Struct2Net: a web service to predict protein–protein interactions using a structure-based approach

Citation: Singh, R. et al. "Struct2Net: a Web Service to Predict Protein-protein Interactions Using a Structure-based Approach." Nucleic Acids Research 38.Web Server (2010): W508–W515. Web. 25 May 2012.

As Published: http://dx.doi.org/10.1093/nar/gkq481

Publisher: Oxford University Press (OUP)

Persistent URL: http://hdl.handle.net/1721.1/70951

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

W508–W515 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 31 May 2010*
doi:10.1093/nar/gkq481

Struct2Net: a web service to predict protein–protein interactions using a structure-based approach

Rohit Singh[1], Daniel Park[1,2], Jinbo Xu[3], Raghavendra Hosur[1] and Bonnie Berger[1,4,*]

[1]Computer Science and Artificial Intelligence Laboratory, [2]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA, [3]Toyota Technological Institute at Chicago, Chicago, IL and [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA

Received February 28, 2010; Revised May 2, 2010; Accepted May 13, 2010

ABSTRACT

Struct2Net is a web server for predicting interactions between arbitrary protein pairs using a structure-based approach. Prediction of protein–protein interactions (PPIs) is a central area of interest and successful prediction would provide leads for experiments and drug design; however, the experimental coverage of the PPI interactome remains inadequate. We believe that Struct2Net is the first community-wide resource to provide structure-based PPI predictions that go beyond homology modeling. Also, most web-resources for predicting PPIs currently rely on functional genomic data (e.g. GO annotation, gene expression, cellular localization, etc.). Our structure-based approach is independent of such methods and only requires the sequence information of the proteins being queried. The web service allows multiple querying options, aimed at maximizing flexibility. For the most commonly studied organisms (fly, human and yeast), predictions have been pre-computed and can be retrieved almost instantaneously. For proteins from other species, users have the option of getting a quick-but-approximate result (using orthology over pre-computed results) or having a full-blown computation performed. The web service is freely available at http://struct2net.csail.mit.edu.

INTRODUCTION

Systems biology research is like solving a jigsaw puzzle: the goal is to figure out how the various parts (i.e. genes and proteins within the cell) interact and work together. The *interactome* of an organism is then analogous to the puzzle's key: it describes the network of all the protein–protein interactions (PPIs) in a cell. As such, identifying all the protein-protein interactions for an organism is of great value, akin to sequencing its genome. Despite the use of high-throughput techniques in discovering PPIs, however, the coverage of experimentally determined PPI data remains poor (Table 1). Such low coverage is partly because the set of possible PPIs to be verified is so large (100 million for a species with 10 000 genes) that any exhaustive experimental verification will take a long time, even with high-throughput techniques. Indeed, the rate of PPI discovery has slowed down in recent years (Figure 1). Furthermore, the experimental approaches have limitations of their own. For example, tandem affinity purification experiments have historically had difficulty identifying transient interactions, while yeast two-hybrid experiments may produce false positives due to promiscuous proteins (1); recently, statistical methods have been proposed to improve confidence in the output of these experiments (2,3)

The paucity of interactome coverage has motivated significant research interest in methods for supplementing experimentally determined PPI data with interactions inferred or predicted from other sources. A wide variety of methods have been proposed. One approach is to use *interologs*, which are basically PPIs mapped from another species to the target species (4,5). The key problem there is to correctly map homologs across species (6,7). Another approach is to use functional genomic data and leverage the observation that a pair of interacting proteins is also likely to have similar GO annotations, occupy the same cellular sub-compartments, or correspond to genes with similar expression profiles (8,9). Consequently, many researchers have described machine learning-based approaches to predict PPI data from functional genomic data such as gene expression, cellular localization and GO annotation.

Predictions from many of these approaches have been aggregated into a number of databases/web services offering predicted PPIs. The STRING database (10)

*To whom correspondence should be addressed. Email: bab@mit.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Table 1. Availability of experimental PPI data for major eukaryotic organisms

Organism	Number of interactions	Percentage of proteins with at least one interaction
Mouse	1486	6.0
Human	26 640	41.8
Worm	4559	14.5
Fly	22 740	52.7
Yeast	48 901	93.5

The statistics here and in Figure 1 have been computed using data from BioGRID (v 2.0.53) (35). Data based on phenotypic suppression/enhancement and synthetic interaction was excluded as these experiments do not provide evidence of a direct physical interaction between proteins.

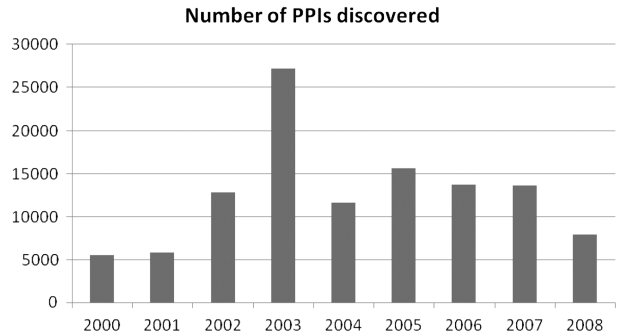

Figure 1. Rate of discovery of new eukaryotic PPI data has slowed.

combines experimental datasets (e.g. KEGG, BioGRID, HPRD) with computational predictions based on co-expression, interologs and text-mining, etc. The entries in this database correspond to functional interactions, and may not always be directly interpretable as PPIs. Another database, IntAct (11), focuses more on inferring interactions from expert curation of data from literature. Other public services include DOMINO (12), InterDom (13) and I2D (14). However, all of these databases suffer from a common selection bias: often, the proteins that have been selected for PPI experiments are usually genes/proteins that have received some attention before and, as such, are also more likely to have functional genomic data.

In this article, we describe Struct2Net, a web service for predicting PPIs using a structure-based approach. Our method predicts interactions by threading each pair of protein sequences onto potential structures in the Protein Data Bank (PDB) (15). Struct2Net provides PPI predictions that are independent of all the non-structure-based approaches and may thus be combined with any of them. Another key advantage of our web server is that, apart from the PDB data, the prediction algorithm only requires protein sequence data as input. It can thus be applied to proteins for which no functional data is available provided there is a suitable PDB structural template available.

The use of structure-based approaches to predict interaction has been previously proposed. Aloy and Russell (16) suggested the use of structure-based approaches to predicting PPIs. Lu *et al.* (17) constructed statistical potential functions to evaluate potential PPIs and later described MultiProspector, a structure-based prediction algorithm (18). In a previous paper, we proposed a prediction algorithm (also used by Struct2Net). Our algorithm builds upon previous work like MultiProspector, by combining a threading approach for template alignment with a novel machine learning approach to estimate a confidence score for the interaction. In our previous proof-of-concept paper, we discussed how Struct2Net's results compare favorably to related work (19).

Unfortunately, the progress made in prediction has not yet translated into comprehensive community resources. Aloy and Russell (20) have described InterPreTS, a web-server to predict PPIs for a given protein, using a homology modeling approach. We have already mentioned Lu *et al.*'s MultiProspector tool which also predicts PPIs (17). More recently, Fukuhara and Kawabata have described HOMCOS (21,22) a web-server that performs a similar task by homology modeling. MODBase is a database of homology models for protein complexes that have high sequence similarity to known structures (23). ADAN is a specialized database for prediction of PPIs mediated by linear motifs and utilizes position-specific matrices to assess putative interactions (24).

We believe that Struct2Net offers a significant advantage over such homology modeling approaches. Successful use of homology modeling requires relatively high sequence similarity between the query and template protein pairs. In contrast, we use a threading-based approach which widens the range of proteins for which predictions can be made. The use of threading also offers us improved performance: Fukuhara *et al.* (22) have reported that HOMCOS achieves a recall of 80% with a precision of about 10%; in comparison, Struct2Net achieves a recall of 80% with a precision of 30% [here, recall = (true positives)/(true positives + false negatives) and precision = (true positives)/(true positives + false positives)].

The Struct2Net approach can also be contrasted with methods that model PPIs based on domain-domain interactions. These approaches argue that the structural basis of protein interaction can be traced to the presence of interacting domains. A domain can be represented simply by its sequence motif or as a structure-fragment. Given a set of known PPIs, one can infer the set of domain pairings that are presumably the underlying cause of interaction. In principle, these pairs can then be used to make prediction for unannotated protein pairs. There has been a significant amount of work on analyzing PPIs using such domain interactions. Some researchers focus solely on the sequence signature of the domains, proposing methods to predict PPIs using these sequence domains (25,26). In previous work, we have discussed how such sequence-domain-based prediction can be combined with our approach in a machine-learning framework (19). We also described some results that suggest that Struct2Net's predictive ability compares well with the sequence-domain approaches.

Other researchers have aimed to understand these domains from a structural perspective. Prieto and Las Rivas (27) have reviewed publicly available databases that facilitate analysis of domain-based PPIs: 3did (28), SNAPPI-DB (29), iPfam (30), PIBASE (31) and PSIBase (32). While our approach has some parallels with these approaches, our goal is significantly different. The domain interaction databases are essentially repositories of known structural data, analyzed specifically from a PPI perspective. Prediction, which is our core goal, is usually out of the scope of these approaches. In the 'Methods Overview' section below, we suggest how Struct2Net could take advantage of some of these databases.

METHODS OVERVIEW

The guiding intuition behind our prediction approach is that if a potential interaction is sufficiently favorable from a thermodynamics perspective, it is likely to be true. We provide a brief description of the algorithm here. For more details, see Singh *et al.* (19), which describes a proof-of-concept implementation of the algorithm.

Our approach proceeds in two broad stages. Given a pair of protein sequences, the first stage predicts the most likely structure of the complex formed by the two proteins and produces a vector of scores that quantitatively represent the thermodynamic suitability of this structure. For this task, we start by analyzing the PDB to construct a database of complex-structure templates; then we thread the two sequences jointly through the various templates in this database and identify the best fitting template. Our threading algorithm formulates the threading problem as an integer linear program (ILP) and uses branch-and-bound techniques to efficiently find the solution. The ideas in this algorithm, when applied to a single-protein threading context in the RAPTOR program, have performed well at various blind tests and competitions (33,34). To speed up prediction, we ran PSI-BLAST (35) before running our threading algorithm. If some templates in our database appear in the list of PSI-BLAST top hits (E-value $<10^{-4}$), we simply thread the sequence pair to these templates instead of the whole template database. This speedup procedure does not lose accuracy since PSI-BLAST is very good at close homolog detection.

We now briefly describe how the database of complex templates was constructed. We begin by using a simple geometric criterion to determine if two protein chains form a complex. This provides an unbiased and objective way of characterizing an interaction. Given two protein chains in the same PDB entry, we first calculate the distance between two (non-hydrogen) atoms from these two chains. We assume that there is an interaction between two residues of different chains if there is at least one pair of atoms from these two residues with distance <3.5 Å. If there are at least 10 interacting residue pairs between two chains A and B, we say these two chains form a complex. To avoid redundancy, we enforce the constraint that any two templates in the database share $<70\%$ sequence identity. Following this procedure, our database currently contains 10 111 dimers. While our template database (and the web server's predictions) are currently built at the chain level, we intend to explore the incorporation of domain–domain interactions (from databases like SNAPPI, 3did, PSIBase, PIBASE, etc.) into it. This may help enlarge the database's coverage.

The second stage of our prediction approach evaluates the likelihood of the interaction based on the predicted structure. We compute various energy scores that evaluate the structure (e.g. the quality of the interfacial region, the quality of fit for the individual proteins). Given these, we use logistic regression to predict whether an interaction will occur. Let y_i be an indicator variable representing protein interaction, i.e. $y_i = 1$ if the protein pair i interacts and 0 otherwise. Let $\boldsymbol{x_i} = \{x_i^1, x_i^2, \ldots\}$ be the vector of scores we use for prediction. We fit the following model:

$$\log \frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)} = \alpha + \beta_1 x_i^1 + \beta_1 x_i^2 + \cdots$$

where α, β_1, β_2, etc. are parameters to be learned from data. To train this model, we constructed positive and negative training sets. Obviously, the choice of these sets can have a substantial impact on the prediction algorithm's quality.

We have developed criteria for constructing these datasets. The exact criteria and a discussion about the rationale behind them are available at the Struct2Net website. Briefly, we require that the positive examples either come from a small set of trustworthy protocols or from low-throughput experiments; or roughly correspond to co-clustered protein pairs in the PPI network. We chose BioGRID (36) as our data-source, but other multi-species genome-wide databases [e.g. MINT (37) or APID (38)] could also be used. For negative examples, we require that the two proteins either be disconnected in the PPI network or be at least 3 hops away from each other. Using these criteria, we had a training set of 62 519 pairs and a test set of 15 635 pairs (with a positive:negative ratio of 1:6 approximately, in both sets). We believe that these datasets provide good evidence of validation. Our construction of the negative dataset was motivated by similar approaches in literature (8). For positive datasets, we believe that our approach identifies true PPIs with better confidence than an alternative approach that would select repeatedly observed PPIs (across multiple experiments). Our scheme emphasizes protocols and studies with low error-rates. In contrast, many high-throughput protocols (e.g. yeast two-hybrid) have systematic biases which may manifest as repeated false positives, even across multiple experiments.

In addition to the energy scores from the first stage, we aimed to enhance the model's predictive power by adding extra terms to it. These included interaction terms, non-linear functions of the energy scores, as well as normalized scores (e.g. interfacial energy normalized by the average of the two proteins' sequence length). We then used the Akaike information criterion (AIC) to select the model with the best trade-off of higher

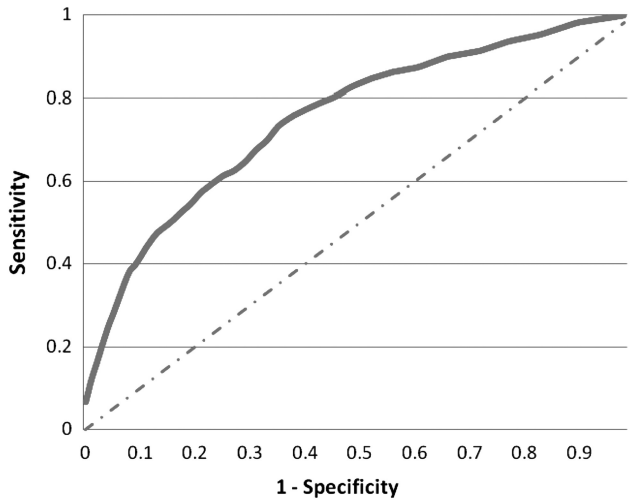

Figure 2. Sensitivity versus specificity. The prediction algorithm can achieve 60% sensitivity while maintaining 75% specificity as measured on the test set. Here, sensitivity = (true positives)/(true positives + false negatives) and specificity = (true negatives)/(true negatives + false positives). We constructed a training set and test set of positive and negative examples from yeast and fly, using criteria we have developed to identify high-confidence positive and negative examples of PPIs (see the website FAQ for details). After training the logistic regression model on the training set, its performance was measured on the test set.

explanatory power and lower complexity. Using this model, we computed the interaction score for the given joint structure.

As seen by the graph in Figure 2, our method has significant predictive power when tested on current data. For further details, including the construction of training/test datasets and evaluation of the algorithm, please see 'About' on the Struct2Net website. As the threshold for the interaction score is increased, the specificity of the model rises. Higher sensitivity, on the other hand, can be achieved by choosing lower specificity. Also, we note here that we do not make a prediction for a candidate protein pair if the first stage of our algorithm fails to predict a structure for them.

WEB SERVER

The Struct2Net server provides multiple querying options. For the most commonly studied organisms (*Saccharomyces cerevisiae, Drosophila melanogaster, Homo sapiens*), PPI predictions have been pre-computed and can be retrieved by gene name or a wide array of gene identifiers, including 'ids' from Ensembl, EMBL, Entrez, UniProtKB, GenBank, FlyBase and Saccharomyces Genome Database (SGD; Figure 3A). For proteins from other organisms, the users can query by sequence in FASTA format (Figure 3B). Users have the option of getting a quick-but-approximate result, by retrieving predictions from the best-hit ortholog over pre-computed results, or have a full-blown computation performed (Figure 3C). Furthermore, with full-blown computations, a batch query option is available for querying multiple sequences at a time. In addition, with orthology-based approximation, users can specify just one protein identifier or FASTA sequence; in that case, all the interactions involving that protein will be returned.

Predictions are retrieved almost instantaneously when querying by ids. When querying by protein sequence and with orthology-based approximation selected, typical run-times are within 20 s. Full-blown computations finish within 45 mins, given query and subject sequences. Because of the potential for long run-times (e.g. if the server is overloaded), we encourage the user to supply an email address to which a job id and a link to the progress page are sent upon submission. Alternatively, users can check the progress of a submitted job by entering a job id in the 'Fetch Job' webpage. Upon completion of a job, an email with a link to the results page will also be sent.

For pre-computed predictions in *S. cerevisiae*, *D. melanogaster* and *H. sapiens*, the output for each query protein sequence consists of a list of all predicted interactions along with their confidence scores (Figure 3D). Struct2Net interactively links each gene hit to various sequence databases along with associated GO annotations and aliases. Results are also cross-referenced with BioGrid in the case where experimental data is available for a predicted interaction. For predictions in other organisms using the Struct2Net algorithm, the output for each sequence pair contains details on the best-fit complex templates used during the computation including sequence alignments, alignment scores, their associated *z*-scores and an interfacial energy calculated between the sequence pair (Figure 3D). In addition, an overall confidence score is provided for each potential interaction. The confidence score ranges from 0 to 1, with 0 indicating minimum confidence and 1 indicating maximum confidence. In the 'About' page of the website, we discuss threshold choices that would allow a user to achieve a desired level of specificity in the output or a desired number of interactions above the threshold. For batch queries, results are separated by each pair of protein sequences.

For users interested in performing large-scale database analysis and classification, bulk download of predictions for *S. cerevisiae*, *D. melanogaster* and *H. sapiens* is also available. We have further made available a script on the Download page that facilitates the integration of Struct2Net's predictions with other tools. In the future, we plan to update our template database every 3 months. Every 6 months, we will update our pre-computed predictions using the latest template database.

In Table 2, we provide an example of our algorithm's results on a set of protein pairs often used as test cases. For comparison, we have also displayed the results of HOMCOS and InterPretS for these pairs. Multi-Prospector no longer seems to be publicly available, and we could not include its results. The test cases we have chosen are the same as chosen by Fukuhara *et al.* for evaluating HOMCOS (22). As can be seen, for pairs that are thought to be interacting (Table 2), the final scores from Struct2Net are, on average, significantly higher than for non-interacting pairs (Table 2). Furthermore, normalizing the difference between the average interacting and non-interacting scores for each method by the

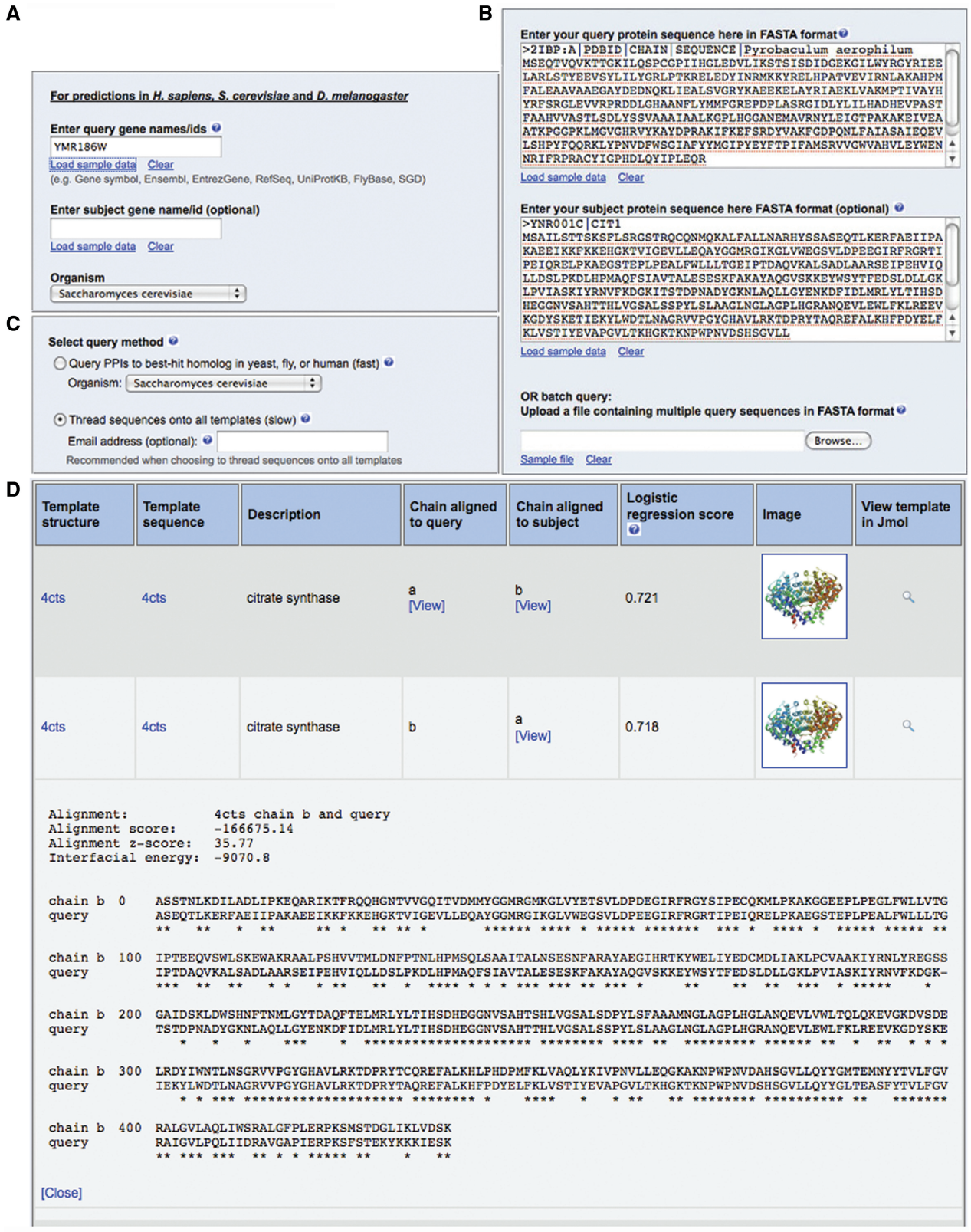

Figure 3. Web interface and output of Struct2Net. (**A** and **B**) Web server entry page. (**C**) A query option for either a quick-but-approximate approach (using orthology over pre-computed predictions from yeast, fly and human) or a full-blown computation using the Struct2Net algorithm. (**D**) Example of an output page when choosing to thread pairs of sequences onto all templates. Confidence scores for a potential interaction are displayed along with associated template–sequence alignments and threading details.

Table 2. Struct2Net results for set of interacting and non-interacting protein pairs

	Struct2Net				HOMCOS		InterPretS
Job ID	Test pairs	Uniprot IDs	Templates	Confidence	Zseqcon	Zcon	Best *Z*-score
Interacting protein pairs							
1OLYN2PM	1b34AB	P62314, P62316	1d3bAB	0.620	−40.9	−3.37	1.62
N9LJTIPG	1g65JK	P22141, P30656	1iruKL	0.590	−62.7	−1.34	No Hits
Q44OTFMD	1gl2BC	O70439, O88384	1gl2BC	0.958	−37.9	−4.38	3.42
HZ0N1HR9	1sxjBC	P40339, P38629	1sxjBC	0.251	−81.3	−3.77	2.87
NQARC82J	1finAB	P24941, P20248	1e9hAB	0.428	−70.7	−2.96	3.04
4LJQHZA	1ukvGY	P39958, P01123	1ukvGY	0.662	−67.3	−6.23	3.90
4LFMIDJ	1bi7AB	Q00534, P42771	1bi7AB	0.385	−51.1	−2.37	0.84
9N2PHLBI	1id3AF	P61830, P02309	1aoiAB	0.989	−45.6	−5.39	4.59
SNTT8NHN	1s1hJN	P38701, P41058	1s1gJN	0.990	−23.7	−0.27	1.28
NBTGSU4P	1ow3AB	Q07960, P61586	1ow3AB	0.425	−62.3	−3.98	2.58
Average				**0.63**	−54.35	−3.57	2.87
Standard Deviation				0.25	14.9	1.72	1.14
Non-interacting protein pairs							
JTP3Q280	1g3nAB	Q00534, P42773	1g3nAB	0.347	−57.1	−2.88	1.61
JCEFCQGQ	1oiuBC	P24941, P20248	1e9hBA	0.428	−70.5	−2.87	3.04
YD4L76VD	1gotAB	P04695, P62871	1gg2AB	0.249	−83.5	−3.39	2.98
YRJQ0JZI	1ow3AB	Q07960, P61586	1ow3AB	0.425	−62.3	−3.98	2.62
JQ260ZEC	1f3mAC	Q13153, Q13153	1f3mAC	0.718	−43.5	−7.05	3.49
VJ8BPGQ2	1a9nAB	P09661, P08579	1a9nAB	0.334	−45.1	−2.82	2.69
0OLMGNWZ	1k5dAC	P62826, P41391	none	0.169	−66.7	−4.15	2.29
8WEA7WWS	1fq1AB	Q16667, P24941	1fq1AB	0.425	−60.9	−1.90	1.83
EWVFV6TL	1fbvAC	P22681, P68036	1fbvAC	0.717	−53.2	−0.87	-0.31
WVW4S9TW	1qbkBC	Q92973, P62826	none	0.180	−61.4	−1.35	2.48
Average				**0.39**	−60.42	−3.13	2.27
Standard Deviation				0.25	14.9	1.72	1.14

We chose sets of interacting and non-interacting protein pairs; these pairs are taken from an analysis by HOMCOS authors (22). Low confidence templates are indicated as 'none'. For comparison, the scores from HOMCOS and InterPretS for these pairs are also shown. Struct2Net provides a confidence score between 0 and 1 (0 indicates minimum confidence while 1 indicates maximum confidence). HOMCOS provides a Zcon measure, while InterPretS provides *Z*-scores. The average positive and negative scores are separated by a larger magnitude in Struct2Net: the separation is about 0.96 SD in Struct2Net; the corresponding separation in HOMCOS is 0.26 SD, and in InterPretS is 0.53 SD. Clearly, the Struct2Net score better distinguishes between interacting and non-interacting pairs.

standard deviation of the method's scores suggests that the discriminatory ability of Struct2Net compares favorably with HOMCOS and InterPretS.

Limitations

A problem common to all structure-based PPI prediction methods is *coverage*: the number of known protein structures is vastly smaller than the number of known protein sequences. As such, no structural template may be available for the protein pair being queried. In contrast to other web services that only use homology modeling, our use of protein threading affords not only greater accuracy but also greater coverage: in yeast and fly, it covers about 10% of the genome. This is because homology modeling matches query proteins based only on sequence alignments to sequences with known structures; in contrast, threading is able to capture alignments more in the 'twilight zone' by matching query sequences to structural templates (19). Furthermore, it has been shown that localized threading using interface profiles can further improve coverage and accuracy (39,40). While Struct2Net can be used for validation purposes (e.g. to double-check entries in BioGRID), its coverage limitations may at the present time make it better suited to be an exploratory tool, especially for unannotated proteins where only sequence information is available, or to be used in conjunction with low-confidence experimental data.

CONCLUSIONS

Although high-throughput biochemical approaches for discovering PPIs have proven very successful, the current experimental coverage of the interactome remains inadequate and would benefit from computational tools. The Struct2Net web server allows the user to easily query for high-probability structure-based interactions as a potentially high-quality, high-coverage data source for large-scale integrative approaches to interactome construction. The predicted interactions also include a numeric score, allowing users to further filter the data. To the best of our knowledge, this web server is the first of its kind and will be of considerable value to systems biologists interested in PPIs, partly because of the effort we have put into identifying high-confidence positive and negative examples of PPIs as inputs to machine learning algorithms and the extensive computational effort involved in making each prediction. A strength of this web service is its ongoing integration of up-to-date structural templates for improving its predictions. Struct2Net's predictions may be used on their own or as one of the

inputs into a computational framework that combines them with other sources (e.g. low-quality experimental data or predictions from functional genomic data). For example, Jensen *et al*. (10), Qi *et al*. (8) and Srinivasan *et al*. (9) have described some general approaches for combining various predictors of PPI data. Struct2Net's predicted interaction scores can easily be integrated into such models.

ACKNOWLEDGEMENTS

Some of computations in this work were performed using the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and the University of Chicago Computation Institute.

FUNDING

This publication was made possible by Grant Number 1R01GM081871; from the National Institute of General Medical Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NIGMS. TTI-C internal research funding (J.X.). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Bjorklund,A.K., Light,S., Hedin,L. and Elofsson,A. (2008) Quantitative assessment of the structural bias in protein-protein interaction assays. *Proteomics*, **8**, 4657–4667.
2. Simonis,N., Rual,J.F., Carvunis,A.R., Tasan,M., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Sahalie,J.M., Venkatesan,K., Gebreab,F. *et al*. (2009) Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nature Methods*, **6**, 47–54.
3. Venkatesan,K., Rual,J.F., Vazquez,A., Stelzl,U., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Zenkner,M., Xin,X., Goh,K.I. *et al*. (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.
4. 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.
5. Uetz,P., Dong,Y.A., Zeretzke,C., Atzler,C., Baiker,A., Berger,B., Rajagopala,S.V., Roupelieva,M., Rose,D., Fossum,E. *et al*. (2006) Herpesviral protein networks and their interaction with the human proteome. *Science*, **311**, 239–242.
6. Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.*, **24**, 427–433.
7. Singh,R., Xu,J. and Berger,B. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA*, **105**, 12763–12768.
8. Qi,Y., Bar-Joseph,Z. and Klein-Seetharaman,J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**, 490–500.
9. Srinivasan,B., Novak,A., Flannick,J., Batzoglou,S. and McAdams,H. (2006) Integrated protein interaction networks for 11 microbes. *LNCS*, **3909**, 1–14.
10. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al*. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
11. 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.
12. Ceol,A., Chatr-aryamontri,A., Santonico,E., Sacco,R., Castagnoli,L. and Cesareni,G. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res.*, **35**, D557–D560.
13. Ng,S.K., Zhang,Z., Tan,S.H. and Lin,K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, **31**, 251–254.
14. Brown,K.R. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.
15. 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.
16. Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
17. Lu,H., Lu,L. and Skolnick,J. (2003) Development of unified statistical potentials describing protein-protein interactions. *Biophys. J.*, **84**, 1895–1901.
18. Lu,L., Arakaki,A.K., Lu,H. and Skolnick,J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome. *Genome Res.*, **13**, 1146–1154.
19. Singh,R., Xu,J. and Berger,B. (2006) Struct2net: integrating structure into protein-protein interaction prediction. *Pac. Symp. Biocomput.*, **11**, 403–414.
20. Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.
21. Fukuhara,N. and Kawabata,T. (2008) HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res.*, **36**, W185–W189.
22. Fukuhara,N., Go,N. and Kawabata,T. (2006) Prediction of interacting proteins from homology-modeled complex structures using sequence and structure scores. *Biophysics*, **3**, 13.
23. Pieper,U., Eswar,N., Webb,B.M., Eramian,D., Kelly,L., Barkan,D.T., Carter,H., Mankoo,P., Karchin,R., Marti-Renom,M.A. *et al*. (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **37**, D347–D354.
24. Encinar,J.A., Fernandez-Ballester,G., Sanchez,I.E., Hurtado-Gomez,E., Stricher,F., Beltrao,P. and Serrano,L. (2009) ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*, **25**, 2418–2424.
25. Lee,H., Deng,M., Sun,F. and Chen,T. (2006) An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, **7**, 269.
26. Wang,H., Segal,E., Ben-Hur,A., Li,Q.R., Vidal,M. and Koller,D. (2007) InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol.*, **8**, R192.
27. Prieto,C., De., Las. and Rivas,J. (2006) Structural domain-domain interactions: assessment and comparison with protein-protein interaction data to improve the interactome. *Nucleic Acids Res.*, **34**, W298–W302.
28. Stein,A., Russell,R. and Aloy,P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.
29. Jefferson,E., Walsh,T., Roberts,T. and Barton,G. (2007) SNAPPI-DB: a database and API of structures, iNterfaces and alignments for protein-protein interactions. *Nucleic Acids Res.*, **35**, D580–D589.
30. Finn,R., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

31. Davis,F. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
32. Gong,S., Yoon,G., Jang,I., Bolser,D., Dafas,P., Schroeder,M., Choi,H., Cho,Y., Han,K., Lee,S. *et al*. (2005) PSIBASE: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*, **21**, 2541–2543.
33. Xu,J. and Li,M. (2003) Assessment of RAPTOR's linear programming approach in CAFASP3. *Proteins*, **53(Suppl. 6)**, 579–584.
34. Xu,J., Peng,J. and Zhao,F. (2009) Template-based and free modeling by RAPTOR++ in CASP8. *Proteins*, **77(Suppl. 9)**, 133–137.
35. 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.
36. Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al*. (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res*., **36**, D637–D640.
37. Chatr-aryamontri,A., Ceol,A., Palazzi,L., Nardelli,G., Schneider,M., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res*., **35**, D572–D574.
38. Prieto,C. and De Las Rivas,J. (2006) APID: Agile Protein interaction data analyzer. *Nucleic Acids Res*., **34**, W298–W302.
39. Pulim,V., Berger,B. and Bienkowska,J. (2008) Optimal contact map alignment of protein-protein interfaces. *Bioinformatics*, **24**, 2324–2328.
40. Pulim,V., Bienkowska,J. and Berger,B. (2008) LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. *Prot. Sci*., **17**, 279–292.