Published online 11 June 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W147–W151***
doi:10.1093/nar/gks553

BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference

Javier Garcia-Garcia[1], Sylvia Schleker[2], Judith Klein-Seetharaman[2,3] and Baldo Oliva[1,*]

[1]Structural Bioinformatics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Research Park of Biomedicine (PRBB), 08003 Barcelona, Catalonia, Spain, [2]Forschungszentrum Jülich, Institute of Complex Systems (ICS-5), 52425 Jülich, Germany and [3]Department of Structural Biology, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received February 25, 2012; Revised May 16, 2012; Accepted May 17, 2012

ABSTRACT

Protein–protein interactions (PPIs) play a crucial role in biology, and high-throughput experiments have greatly increased the coverage of known interactions. Still, identification of complete inter- and intraspecies interactomes is far from being complete. Experimental data can be complemented by the prediction of PPIs within an organism or between two organisms based on the known interactions of the orthologous genes of other organisms (interologs). Here, we present the BIANA (Biologic Interactions and Network Analysis) Interolog Prediction Server (BIPS), which offers a web-based interface to facilitate PPI predictions based on interolog information. BIPS benefits from the capabilities of the framework BIANA to integrate the several PPI-related databases. Additional metadata can be used to improve the reliability of the predicted interactions. Sensitivity and specificity of the server have been calculated using known PPIs from different interactomes using a leave-one-out approach. The specificity is between 72 and 98%, whereas sensitivity varies between 1 and 59%, depending on the sequence identity cut-off used to calculate similarities between sequences. BIPS is freely accessible at http://sbi.imim.es/BIPS.php.

INTRODUCTION

The number of protein–protein interactions (PPIs) experimentally obtained for proteomes of entire species is smaller than the number of expected PPIs (1). Even the number of interactions between proteins in yeast, one of the most studied organisms, is still believed to be underestimated (2). To complement experimental data, several computational methods have been developed to predict PPIs (3,4). One example is the use of experimentally identified interactions in one organism to predict the interactions in other organisms assuming that homolog proteins preserve their ability to interact (5,6). The basis of this hypothesis is to assume that homologs have similar functional behaviour; therefore, they preserve the same PPIs. The prediction of the interaction between the homologs of two interacting proteins is defined as interolog (conservation of PPIs). Several works have used interologs to extend our knowledge about interactomes (7,8) or to predict pathogen/host cross-species PPIs (9,10). Up to now, most predictions based on interologs are restricted to a few species or a limited set of template interactions. For example, Yu *et al.* (5) transferred the known interactions of yeast to four different species. Similarly, Wiles *et al.* (11) developed InterologFinder, a tool to map the interactions integrated in MiMi (12) to five species. Interestingly, PPISearch (13) implements the interolog approach providing different scoring functions, but is restricted to the analysis of a single protein pair at any submission instance. Recently, Gallone *et al.* (14) developed a Perl module to automate predictions based on interologs, using optional metadata to rank the interactions. However, this still requires programming skills. The current limitations stem from the fact that information of known PPIs is spread among several repositories, and sets of PPIs from different databases show a low intersection (15,16). This challenge has led to the development of data integration strategies, such as Biologic Interactions and Network Analysis (BIANA) (17). BIANA is a program framework used in the integration of biological databases mostly focused on PPI databases.

Here, we present the BIANA Interolog Prediction Server (BIPS). BIPS offers a web interface to facilitate the prediction of PPIs based on interologs for a set of proteins provided by the user as input, including entire proteomes. The server benefits from the integration capabilities of BIANA to use a large data set of experimentally identified PPIs. BIANA also offers additional

*To whom correspondence should be addressed. Tel: +34 93 316 05 09; Fax: +34 93 316 05 50; Email: baldo.oliva@upf.edu

information such as gene ontology (GO) terms, clusters of orthologous genes and many other attributes such as predicted number of transmembrane helices, which gives the user the freedom to restrict the predictions according to selected features.

MATERIALS AND METHODS

The interolog hypothesis implies that two proteins (A and B) are predicted to interact if a known interaction between two proteins (A' and B') exists, such that A is similar to A' and B similar to B'. The interaction between the proteins A and B is called target interaction (with A and B being defined as protein targets), whereas the interaction between proteins A' and B' is called template interaction (with A' and B' being defined as templates). In the BIPS server, protein A is the query protein submitted by the user, and protein B is the predicted partner. This broad definition of interologs implies that the hypothesis works not only for orthologs but also for paralogs of the same species.

Sequence similarity measure

Sequence similarity between proteins relies on basic local alignment search tool (BLAST) alignments (18). *Query* protein sequences are aligned against all sequences with known interactions stored in the BIANA MySQL database (17). The alignments provide a similarity measure based on the percentage of identical residues aligned and the percentage of the sequence length of the *queries* and *templates* covered by the alignment (*query* and *template* coverage, respectively). We use a threshold of 90% of template coverage to ensure that the prediction is not inferred from local regions of the template interaction. The geometric mean of individual identities (joint identities) and the geometric mean of individual BLAST E-values (joint E-value) are also considered, as proposed by Yu *et al*. (5). The BIPS server uses a local database of stored similarity measures to avoid unnecessary repeated BLAST searches. This speeds up the server, allowing users to obtain predictions of interactions of full proteomes in manageable time. In this manner, only entirely new sequences consume extra time in the first run.

Domain interactions

We hypothesize that protein A interacts with protein B if a domain A' can be assigned to A and a domain B' to B, such that A' and B' are interacting domains in the iPfam (19) or the 3DID (20) database. We measure the similarity of the target sequences (A and B) with Pfam domains as a function of the E-value calculated with the package HMMER 3.0 (21). We assign Pfam domains using an E-value cut-off of 10^{-5} in the Pfam A database.

Source databases

Template interactions were extracted using the BIANA framework (17) integrating the following 10 databases: DIP (22), HPRD (23), IntAct (24), MINT (25), MPact (26), PHI_base (27), PIG (28), BioGRID (29), BIND (30) and VirusMINT (31). It has been noted that PPI databases share little overlap (16). Therefore, using the integration of multiple sources instead of a single source greatly enlarges the set of predictions. Furthermore, we have used BIANA to include information from other databases such as Uniprot (32) and GO (33). This additional information can be used to interpret the prediction results and select predicted interactions deemed interesting e.g. for experimental validation. Finally, sources of domain–domain interactions are also included, using iPfam (19) and 3DID (20).

Functional annotation

Interacting proteins likely share biological processes or share similar locations compared with non-interacting proteins (4). Thus, we can use a number of similar functional annotations between each pair of proteins predicted to interact to rank the predictions. BIPS uses GO annotations to select the most probable prediction for a query protein by selecting those partners that share similar GO terms. The similarity between GO terms implies that either they are equal or there is a parenthood relationship between them. In addition, GO term semantic similarity and the functional similarity of genes are computed using the method proposed by Wang *et al*. (34)

Clusters of orthologous genes

Two proteins are considered orthologous if they are included in the same cluster of orthologous genes. Ortholog definitions between proteins are extracted from eggNOG (35) and Ensembl Compara (36) databases. Our predictions can be filtered assuming the traditional definition of interologs: two target proteins are supposed to interact if they are orthologous to two known interaction partners.

DESCRIPTION OF THE WEBSERVER

Input

Proteins for which the user wants to predict putative binding partners can be uploaded as a list of sequences in FASTA format or a list of protein identifiers (i.e. UniProt Accession, Uniprot entry and gene symbol, etc.).

Output

The output is a list of predictions that can be viewed or downloaded. The user can browse the data associated with the predicted partners. Some details of the template interaction, such as the source database of the interaction and the method of detection, are provided. The user can select several parameters helping estimate the reliability of the predictions: (i) sequence similarity measures, (ii) checking domain–domain interactions (either using domains from 3DID or iPfam), (iii) checking common GO terms between the predicted partners of an interaction and (iv) using clusters of orthologous genes for the prediction.

Additionally, template interactions can be restricted to a subset of proteins to improve the reliability of the predictions. For example, the user can select interactions

based on the experimental methods by which the template interactions were observed, the number of experiments confirming the template interaction or the number of species in which the interaction between homolog pairs of the template was observed.

Finally, the user can restrict the list of predictions, reducing the number of predictions to a manageable size. The user can select specific partners: (i) those with specific keywords in their descriptive attributes, (ii) those associated with a certain pathology, (iii) those belonging to a specific taxon, including the case in which *query* proteins are from a particular pathogen, and the predicted partners are from selected hosts, (iv) those belonging to a subset of proteins uploaded by the user and (v) those with transmembrane predicted regions [calculated with TMHMM (37)].

BENCHMARK

We have checked the validity of our predictions by two approaches. In the first approach, specific known interactomes reported in BIANA were predicted using the leave-one-out strategy. For each interaction being tested, all interactions reported in BIANA were used as templates including the interactions of the same organism, but excluding the interaction being tested. Several organisms covering different taxonomy groups were considered (see Table 1). Sensitivity was calculated by testing the percentage of known interactions correctly predicted over the total of known interactions. Between 1 and 59% of known interactions were predicted by using a sequence identity cut-off ranging from 30 to 90%. In a complementary approach, we used the negatome database (38) as a source for non-interactions to calculate specificity. Specificity was calculated as the percentage of correctly predicted negative interactions (true negatives) out of 1291 non-interacting pairs. The specificity ranged between 72 and 98%. However, the study of specificity is not enough to quantify false positive predictions in the PPI context because the number of non-interacting protein pairs is larger than the number of interacting pairs. Indeed, a high specificity does not necessarily mean that a large proportion of the predicted interactions is correct [i.e. positive predictive value (PPV)]. We selected the *Arabidopsis* interactome (39) to estimate an example of the PPV of our predictions (see Table 1). This is one of the newest data sets of PPIs available, and we have to note that its authors estimated that their experiment had already a precision of 80% and a sensitivity of 16%. A more detailed comparison with a previous interolog approach is shown in Supplementary Table S1.

DISCUSSION

We have presented a PPI prediction server, BIPS, which is based on the similarity between protein sequences and PPIs reported in several biological databases integrated using the BIANA framework. BIPS benefits from the large amount of information deposited in these databases. By increasing the number of template interactions, coverage of the predictions is greatly improved. Traditionally, the interolog approach has been defined as the transference of interactions between orthologs from different species. However, the distinction between orthologs (gene pairs that trace back to speciation) and paralogs (gene pairs resulting from duplication events) is not always clear, as it depends on the last common ancestor applied. BIPS makes no distinction between orthologs and paralogs. In contrast, BIPS relies directly on pair-to-pair similarities to perform its inferences. Comparison of predictions based only on groups of orthologous genes show comparable results when applying a pair-to-pair sequence similarity cut-off between 30 and 70% identity (see Table 1). The main advantages of BIPS over other servers are the capability to predict interactions on a large scale such as for whole proteomes, in a reasonable time, the use of up-to-date database information and the option for the user to select several filters to improve confidence in the results. The latter is based on the notion that using additional information about the protein targets and template interactions increases the reliability of the predictions. The most trustable predictions are observed when the sequence similarity measure is restrictive and some filters

Table 1. Sensitivity (a), specificity (b) and PPV (c) of the prediction of different data sets by varying the filtering conditions

Data set	90% identity				70% identity				30% identity				eggNOG
	All	Dom	GO	COG	All	Dom	GO	COG	All	Dom	GO	COG	All
Arabidopsis (a)	17%	3%	8%	7%	31%	5%	16%	13%	53%	8%	25%	21%	32%
Yeast (a)	13%	0%	4%	6%	17%	0%	5%	6%	30%	1%	10%	14%	33%
Human (a)	11%	1%	5%	4%	23%	2%	10%	8%	42%	4%	17%	15%	28%
Mouse (a)	17%	1%	6%	8%	36%	2%	12%	14%	59%	3%	18%	20%	27%
Drosophila (a)	4%	0%	1%	0%	5%	0%	1%	1%	10%	1%	2%	2%	8%
Worm (a)	1%	0%	0%	0%	4%	1%	2%	1%	12%	2%	6%	4%	11%
Negatome (b)	92%	98%	95%	97%	88%	97%	91%	96%	72%	92%	78%	97%	77%
Arabidopsis (c)(d)	1.27%	4.92%	2.33%	1.67%	0.56%	4%	1.23%	0.63%	0.07%	0.56%	0.17%	0.06%	0.03%

The percentage indicates sensitivity (a), specificity (b) or PPV (c). (d) Main screen data set from the *Arabidopsis* interactome map for 8 596 *Arabidopsis* proteins, with a precision of 80% and a sensitivity of 16% (38). *All*, all predictions without applying any restriction; *Dom*, predictions filtered by known interacting domains reported in 3DID or iPfam; *GO*, predictions filtered by GO term similarity (biological process or cellular compartment); *COG*, predictions filtered by known interacting proteins in the same clusters of orthologous groups. Clusters of orthologous genes were as defined in the eggNOG database excluding non-supervised clusters.

are applied (see sensitivity and specificity values in Table 1). All these capabilities provide the user with a useful tool to select predictions and focus on a specific research area.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

FUNDING

Spanish Ministry of Science and Innovation (MICINN); German Federal Ministry of Education and Research (BMBF); partners of the ERASysBio+ initiative supported under the EU ERA-NET Plus scheme in FP7 (SHIPREC) Euroinvestigación [EUI2009-04018], as well as FEDER [BIO2011-22568, PSE-0100000-2009]. Funding for the open access charge: ERASysBio+ initiative supported under the EU ERA-NET Plus schme in FP7 (SHIPREC) Euroinvestigación [EUI2009-04018].

Conflict of interest statement. None declared.

REFERENCES

1. Aloy,P. and Russell,R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.*, **22**, 1317–1321.
2. Sambourg,L. and Thierry-Mieg,N. (2010) New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics*, **11**, 605.
3. Qi,Y., Klein-Seetharaman,J. and Bar-Joseph,Z. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput.*, 531–542.
4. Jain,S. and Bader,G.D. (2010) An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, **11**, 562.
5. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
6. Matthews,L.R., Vaglio,P., Reboul,J., Ge,H., Davis,B.P., Garrels,J., Vincent,S. and Vidal,M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.*, **11**, 2120–2126.
7. Wang,F., Liu,M., Song,B., Li,D., Pei,H., Guo,Y., Huang,J. and Zhang,D. (2012) Prediction and characterization of protein-protein interaction networks in swine. *Proteome Sci.*, **10**, 2.
8. Shin,C.J., Davis,M.J. and Ragan,M.A. (2009) Towards the mammalian interactome: inference of a core mammalian interaction set in mouse. *Proteomics*, **9**, 5256–5266.
9. Schleker,S., Garcia-Garcia,J., Klein-Seetharaman,J. and Oliva,B. (2012) Prediction and comparison of *Salmonella*-human and *Salmonella-Arabidopsis* interactomes. *Chem Biodivers*, **9**, 991–1018.
10. Krishnadev,O. and Srinivasan,N. (2011) Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. J. Biol Macromol.*, **48**, 613–619.
11. Wiles,A.M., Doderer,M., Ruan,J., Gu,T.T., Ravi,D., Blackman,B. and Bishop,A.J. (2010) Building and analyzing protein interactome networks by cross-species comparisons. *BMC Syst. Biol.*, **4**, 36.
12. Tarcea,V.G., Weymouth,T., Ade,A., Bookvich,A., Gao,J., Mahavisno,V., Wright,Z., Chapman,A., Jayapandian,M., Ozgur,A. *et al.* (2009) Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res.*, **37**, D642–D646.
13. Chen,C.C., Lin,C.Y., Lo,Y.S. and Yang,J.M. (2009) PPISearch: a web server for searching homologous protein-protein interactions across multiple species. *Nucleic Acids Res.*, **37**, W369–W375.
14. Gallone,G., Simpson,T.I., Armstrong,J.D. and Jarman,A.P. (2011) Bio::Homology::InterologWalk—a Perl module to build putative protein-protein interaction networks through interolog mapping. *BMC Bioinformatics*, **12**, 289.
15. Mathivanan,S., Periaswamy,B., Gandhi,T.K., Kandasamy,K., Suresh,S., Mohmood,R., Ramachandra,Y.L. and Pandey,A. (2006) An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, **7 (Suppl. 5)**, S19.
16. Aragues,R., Garcia-Garcia,J. and Oliva,B. (2008) Integration and prediction of PPI using multiple resources from public databases. *J. Proteomics. Bioinform.*, **1**, 166–187.
17. Garcia-Garcia,J., Guney,E., Aragues,R., Planas-Iglesias,J. and Oliva,B. (2010) Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, **11**, 56.
18. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
20. Stein,A., Ceol,A. and Aloy,P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
21. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
22. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
23. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
24. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
25. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2011) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
26. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
27. Winnenburg,R., Baldwin,T.K., Urban,M., Rawlings,C., Kohler,J. and Hammond-Kosack,K.E. (2006) PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res.*, **34**, D459–D464.
28. Driscoll,T., Dyer,M.D., Murali,T.M. and Sobral,B.W. (2009) PIG—the pathogen interaction gateway. *Nucleic Acids Res.*, **37**, D647–D650.
29. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
30. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
31. Chatr-aryamontri,A., Ceol,A., Peluso,D., Nardozza,A., Panni,S., Sacco,F., Tinti,M., Smolyar,A., Castagnoli,L., Vidal,M. *et al.* (2009) VirusMINT: a viral protein interaction database. *Nucleic Acids Res.*, **37**, D669–D673.
32. UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
33. 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. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

34. Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S. and Chen,C.F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.
35. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
36. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
37. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
38. Smialowski,P., Pagel,P., Wong,P., Brauner,B., Dunger,I., Fobo,G., Frishman,G., Montrone,C., Rattei,T., Frishman,D. *et al.* (2010) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res.*, **38**, D540–D544.
39. *Arabidopsis* Interactome Mapping Consortium. (2011) Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, **333**, 601–607.