Human Genomics

REVIEW Open Access

# GWIDD: a comprehensive resource for genome-wide structural modeling of protein-protein interactions

Petras J Kundrotas[1], Zhengwei Zhu[1,3] and Ilya A Vakser[1,2*]

## Abstract

Protein-protein interactions are a key component of life processes. The knowledge of the three-dimensional structure of these interactions is important for understanding protein function. Genome-Wide Docking Database (http://gwidd.bioinformatics.ku.edu) offers an extensive source of data for structural studies of protein-protein complexes on genome scale. The current release of the database combines the available experimental data on the structure and characteristics of protein interactions with structural modeling of protein complexes for 771 organisms spanned over the entire universe of life from viruses to humans. The interactions are stored in a relational database with user-friendly interface that includes various search options. The search results can be interactively previewed; the structures, downloaded, along with the interaction characteristics.

**Keywords:** Protein-protein interactions, Structural modeling, Protein docking, Structural genomics, Interactome

## Introduction

Proteins function by interacting with other biologically relevant molecules. Understanding the mechanisms of protein-protein interactions (PPI) is essential for studying life processes at the molecular level. Genome sequencing provided a vast amount of information on proteins at the sequence level. Currently, efforts focus on the function assignment of these proteins based on their three-dimensional (3D) structures and interactions. Interaction maps for specific organisms and biochemical pathways need to be complemented by the structural information. Experimental techniques are limited in their ability to produce the structures on the genome scale. Thus, computational methods are essential for this task [1].

Structural modeling of PPI has its origins in *ab initio* techniques based on shape and physicochemical complementarity. More recent approaches take advantage of statistical potentials and machine learning [2,3]. Despite progress in development of such template-free algorithms, their accuracy in the high-throughput structure determination is limited.

Rapidly increasing amount of data on PPI makes possible application of the template-based methods. Such approaches are based on the observation that monomers with similar sequences and/or structures, generally, have similar binding modes. Several groups assessed the quality of PPI modeling based on sequence alignment to complexes with known structure [4-9]. Studies showed that the majority of such homology-docking models are of acceptable and medium quality, according to the established criteria [3]. An alternative template-based approach takes advantage of the structural similarity between the target and the template complexes [10-13].

The progress in 3D modeling of PPI is reflected in the Genome-Wide Docking Database (GWIDD) [14], which provides annotated collection of experimental and modeled PPI structures from the entire universe of life spanning from viruses to humans. The resource has user-friendly search interface, providing preview and download options for experimental and modeled PPI structures.

## Database design

GWIDD imports PPI from external sources, including the last free release of BIND [15] and DIP [16,17].

* Correspondence: vakser@ku.edu
[1]Center for Bioinformatics, The University of Kansas, 2030 Becker Dr., Lawrence, KS 66047, USA
[2]Department of Molecular Biosciences, The University of Kansas, 2030 Becker Dr., Lawrence, KS 66047, USA
Full list of author information is available at the end of the article

Currently, we are working on interfacing GWIDD with MINT [18], BioGRID [19], and IntAct [20]. To provide the structures to PPI, the following scheme is utilized. If the complex is found in the Protein Data Bank (PDB), the X-ray structure is used, and no modeling is performed (10,924 GWIDD entries). Otherwise, a search for a pair of homologous sequences from complexes with known structure is performed, and the model is built by homology docking [6,7]. Statistical significance of the sequence alignments is assigned [7], with an additional requirement that both alignments contain at least 80% of the target sequences. This provides structures for 12,646 PPI. For the interactions not covered by these two steps, the interacting monomers are modeled independently by homology modeling, with subsequent docking of the models by structural alignment [12]. Incorporation of the structural alignment predictions (28,811 entries) into GWIDD is currently in progress (the structures are available from the authors by request). The graphical summary of the GWIDD coverage of genomes is in Figure 1.

## User interface

The database (http://gwidd.bioinformatics.ku.edu) user interface (Figure 2) offers search by keywords, sequences (explicit input or upload in FASTA format), or structures (upload in PDB format), for one or both interacting proteins. The search by keywords can be performed using any word in the protein description (name of organism, cellular location, biological function etc.) or by selection from drop-down menus that are listing organisms currently in GWIDD. Repeated selection of the box 'Add another organism to the list' allows expansion of the search to several organisms. An option for search by standard taxonomy identification (ID) with link to taxonomy database http://www.uniprot.org is also provided. In case of input PDB file, the sequence is extracted from SEQRES tags or, if the SEQRES is not available, from ATOM tags of $C^{\alpha}$ atoms. The sequences from different sources can differ in length even for the same protein (e.g., due to unresolved fragments of the X-ray structure). Thus, advanced sequence search options are available. Figure 2 shows an example of search by organism.

The user can enable the second half of the search interface if information related to the interaction partner is available ('protein B,' Figure 2). The search results can be filtered by the structure availability (experimental, modeled, or no structures). Online help is provided in pop-up windows. The search result screen displays all interactions in the database satisfying the input search criteria in the form of an expandable list of GWIDD interaction IDs. For the homology-docking models, the alignments used to build the model are provided, and the model quality is assessed by the sequence identity criteria [5]. Links are provided to download the PDB-format files, along with the text file containing relevant information. Visualization screen is available to display the structures by different interactive representations. A link is provided to download the entire set of sequence-homology models in one gzipped archive.

## Implementation

GWIDD unifies different external PPI data formats into a single data set, removing redundancy and retaining common data fields for all the sources. The

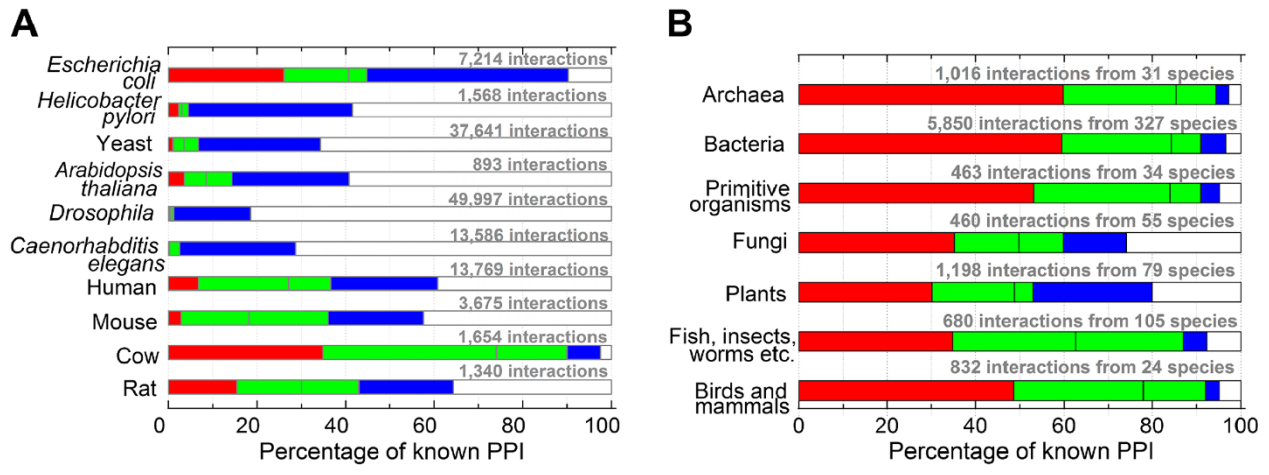


**Figure 1 Structural coverage of different genomes in GWIDD.** X-ray structures of complexes are in *red*, sequence-based models of complexes are in *green*, and interactions with structural models of the monomers are in *blue*. (**A**) Ten genomes with the largest number of known PPI. (**B**) The rest of the genomes (the data from A excluded).

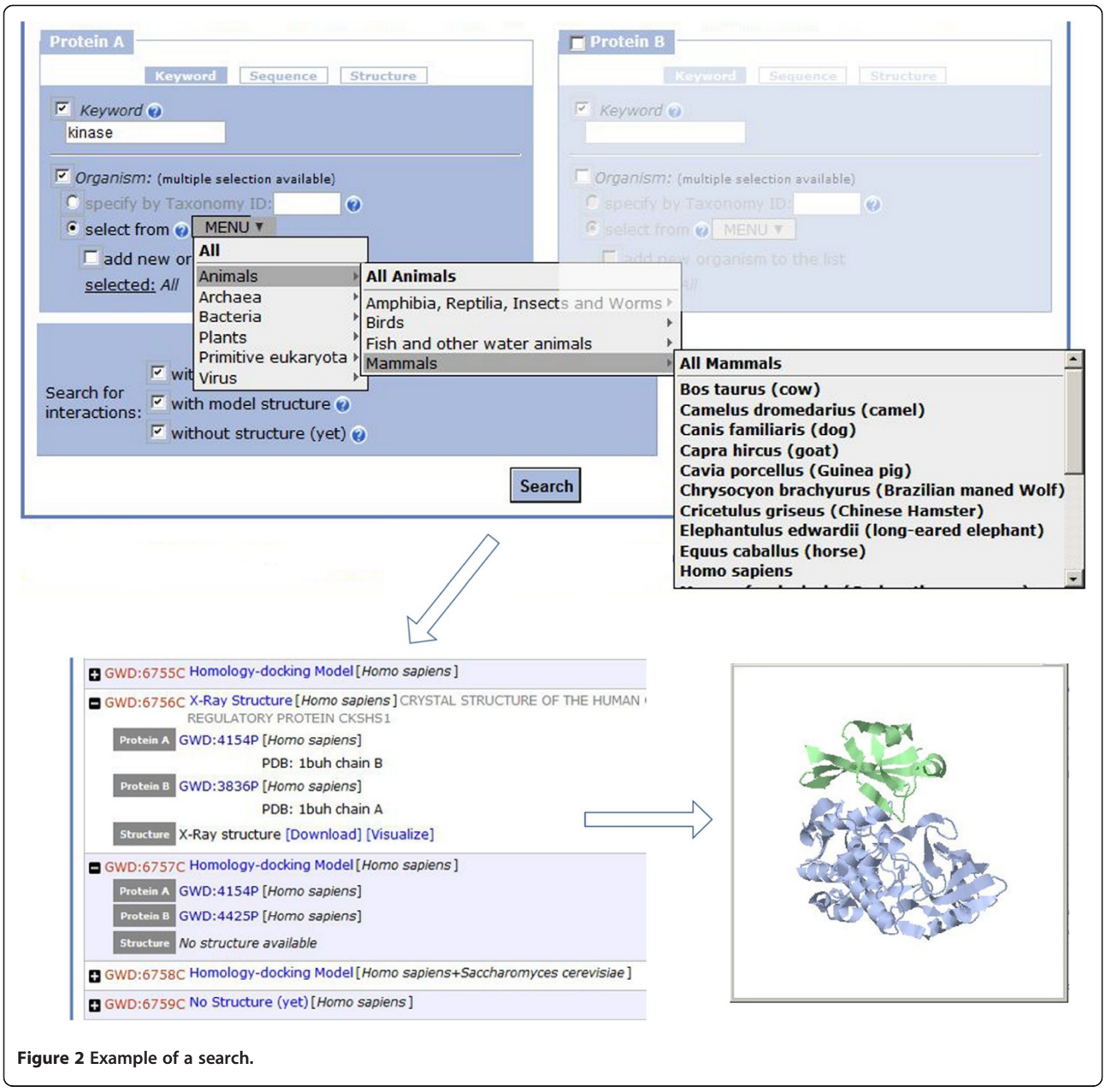


**Figure 2** Example of a search.

interaction data are stored in a relational database, except for large files, such as structure coordinates, which are stored directly in the file system and are linked from the relational database. The web interface is implemented on the Linux-Apache-PostgreSQL -PHP software stack. Web user interface is built using hypertext preprocessor (PHP) and jQuery library, where PHP is for web presentation and logic as well as back-end database access; jQuery is responsible for AJAX and other JavaScript-based dynamic features. Visualization of protein structures is implemented in Jmol (www.jmol.org). Homology docking was performed by NEST [21], BLAST [22], and in-house profile-to-profile alignment program. The procedures are joined by Python scripts.

## Future directions

GWIDD development will incorporate other structural modeling techniques, such as multi-template/threading modeling of interacting proteins, partial structural alignment [12], and template-free docking by GRAMM [23-25]. A major expansion of GWIDD will be the incorporation of new PPI sources from other publicly available PPI databases. Large-scale systematic benchmarking of the high-through put modeling will be used to assign a confidence score to the modeled structures.

**Competing interests**
The authors declare that they have no competing interests.

**Acknowledgments**
Andrey Tovchigrechko and Tatiana Baronova made important contributions to the GWIDD project at the earlier stages of development. This work was supported by the National Institutes of Health grant R01 GM074255.

**Author details**
[1]Center for Bioinformatics, The University of Kansas, 2030 Becker Dr., Lawrence, KS 66047, USA. [2]Department of Molecular Biosciences, The University of Kansas, 2030 Becker Dr., Lawrence, KS 66047, USA. [3]Current address: Department of Genetics, Room 716B, Abramson Research Center, University of Pennsylvania, 3615 Civic Center Blvd., Philadelphia, PA 19104, USA.

**Authors' contributions**
ZZ performed calculations and implemented the web interface. PJK developed calculation pipelines, designed the web interface, analyzed data, and drafted the manuscript. IAV designed the research, analyzed data, and wrote the paper. All authors read and approved the final manuscript.

Received: 27 June 2012 Accepted: 11 July 2012
Published: 11 July 2012

**References**

1. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A: **A structural perspective on protein–protein interactions.** *Curr Opin Struct Biol* 2004, **14:**313–324.
2. Vakser IA, Kundrotas P: **Predicting 3D structures of protein-protein complexes.** *Curr Pharm Biotech* 2008, **9:**57–66.
3. Lensink MF, Wodak SJ: **Docking and scoring protein interactions: CAPRI 2009.** *Proteins* 2010, **78:**3073–3084.
4. Aloy P, Pichaud M, Russell RB: **Protein complexes: structure prediction challenges for the 21st century.** *Curr Opin Struct Biol* 2005, **15:**15–22.
5. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology.** *Proc Natl Acad Sci USA* 2002, **99:**5896–5901.
6. Kundrotas PJ, Alexov E: **Predicting 3D structures of transient protein-protein complexes by homology.** *Bioch Biophys Acta* 2006, **1764:**1498–1511.
7. Kundrotas PJ, Lensink MF, Alexov E: **Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles.** *Int J Biol Macromol* 2008, **43:**198–208.
8. Lu L, Lu H, Skolnick J: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.** *Proteins* 2002, **49:**350–364.
9. Mukherjee S, Zhang Y: **Protein-protein complex structure predictions by multimeric threading and template recombination.** *Structure* 2011, **13:**955–966.
10. Gunther S, May P, Hoppe A, Frommel C, Preissner R: **Docking without docking: ISEARCH - prediction of interactions using known interfaces.** *Proteins* 2007, **69:**839–844.
11. Keskin O, Nussinov R, Gursoy A: **PRISM: protein-protein interaction prediction by structural matching.** *Methods Mol Biol* 2008, **484:**505–521.
12. Sinha R, Kundrotas PJ, Vakser IA: **Docking by structural similarity at protein-protein interfaces.** *Proteins* 2010, **78:**3235–3241.
13. Korkin D, Davis FP, Alber F, Luong T, Shen M, Lucic V, Kennedy MB, Sali A: **Structural modeling of protein interactions by analogy: application to PSD-95.** *PLoS Comp Biol* 2006, **2:**1365–1376.
14. Kundrotas PJ, Zhu Z, Vakser IA: **GWIDD: genome-wide protein docking database.** *Nucl Acid Res* 2010, **38:**D513–D517.
15. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucl Acid Res* 2005, **33:**D418–D424.
16. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucl Acid Res* 2004, **32:**D449–D451.
17. Xenarios I, Rice DW, Salwinski L, Baron NK, Marcotte EM, Eisenberg D: **DIP: the Database of Interacting Proteins.** *Nucleic Acids Res* 2000, **28:**289–291.
18. Ceol A, Aryamontri AC, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2009 update.** *Nucl Acid Res* 2010, **38:**D532–D539.
19. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M: **The BioGRID Interaction Database: 2011 update.** *Nucl Acid Res* 2011, **39:**D698–D704.
20. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H: **The IntAct molecular interaction database in 2010.** *Nucl Acid Res* 2010, **38:**D525–D531.
21. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IY, Alexov E, Honig B: **Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling.** *Proteins* 2003, **53:**430–435.
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of database programs.** *Nucleic Acids Res* 1997, **25:**3389–3402.
23. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA: **Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.** *Proc Natl Acad Sci USA* 1992, **89:**2195–2199.
24. Vakser IA, Matar OG, Lam CF: **A systematic study of low-resolution recognition in protein-protein complexes.** *Proc Natl Acad Sci USA* 1999, **96:**8477–8482.
25. Tovchigrechko A, Wells CA, Vakser IA: **Docking of protein models.** *Protein Sci* 2002, **11:**1888–1896.

doi:10.1186/1479-7364-6-7
**Cite this article as:** Kundrotas *et al.*: GWIDD: a comprehensive resource for genome-wide structural modeling of protein-protein interactions. *Human Genomics* 2012 **6**:7.