Structural bioinformatics

Advance Access publication July 9, 2013

iLoops: a protein–protein interaction prediction server based on structural features

Joan Planas-Iglesias[†], Manuel A. Marin-Lopez[†], Jaume Bonet[†], Javier Garcia-Garcia and Baldo Oliva[*]

Structural Bioinformatics Laboratory, Universitat Pompeu Fabra, 08003 Barcelona, Spain

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Protein–protein interactions play a critical role in many biological processes. Despite that, the number of servers that provide an easy and comprehensive method to predict them is still limited. Here, we present iLoops, a web server that predicts whether a pair of proteins can interact using local structural features. The inputs of the server are as follows: (i) the sequences of the query proteins and (ii) the pairs to be tested. Structural features are assigned to the query proteins by sequence similarity. Pairs of structural features (formed by loops or domains) are classified according to their likelihood to favor or disfavor a protein–protein interaction, depending on their observation in known interacting and non-interacting pairs. The server evaluates the putative interaction using a random forest classifier.

Availability: iLoops is available at http://sbi.imim.es/iLoops.php

Contact: baldo.oliva@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 8, 2013; revised on June 21, 2013; accepted on July 5, 2013

1 INTRODUCTION

Protein–protein interactions (PPIs) mediate almost all processes in living cells. Thus, the discovery of new PPIs is the key to understanding the complexity of biological systems. Several experimental methods have been developed to identify new PPIs; among them, yeast two-hybrid and tandem affinity purification are the most used high-throughput methods (Yu *et al.*, 2008). However, these methods are still economically and time-costly, and are hindered by the high amount of false-negative interactions (Braun *et al.*, 2009).

Mirroring the experimental techniques, computational methods have also been developed to identify new PPIs. These computational methods can be divided into three main approaches, depending on the contextual properties they exploit: structural, genomic or biological (Skrabanek *et al.*, 2008). Structural context methods, such as InterPreTS (Aloy and Russell, 2003), PIPE (Pitre *et al.*, 2006) or Struct2Net (Singh *et al.*, 2010), extrapolate structural information of a protein directly from its sequence and predict or score PPIs based on the molecular composition and structural conformation of the partners. Otherwise, genomic context methods like STRING (Szklarczyk *et al.*, 2011) or Predictome (Mellor *et al.*, 2002) provide predictions of PPIs based on gene fusion, gene co-localization and phylogenetic profiles. Finally, biological context methods [e.g. GeneCensus (Jansen *et al.*, 2003)] use Bayesian networks to produce reliable predictions.

To properly assess the success of any PPI prediction method, a reference set of non-interacting protein pairs (NIPs) is required (Ben-Hur and Noble, 2006; Trabuco *et al.*, 2012). The Negatome database (Smialowski *et al.*, 2010) is a set of protein pairs that are unlikely to engage in physical direct interactions compiled through manually curated literature and crystallographic data.

Here, we present the iLoops web server, a web implementation of our recently published structural context method (Planas-Iglesias *et al.*, 2013) that exploits ArchDB classification of loops (Espadaler *et al.*, 2004) to predict PPIs. Our method explores the balance of structural features (SFs) observed in PPIs and/or NIPs. The server provides PPI predictions with an associated precision using a random forest (RF) classifier (Hall *et al.*, 2009) and considering different ratios between PPIs and NIPs.

2 METHODS

2.1 Signatures

We use the classification of loops from ArchDB (Espadaler *et al.*, 2004) and domains from SCOP (Andreeva *et al.*, 2008) to define the local SFs; hence, SFs are domains or loops. Protein signatures are defined as groups of 1–3 SFs (either loops or domains). Interaction signatures are defined as pairs of protein signatures of the same type between two proteins, interacting or not interacting.

2.2 Alignment of protein signatures

Structural features are annotated on the sequences using BLAST (Altschul *et al.*, 1997). SFs are assigned to a query protein if the sequence alignment is above the twilight zone (Rost, 1999) and the coverage of the SF is high enough (100% for loops, 75% for domains) (Fig. 1a). The prediction is limited to pairs of proteins with some SF.

2.3 Evaluation of interaction signatures

Interaction signatures are scored as favoring or disfavoring the interaction (Fig. 1b) as in Planas-Iglesias *et al.* (2013), using the Negatome database (Smialowski *et al.*, 2010) for NIPs and the integration of several sources of PPIs with BIANA (Garcia-Garcia *et al.*, 2010) for PPIs. In this server, we use larger sets of PPIs and NIPs as training and testing sets, allowing

[*]To whom correspondence should be addressed.

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

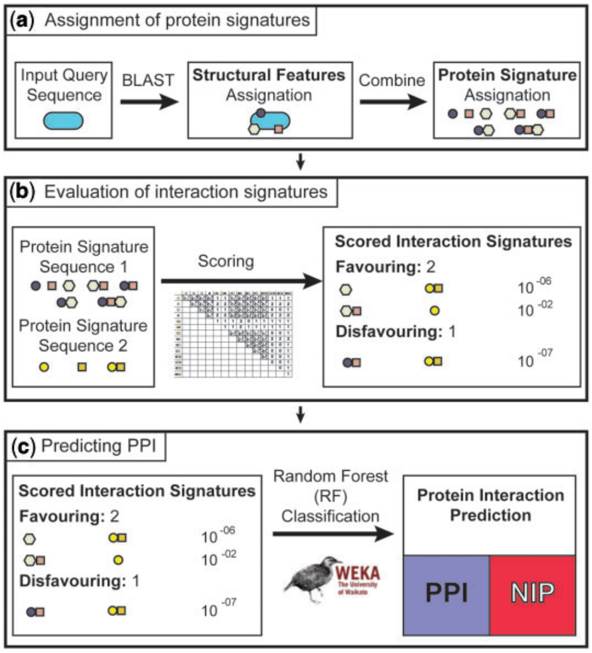
(a) Assignment of protein signatures
Input Query Sequence
BLAST
Structural Features Assignation
Combine
Protein Signature Assignation
(b) Evaluation of interaction signatures
Protein Signature Sequence 1
Protein Signature Sequence 2
Scoring
Scored Interaction Signatures
Favouring: 2
10^{-06}
10^{-02}
Disfavouring: 1
10^{-07}
(c) Predicting PPI
Scored Interaction Signatures
Favouring: 2
10^{-06}
10^{-02}
Disfavouring: 1
10^{-07}
Random Forest (RF) Classification
WEKA
The University of Waikato
Protein Interaction Prediction
PPI
NIP

Fig. 1. Schema of the iLoops server prediction pipeline

for different ratios of PPIs and NIPs (Planas-Iglesias *et al.*, 2013). The prediction is limited to pairs with scored interaction signatures.

2.4 Predicting a PPI

We use RF models for the prediction. RF models were generated with the WEKA package (Hall *et al.*, 2009) by learning from the signatures of the training sets. Several RF models were obtained using different relative costs (RCs) to penalize the ratio of false-positive predictions. Associated precisions were computed for different unbalance ratios (URs) of PPIs and NIPs (Planas-Iglesias *et al.*, 2013). The actual server uses these RF models to provide the prediction of the input protein pairs (Fig. 1c). Each prediction has an inferred precision derived from our tests with different URs.

3 RESULTS: SERVER USAGE

The input for the iLoops web server is a set of FASTA sequences and the list of pairs of proteins to test. Data are provided through a text area, and the user selects the type of SFs to use for the prediction. Each submission is limited to 25 protein pairs. The server provides a job identification code that can be used to retrieve the predictions. Predictions are browsed through the web interface or can be downloaded in a compressed xml file. They are provided as a Boolean decision (YES/NO) for each query pair of the input list, plus the score given by the RF model. An inferred precision is also provided depending on the expected UR between PPIs and NIPs. This UR is selected by the user and depends on the experiment conditions (e.g. for any pair of co-localized human proteins the UR is ~1/50). The server selects by default the best RC for a given UR. Advanced options allow the user to select different RCs for the RF classifier. Details for each prediction display the SFs assigned to each query protein and a list of favoring (positive) and disfavoring (negative) interaction signatures sorted by their *P*-value. Help and FAQ sections in the server provide detailed information for setting parameters and examples of use.

4 DISCUSSION AND CONCLUSION

This work presents a server that predicts whether two proteins can interact based on the identification of SFs. The objective is to use loop or domain patterns learnt from PPIs and NIPs to predict the binding between protein pairs. The server provides a user-friendly interface and a comprehensive results page. Predictions can be traced back to the original databases to understand how they were obtained. Such traceability allows the user to comprehend the results and devise new experiments for a particular interaction (e.g. identifying interaction signatures that are relevant for the RF decision). The iLoops server offers the possibility to select the expected ratio between PPIs and NIPs in the predictions. Thus, iLoops allows the user to address different questions, such as predicting the largest amount of real interactions or minimizing the number of false predictions, providing an inferred precision of the prediction. For instance, predictions within a balanced set (1 PPI for each 1 NIP) can achieve 89% of precision and 81% of recall; conversely, predictions made on a set containing 1 PPI for each 50 NIPs (see earlier in the text) may expect 38% of precision and 39% of recall with an RC of 20 (Planas-Iglesias *et al.*, 2013).

Funding: This work was supported by the Spanish Government (MICINN) [FEDER BIO2008-0205 and BIO2011-22568, EUI2009-04018] and the Generalitat de Catalunya [FI-DGR 2012].

Conflict of Interest: none declared.

REFERENCES

Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andreeva,A. *et al.* (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

Ben-Hur,A. and Noble,W.S. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7** (**Suppl. 1**), S2.

Braun,P. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, **6**, 91–97.

Espadaler,J. *et al.* (2004) ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res.*, **32**, D185–D188.

Garcia-Garcia,J. *et al.* (2010) Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, **11**, 56.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *SIGKDD Explor.*, **11**, 10–18.

Jansen,R. *et al.* (2003) A bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.

Mellor,J.C. *et al.* (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, **30**, 306–309.

Pitre,S. *et al.* (2006) PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**, 365.

Planas-Iglesias,J. *et al.* (2013) Understanding protein-protein interactions using local structural features. *J. Mol. Biol.*, **425**, 1210–1224.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Singh,R. *et al*. (2010) Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Res*., **38**, W508–W515.

Skrabanek,L. *et al*. (2008) Computational prediction of protein-protein interactions. *Mol. Biotechnol*., **38**, 1–17.

Smialowski,P. *et al*. (2010) The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*., **38**, D540–D544.

Szklarczyk,D. *et al*. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*., **39**, D561–D568.

Trabuco,L.G. *et al*. (2012) Negative protein–protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*, **58**, 343–348.

Yu,H. *et al*. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.