp1 component	IPR001232 - SKP1 component IPR008945 - Skp1-Skp2 dimerisation	

Homologous interactions

[Help](#)

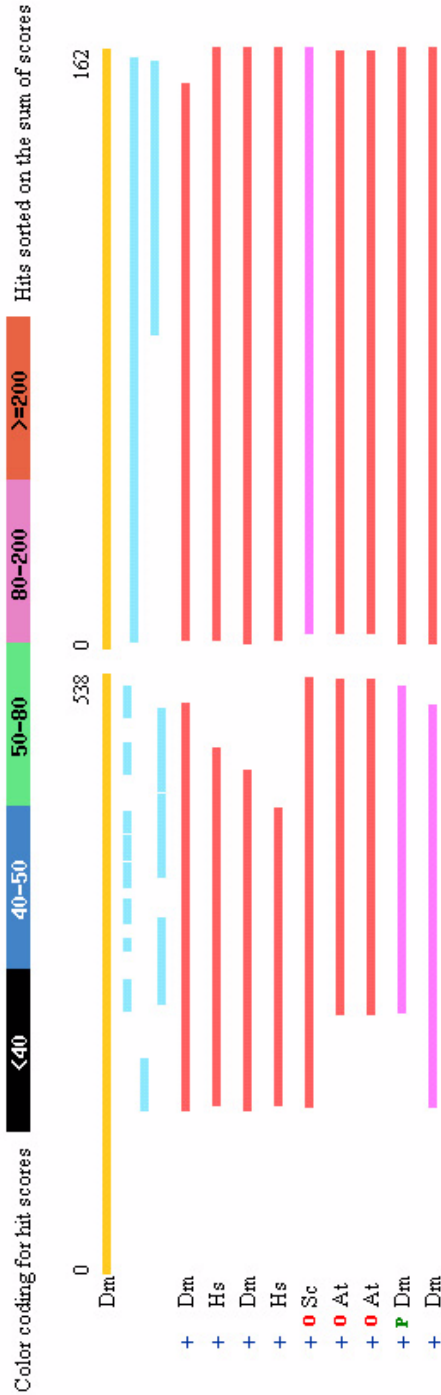


Figure 2 Web Interface of the Database of Homologous Interactions with interaction details and GO, InterPro and KOG annotations of the interacting proteins. Also shown is a graph of the homologs with the query interaction in yellow, InterPro domains of the interacting proteins in blue and the homologs of the proteins colored differently with varying hit scores, similar to BLAST results¹⁰. The species in which the homolog is found is given by a 2 letter code as given in Table 1. If the interaction is an ortholog or a paralog of the query interaction, it is indicated by an 'O' or a 'P' along side each homolog. Clicking on the '+' gives further details of the homolog such as E-value, percent identity, hit score and KOG, where available. Tool tips are provided in various places.

able to provide a list of similar interactions found in the same or in different species. This is of considerable use in comparative genomic analyses. In future, we plan to use HINT in the identification of true positives in high-throughput interaction data sets and in the formation of interaction families.

Acknowledgments

This study has been supported by grant-in-aid from Institute for Bioinformatics Research and Development, Japan Science and Technology Agency and by grant-in-aid for Scientific Research on priority areas No. 12144206 from the Ministry of Education, Science, Sports and Culture of Japan.

References

1. Aloy, P. & Russell, R. B. Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* **22**, 1317–1321 (2004).
2. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
3. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. & Apweiler, R. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
4. Bader, G. D., Betel, D. & Hogue, C. W. V. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
5. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
6. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. & Natale, D. A. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
7. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J. A., Vaughan, R. & Zdobnov, E. M. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318 (2003).
8. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
9. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. & Apweiler, R. The HUPO PSI's Molecular Interaction format — a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
10. <http://www.ncbi.nlm.nih.gov/BLAST/>.