SOURCE CODE FOR BIOLOGY AND MEDICINE

SOFTWARE REVIEW Open Access

Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information

Chandra Sekhar Pedamallu, Janos Posfai*

Abstract

Background: Protein-protein interactions are crucially important for cellular processes. Knowledge of these interactions improves the understanding of cell cycle, metabolism, signaling, transport, and secretion. Information about interactions can hint at molecular causes of diseases, and can provide clues for new therapeutic approaches. Several (usually expensive and time consuming) experimental methods can probe protein - protein interactions. Data sets, derived from such experiments make the development of prediction methods feasible, and make the creation of protein-protein interaction network predicting tools possible.

Methods: Here we report the development of a simple open source program module (*OpenPPI_predictor*) that can generate a putative protein-protein interaction network for target genomes. This tool uses the orthologous interactome network data from a related, experimentally studied organism.

Results: Results from our predictions can be visualized using the *Cytoscape* visualization software, and can be piped to downstream processing algorithms. We have employed our program to predict protein-protein interaction network for the human parasite roundworm *Brugia malayi*, using interactome data from the free living nematode *Caenorhabditis elegans*.

Availability: The *OpenPPI_predictor* source code is available from http://tools.neb.com/~posfai/.

Introduction

The cell is the structural and functional unit of living organisms. Cells carry out numerous functions, from DNA replication, cell replication, protein synthesis, and energy production to molecule transport, to various inter- and intra-cellular signaling. Many of these fundamental processes require cascades of biochemical reactions that are catalyzed by interacting protein enzymes. Other interacting proteins provide structural support for the cells, form scaffolds for intracellular localization, and serve as chaperones or as transporters. The large-scale study of all cellular proteins is known as proteomics [1,2]. Since aspects of protein function can be inferred from the protein's complex interactions, from its position in interaction networks, one of the main goals of proteomics is to map the interactions of proteins. Uncovering protein-protein interaction information is a major undertaking in basic biological research, helps in the discovery of novel drug targets for the treatment of various diseases. Interaction networks (interactomes) for many model organisms have been established experimentally. Experimental probing of protein-protein interactions requires labor-intensive techniques, such as co-immunoprecipitation, or affinity chromatography [3]. High-throughput experimental techniques, such as yeast two-hybrid screens [4] and mass spectrometry [5] are also available for large-scale detection of protein-protein interactions, for the exploration of protein's amino acid sequences, their structures, and relationships [3]. Following these advances, numerous computational methods have been developed to predict protein-protein interaction networks. These use or combine phylogenetic profiling [6], homologous interacting partner analysis [7], structural pattern comparisons [8-10], bayesian network modeling [11], literature mining [12], codon usage analysis [13], and so on. Surveys on computational methods for prediction of protein-protein interactions

* Correspondence: posfai@neb.com
New England Biolabs, Ipswich, MA, USA

are available in the literature [3,14]. Complementing efforts centralize protein-protein interaction data through the construction of databases, such as *STRING* [15], *MINT* [16], *BioGRID* [17], *DIP* [18], *POINT* [19] and *IntAct* [20].

Most of these reviewed prediction methods are implemented as web servers, which are convenient for the in-depth analysis of selected nodes and features, but offer little flexibility when the prediction of a complete cellular interactome is an intermediate goal, embedded in an involved discovery scheme. In this paper, we report the development of a simple open source tool (*OpenPPI_predictor*) for such intermediate role. The tool predicts the complete protein-protein interaction network for target genomes, using interactome data from related organisms (i.e. reference genomes). For further analysis, the generated putative interactome can be visualized using the *Cytoscape* software [21], and can be forwarded to follow-up program modules. We have developed this program to predict the protein-protein interaction network of the human parasite *B. malayi*. The predictions rely on the available interactome data of the close relative nematode *C. elegans*. The predicted number of interactions, and types of interacting partners, the distinguishing features from the human interactome guide our wet lab researchers in the selection of protein targets which seem essential for the parasite, so blocking them would disrupt its cell cycle, yet the intervention would not interfere with human protein complexes.

Design and Implementation

This tool comprises of two modules: (a) Ortholog (diverged from the same immediate ancestor) protein identifier, and (b) Protein - Protein interaction predictor. The tool requires four kinds of inputs:

i. Sequences of proteins from the reference genome,
ii. Interactome for the reference genome (also called as orthologous interactome),
iii. Sequences of proteins from the genome of interest.
iv. Protein ortholog assignments between organism of interest and reference organism.

The ortholog protein identifier extracts information from an ortholog database, and makes connections across the reference genome and the genome of interest. The output from this module is a list of connections between proteins in the genome of interest and their corresponding orthologous relatives in the reference genome.

The protein - protein interaction predictor module uses the already known interactome of the reference genome. Interactions in the reference set are projected back to the corresponding orthologous proteins of the genome of interest.

More formally, the workflow of our method is as follows: assume we have two query proteins Q1 and Q2, with corresponding orthologous proteins R1 and R2 in the reference genome. If R1 and R2 interact in the reference organism, then the prediction is made, that Q1 and Q2 also interact. Knowledge about the relationship of R1 and R2 are transferred to a predicted relationship between Q1 and Q2.

Figure 1 describes the overall implementation in *OpenPPI_predictor* tool. The algorithm used in this pipeline is divided into following steps:

Step 1: Create the ortholog connections between reference genome and genome of interest sequence, using ortholog identifier component from *OrthoMCL-DB*.

Step 2: Use protein-protein interaction predictor to predict interactome for genome of interest from interaction data in the reference genome, and the ortholog connections created in Step 1. The predicted interactome is in format that is compatible to *Cytoscape* software.

Step 3: Use *Cytoscape* software to visualize and analyze the predicted interactome generated in Step 2.

OpenPPI_predictor is implemented in AWK and C shell language and installed on Linux. The source code can be downloaded from http://tools.neb.com/~posfai/. The program has been tested by generating predictions for pairs of yeast and mammalian model organisms, using several resources listed in the Introduction.

Results and Discussion

We have used *OpenPPI_predictor* to predict the *B. malayi* protein - protein interaction network. Genome and proteome data was fetched from NCBI (http://ncbi.nlm.nih.gov) ortholog assignments from *OrthoMCL-DB* [22], while reference *C. elegans* interactome data was downloaded from Worm Interactome Database [23].

The filarial nematode *B. malayi* is a human parasite. It causes elephantiasis, a wide spread and devastating disease, characterized by swelling of the lower limbs. Other filarial parasites, *Wuchereria bancrofti* and *Brugia timori* are also widespread, and cause serious diseases. Though these latter organisms differ from *B. malayi* morphologically, symptomatically, and in geographical extent [24], our target selection method can be followed in their cases as well. *C. elegans* is a free living nematode, and one of the most studied organisms, with available experimental genome, proteome, and interactome data. *C. elegans* interactome is used here to predict the *B. malayi* interactome, because of the high level of genomic conservation between these species [25]. The *C. elegans* interactome is composed of 178151 interactions, from the 20100 proteins encoded in the genome.

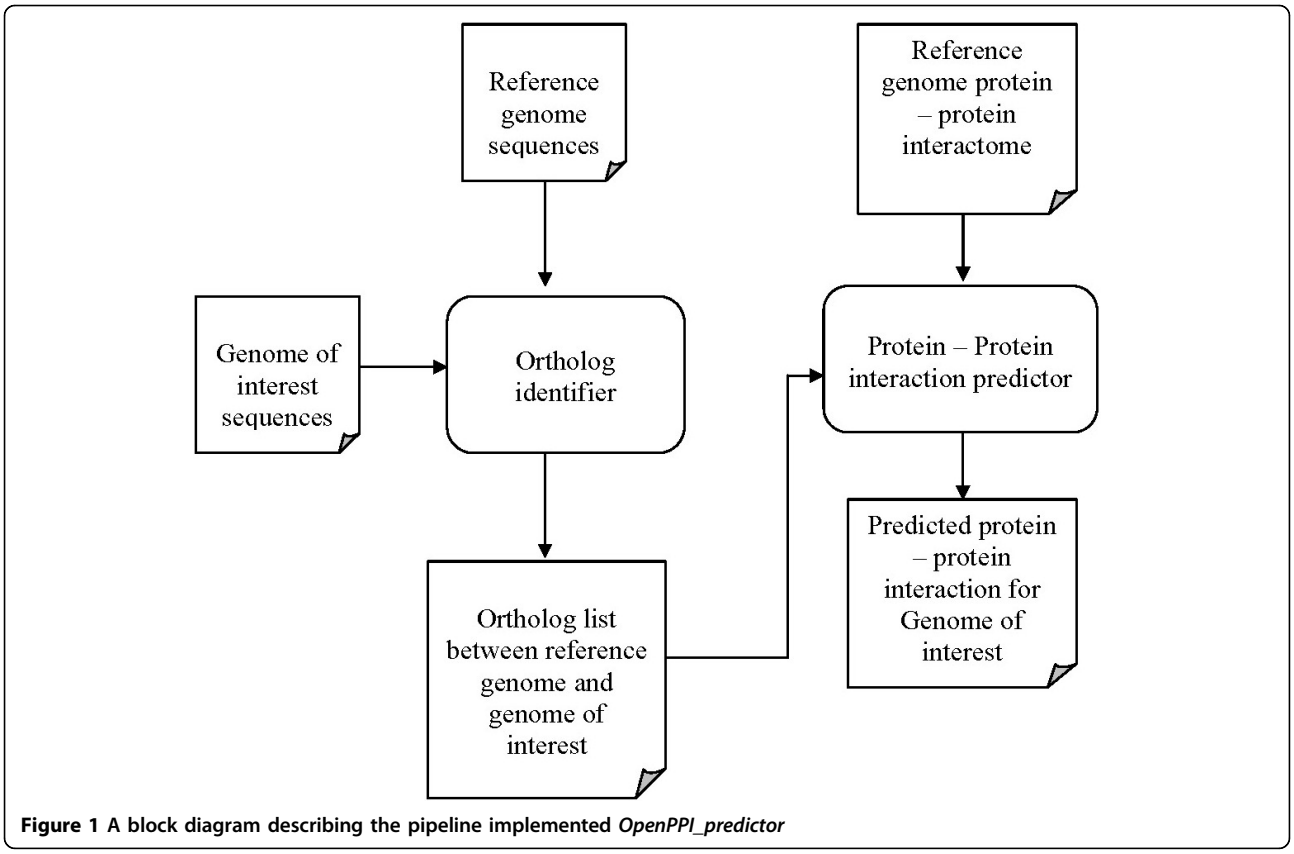

Figure 1 A block diagram describing the pipeline implemented *OpenPPI_predictor*

The interactions have been established through large scale projects using different methods.

Orthology resources typically employ all-versus-all BLASTP analysis (Washington University, http://blast.wustl.edu), followed by some form of clustering (Jaccard clustering, bidirectional best hit clustering; [25]). Some tools, including *ProGMap* [26], *Berkeley PHOG* [27], *TdrTargets* [28], *BLASTO* [29], use additional, complementing sequence and structural information to identify orthologs across multiple organisms. Several ortholog databases have been compiled, using variants of the above procedures. For ortholog information between *C. elegans* and *B. malayi* genome we considered the Clusters of Orthologous Groups of proteins (*COG*s, [30]), and the Princeton Protein Orthology Database (*P-POD*, [31]), but settled on the more up-to-date and more accessible *OrthoMCL-DB* database ([22], http://www.orthomcl.org/common/ downloads/).

Figure 2 and Figure 3 illustrate the *C. elegans* interactome and the predicted *B. malayi* interactome, using *Cytoscape* software. The predicted *B. malayi* interactome is composed of 164187 interactions from 11460 protein coding sequences. From our predictions, the *B. malayi* interactome seems sparser then the *C. elegans* interactome. This difference may be due to the fact, that *B. malayi* is a parasite, which exploits a host organism, hijacks some of its functions, metabolites, and processes. Incompleteness of the *B. malayi* genome sequence, and also the limited accuracy in the identification of ortholog relationships across *C. elegans* and *B. malayi* may contribute to sparseness.

For a post-prediction analysis, we have used *Mcode* [32] to find clusters (highly connected regions) in the interaction network. Such clusters often correspond to protein complexes, and are parts of distinct metabolic pathways. *Mcode* identifies 118 and 143 clusters in *B. malayi* and *C. elegans* interactomes, respectively. The highly connected region contains 363 and 340 proteins in *B. malayi* and *C. elegans* interactome. This observation suggests that core cellular functions of the two related organisms have similar complexity. Figure 4 illustrates the distribution of clusters and number of cluster members. Further analysis of these highly connected regions may provide clues about genes missing from a conserved pathway, or proteins missing from a complex. The predicted interactome could be used to attribute protein function as well [33-35].

The utility of predictions depends on several factors. The establishment of orthology (relatedness through descent from the same common ancestor) carries less uncertainty, if a closely related reference organism can be found. Data from multiple related reference organisms

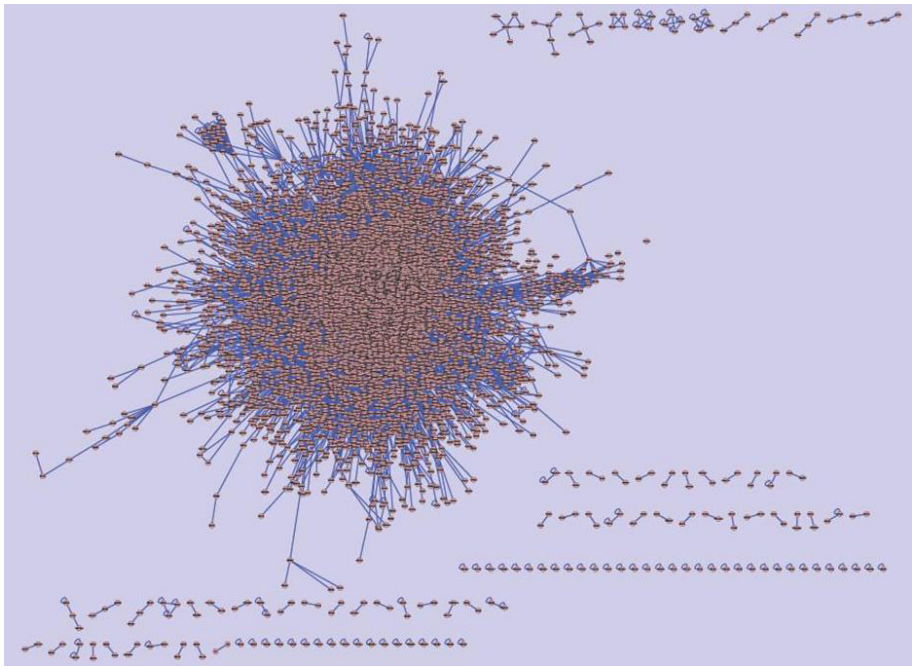

Figure 2 ***C. elegans*** **interactome**

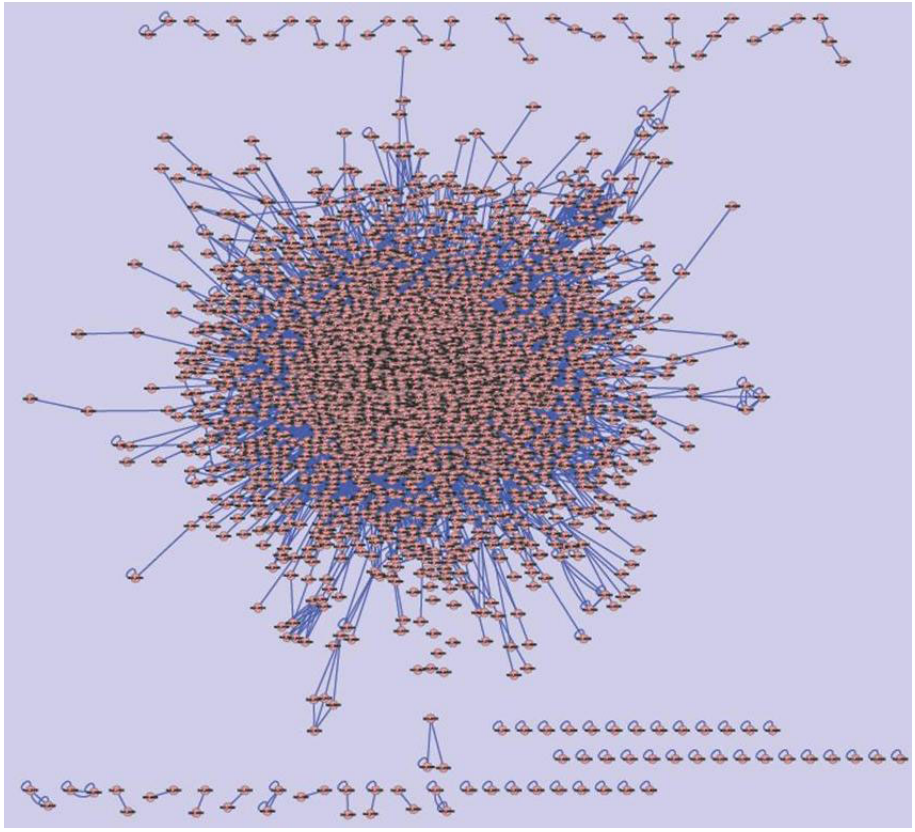

Figure 3 **Predicted** ***B. malayi*** **interactome**

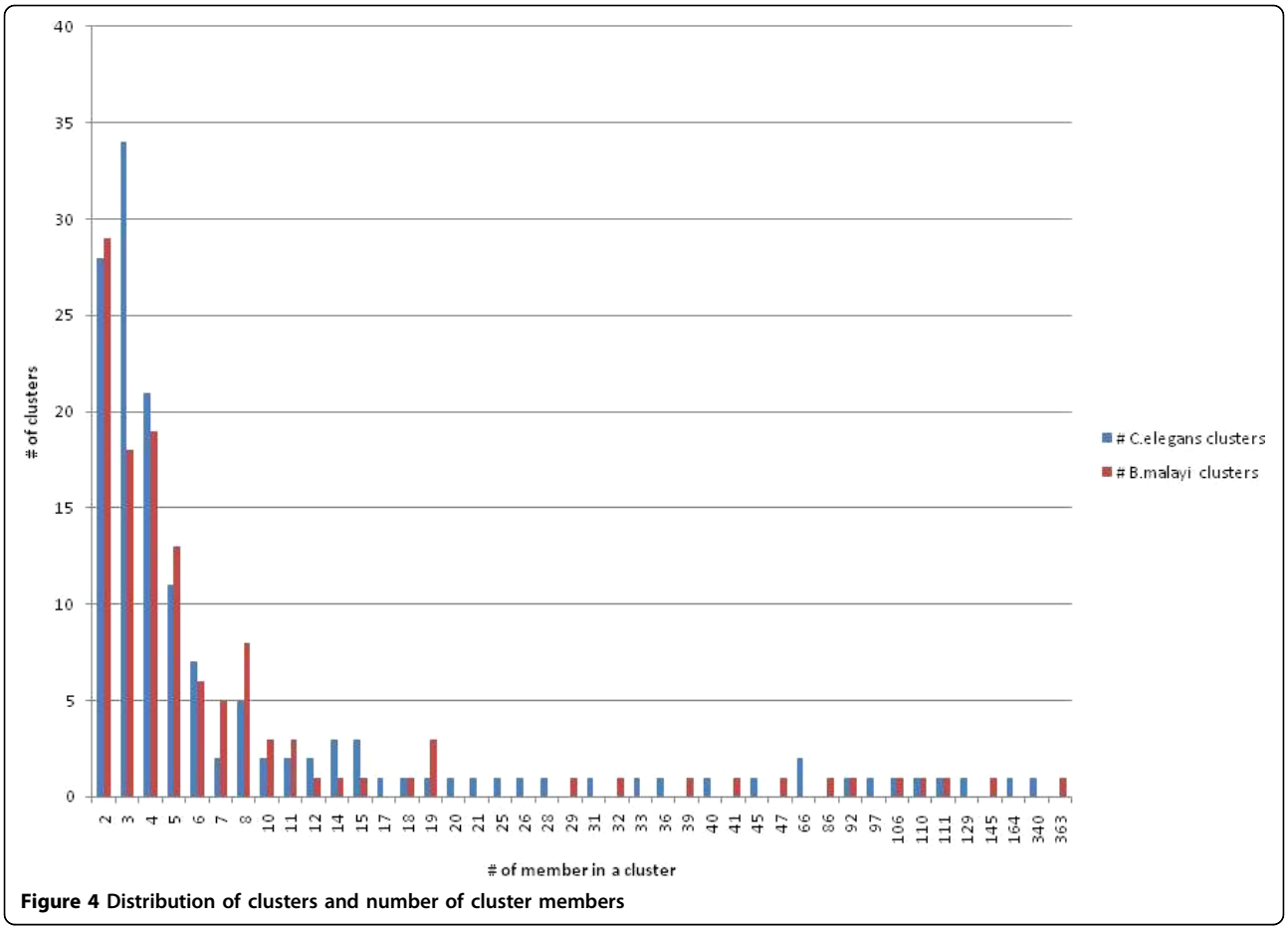

Figure 4 Distribution of clusters and number of cluster members

would increase the signal to noise ratio, and improve the value of predictions. Experimentally verified test tube interactions may not be projected unconditionally to in vivo conditions: *i*. proteins interacting in a screen may not co-exist or co-localize in the living cell, they may be synthesized in different phases of the cell cycle, or they can be transported to different intracellular compartments, *ii*. post-translationally modified, mutated, alternatively spliced proteins may not interact with the same partners, *iii*. presence and binding of co-factors can change protein structure, hence interaction partnerships, *iv*. quorum signals can turn on and off interactions in bacteria, *v*. cell type and expression levels can modify interactions, *vi*. non-binary effects appear. Since the prediction tool uses such uncertain data, we should expect a degree of uncertainty in our predictions, and the results should be considered putative.

Conclusions

Here we report the development of the *OpenPPI_predictor* tool. The tool predicts the protein interactome for a genome of interest, using the interactome data from a closely related organism, and protein orthology information between the two species. The tool is designed for genome wide interactome predictions, and provides a simple, flexible and easy to use platform for proteomic research.

Making predictions about possible protein-protein interactions is only an intermediate step in understanding protein function or in the search of drug targets. Upstream and downstream steps, biochemical and physiological considerations (many listed in earlier paragraphs) in finding applicable datasets, in filtering input data, in interpreting results and in drawing inferences make only the predictions relevant.

In the future, we plan to enhance both the utility and the coverage of our predictions using data from multiple related organisms, taking into account the phylogenetic distances between the interrogated pairs. We plan on ranking, or categorizing the predicted interactions according the consistency with which the predictions appear in the pair-wise predictions.

Acknowledgements

We wish to thank Kshitiz Chaudhary and Tilde Carlow at New England Biolabs for scientific discussions. We also thank Tamas Vincze at New England Biolabs for technical support.

Authors' contributions
CSP wrote source code for *openPPI_predictor*. CSP and JP wrote the manuscript. All authors have read and approved the final manuscript.

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

Received: 12 March 2010 Accepted: 4 August 2010
Published: 4 August 2010

References

1. Anderson NL, Anderson NG: **Proteome and proteomics: new technologies, new concepts, and new words.** *Electrophoresis* 1990, **19(11)**:1853-61.
2. Blackstock WP, Weir MP: **Proteomics: quantitative and physical mapping of cellular proteins.** *Trends Biotechnol* 1999, **17(3)**:121-7.
3. Skrabanek L, Saini HK, Bader GD, Enright AJ: **Computational prediction of protein-protein interactions.** *Mol Biotechnol* 2008, **38(1)**:1-17.
4. Young KH: **Yeast Two-Hybrid: So Many Interactions, (in) So Little Time.** *Biology of Reproduction* 1998, **58**:302-311.
5. Figeys D, McBroom LD, Moran MF: **Mass Spectrometry for the Study of Protein-Protein Interactions.** *Methods* 2001, **24(3)**:230-239.
6. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-8.
7. Aloy P, Russell RB: **InterPreTS: Protein Interaction Prediction through Tertiary Structure.** *Bioinformatics* 2003, **19(1)**:161-162.
8. Aytuna AS, Keskin O, Gursoy A: **Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.** *Bioinformatics* 2005, **21(12)**:2850-2855.
9. Keskin O, Ma B, Nussinov R: **Hot regions int protein-protein interactions: The organization and contribution of structurally conserved hot spot residues.** *J Mol Biol* 2004, **345**:1281-1294.
10. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A: **PRISM: protein interactions by structural matching.** *Nucl Ac Res* 2005, , **33 Web Server:** W331-336.
11. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644)**:449-453.
12. Li Y, Hu X, Lin H, Yang Z: **Learning an enriched representation from unlabeled data for protein-protein interaction extraction.** *BMC Bioinformatics* 2010, **2**:S7.
13. Najafabadi HS, Salavati R: **Sequence-based prediction of protein-protein interactions by means of codon usage.** *Genome Biol* 2008, **9(5)**:R87.
14. Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, Golshani A: **Computational methods for predicting protein-protein interactions.** *Adv Biochem Eng Biotechnol* 2008, **110**:247-67.
15. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, , **37 Database:** D412-6.
16. Chatraryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res* 2007, **35**:D572-574.
17. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**: D535-539.
18. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**: D449-D451.
19. Huang TW, Tien AC, Huang WS, Lee YC, Peng CL, Tseng HH, Kao CY, Huang C-YF: **POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome.** *Bioinformatics* 2004, **20(17)**:3273-3276.
20. 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* 2004, **32**: D452-455.
21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13(11)**:2498-504.
22. Chen F, Mackey AJ, Christian JStoeckert Jr, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34**:D363-8.
23. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, Chen WJ, Cunningham F, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Pai S, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg PW, Stein LD: **WormBase: a comprehensive data resource for Caenorhabditis biology and genomics.** *Nucleic Acids Res* 2005, , **33 Database:** D383-9.
24. John DT, William AP: *Markell and Voge's Medical Parasitology* St. Louis: Saunders Elsevier, 9 2006.
25. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, Angiuoli SV, Creasy T, Amedeo P, Haas B, El-Sayed NM, Wortman JR, Feldblyum T, Tallon L, Schatz M, Shumway M, Koo H, Salzberg SL, Schobel S, Pertea M, Pop M, White O, Barton GJ, Carlow CKS, Crawford MJ, Daub J, *et al*: **Draft genome of the filarial nematode parasite Brugia malayi.** *Science* 2007, **317(5845)**:1756-60.
26. Kuzniar A, Lin K, He Y, Nijveen H, Pongor S, Leunissen JA: **ProGMap: an integrated annotation resource for protein orthology.** *Nucleic Acids Res* 2009, , **37 Web Server:** W428-34.
27. Datta RS, Meacham C, Samad B, Neyer C, Sjölander K: **Berkeley PHOG: PhyloFacts orthology group prediction web server.** *Nucleic Acids Res* 2009, , **37 Web Server:** W84-9.
28. Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, Crowther GJ, Doyle MA, Hertz-Fowler C, Hopkins AL, McAllister G, Nwaka S, Overington JP, Pain A, Paolini GV, Pieper U, Ralph SA, Riechers A, Roos DS, Sali A, Shanmugam D, Suzuki T, Van Voorhis WC, Verlinde CL: **Genomic-scale prioritization of drug targets: the TDR Targets database.** *Nat Rev Drug Discov* 2008, **7(11)**:900-7.
29. Zhou Y, Landweber LF: **BLASTO: a tool for searching orthologous groups.** *Nucleic Acids Res* 2007, , **35 Web Server:** W678-W682.
30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41, Epub.
31. Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli SV, White O, Botstein D, Dolinski K: **The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists.** *PLoS One* 2007, **2(1)**: e766.
32. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4(1)**:2.
33. Letovsky S, Kasif S: **Predicting protein function from protein/protein interaction data: a probabilistic approach.** *Bioinformatics* 2003, **19**: i197-i204.
34. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting Protein Function and Protein-Protein Interactions from Genome Sequences.** *Science* 1999, **285(5428)**:751-753.
35. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks.** *Nature Biotechnology* 2003, **21**:697-700.

doi:10.1186/1751-0473-5-8
Cite this article as: Pedamallu and Posfai: **Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information.** *Source Code for Biology and Medicine* 2010 **5**:8.