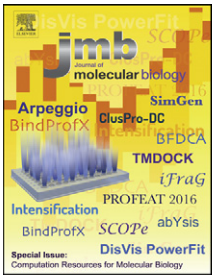
Article



iFrag: A Protein–Protein Interface Prediction Server Based on Sequence Fragments

Javier Garcia-Garcia [1], Victòria Valls-Comamala [2], Emre Guney [3,4], David Andreu [5], Francisco J. Muñoz [2], Narcis Fernandez-Fuentes [6] and Baldo Oliva [1]

*1 - **Structural Bioinformatics Laboratory,** Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain*
*2 - **Laboratory of Molecular Physiology,** Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain*
*3 - **Center for Complex Network Research and Department of Physics,** Northeastern University, Boston, 02115, MA, USA*
*4 - **Joint IRB-BSC-CRG Program in Computational Biology,** Institute for Research in Biomedicine (IRB Barcelona), C/Baldiri Reixac 10-12, 08028 Barcelona, Spain*
*5 - **Laboratory of Proteomics and Protein Chemistry,** Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Doctor Aiguader 88, 08003 Barcelona, Spain*
*6 - **Institute of Biological,** Environmental and Rural Sciences, Aberystwyth University, Gogerddan Campus, SY23 3EB Aberystwyth, UK*

***Correspondence to Narcis Fernandez-Fuentes and Baldo Oliva:**. baldo.oliva@upf.edu*
http://dx.doi.org/10.1016/j.jmb.2016.11.034
Edited by Michael Sternberg

Abstract

Protein–protein interactions (PPIs) are crucial in many biological processes. The first step towards the molecular characterisation of PPIs implies the charting of their interfaces, that is, the surfaces mediating the interaction. To this end, we present here iFrag, a sequence-based computational method that infers possible interacting regions between two proteins by searching minimal common sequence fragments of the interacting protein pairs. By utilising the sequences of two interacting proteins (queries), iFrag derives a two-dimensional matrix computing a score for each pair of residues that relates to the presence of similar regions in interolog protein pairs. The scoring matrix is represented as a heat map reflecting the potential interface regions in both query proteins. Unlike existing approaches, iFrag does not require three-dimensional structural information or multiple sequence alignments and can even predict small interaction sites consisting only of few residues. Thus, predicted interfaces range from short fragments composed of few residues to domains of proteins, depending on available information on PPIs, as we demonstrate in several examples. Moreover, as a proof of concept, we include the experimental validation on the successful prediction of a peptide competing with the aggregation of β-amyloid in Alzheimer's disease. iFrag is freely accessible at http://sbi.imim.es/iFrag.

Introduction

To understand the mechanisms that give rise to protein–protein interactions (PPIs) and their regulation, it is important to know the molecular details of an interaction. The first step towards this end is usually to define the regions that mediate the PPI, that is, the interface(s). Several experimental techniques help to determine the interacting regions of PPIs, such as conducting domain deletion experiments, site-directed mutagenesis to disrupt the interaction, domain-based yeast two-hybrid mapping [1], or footprinting [2], among others. The interface between two proteins is the result of the specific interaction between residue pairs playing a structural and functional role for the interaction. The properties exhibited by protein interfaces are unique and distinct from other non-interacting surfaces, and these are exploited to predict them. For example, evolutionary constraints lead to higher conservation of residues involved in interactions [3], to different amino acid composition propensities [4], or to a co-evolving behaviour of residues located in the interface [5]. Many of these features are present in specific sequence fragments of the proteins involved in a specific interface. Thus, the preservation of an interaction between homolog proteins, which has been customarily used to predict PPIs [6], implies the preservation of the local sequence fragments. 3D information has also been exploited, such as the use of atomic solvation parameters [7] or the extraction of structural interacting sequence patterns [8,9]. Correlations found between structural features (structural loops) and PPIs were similarly applied to predict interacting regions [10]. Other methods developed to predict PPI interfaces include DOMINE [11], focused on domain–domain interaction predictions by scoring a comprehensive collection of known and predicted domain–domain interactions; PIPE-Sites [12] and SLIDER [13], which infer potential interfaces by detecting recurring polypeptide sequences or overrepresented pairs of patterns in sequences of interacting proteins; PPIPP [14], based on a two-stage neural network trained with known interacting residue pairs from the Docking Benchmark 4.0 [15]; and VORFFIP [16] and M-VORFFIP [17], based on Random Forest classifiers integrating evolutionary, experimental, structure, and energy-based information. Also, several groups attempted to predict binding interfaces by means of sequence co-evolutionary approaches [18], although most servers and programs are still exclusively committed to protein structure prediction (e.g., PconsC [19]). For further information, see the recent review by Esmaielbeiki *et al.* [20].

Here, we present iFrag, a computational tool that searches minimal sequence similarity between pairs of interacting proteins to predict their binding region. iFrag makes no assumptions about protein domain composition, neither does it use protein structural information nor multiple sequence alignments. iFrag scores pairs of residues, one from each protein partner of a PPI, to unveil the smallest common sequence fragment of homolog pairs of interacting proteins. iFrag output consists on a heat map with the scores of residue pairs that potentially highlights the interacting regions between two proteins. iFrag also predicts putative interfaces based on sequence similarity with protein complexes with known 3D structure (i.e., templates) if they are available (named BlastPDB in iFrag server). iFrag-BlastPDB approach requires a minimum percentage of sequence identity on the alignment with the templates (30%). The output is complemented with sequence feature annotations from Uniprot database [21] and PFAM domains [22], helping the user to identify or discard regions of interest useful for designing validation experiments.

The method is evaluated and compared with other state-of-the-art applications using a non-redundant set of proteins for which the interface is known. Using a non-redundant set for evaluating PPI site prediction methods is essential since failure to do so results in overly optimistic applicability measures. The methods that predict protein–protein interfaces based only on sequence information typically have lower accuracies than methods incorporating evolutionary and structural information. Nevertheless, they are still widely used. Furthermore, the need for these methods is becoming more important with the increase of our knowledge on network rewiring produced by either mutations or alternative forms of proteins [23,24]. To demonstrate their use, methods based on sequence information are usually compared with random predictions that tend to have intrinsically very low accuracies. In this work, we also show an example of successful application of iFrag when one of the partners is a small peptide such as the β-amyloid, with only 42 aa, detecting a potential peptide able to prevent its aggregation.

Results

Performance and comparison of iFrag predictions with other methods

The prediction of protein interfaces using only sequence information is challenging; the probability of finding a limited set of pairs of contacting residues (interface) among all possible combinations of pairs of two sequences is extremely low. It is then not surprising that the overall performance of iFrag is low, but so are the rest of the current methods analysed in the work (Table 1). Most results obtained in the human interactome benchmark correspond to predictions of domains. In this sense, iFrag results are comparable in many examples to DOMINE, as it predicts the domains involved in a PPI. However, the combination of accuracy and applicability results in improved AUC-ROC and AUC-PR of iFrag with respect to DOMINE. iFrag is also competitive in terms of accuracy and applicability with the rest of approaches (SLIDER, PIPE-Sites, and PPIPP). Only the approach of BLAST (simple search of similarity among complexes with known structure; iFrag-BlastPDB) is better. This is highlighted as well by the improvement ratio, which is the highest for the iFrag-BlastPDB approach. Certainly, finding homologs involved in interactions in a known complex has the best performance; such data, however, are not always available (the exceptionally high applicability in Table 1 is caused by a bias in the validation set, because the knowledge of 3D structure is a requisite for the validation). Also, failing to obtain a prediction was only considered in the calculation of the applicability; for the rest of statistic measures, the prediction was neglected (i.e., neither as true nor false on negative and positive predictions). Only the PPIPP method achieved 100% applicability.

The other aspect that influences the level of success is to define how a prediction can be considered successful. We should consider that the probability of

Table 1. Summary of comparison of sequence-based prediction methods

Method	AUC ROC	AUC PR	MCC	F_1	PPV	Applicability (%)	Improvement ratio
DOMINE	0.66	0.00141	0.0192	0.0032	0.0016	47	1.97
BLAST	0.67	0.05150	0.2222	0.0204	0.1422	78	155.45
PIPE-Sites	0.55	0.00158	0.0204	0.0101	0.005	70	6.11
PPIPP	0.53	0.00680	0.0105	0.0165	0.0087	100	1.29
SLIDER	0.50	0.00057	0.0028	0.0034	0.0025	6	4.52
iFrag	0.74	0.00241	0.0107	0.0023	0.0011	52	1.45

AUC-ROC, AUC-PR, MCC, F1, PPV stand for area under the ROC curve, area under the precision-recall curve, Matthew's correlation coefficient, F_1-scores, and positive prediction value, respectively. The applicability is calculated as the percentage of protein pairs in the validation set in which each method can be applied (i.e., it returns a prediction), and the improvement ratio is calculated as the ratio between the total of residue pairs of the interfaces and the total of possible residue pairs summed for all PPIs.

finding a limited set of pairs of contacting residues among all possible combinations of pairs of two sequences is extremely low. Thus, several criteria have been used to evaluate the quality of a prediction. In PIPE-Sites, a prediction was considered successful if the predicted residues were in a range near the real solution using a neighbouring distance-measure criterion. In DOMINE, a prediction was successful if the residues of the interface were within the margins of the predicted PFAM domain. In our comparison, we have only used the native interface of residues (as defined by the 3D structure), which decreases the probability of success and produces low accuracy. Consequently, we should also compare the results of all approaches with a random prediction (i.e., by means of the improvement ratio; see Table 1). All approaches perform better than random, but the largest improvement is obtained when running the iFrag server using the BlastPDB approach; when searching for domains, iFrag improvement ratio is comparable to that of DOMINE.

Examples of predicted interfaces: from short fragments to entire domains

iFrag is a versatile approach to predict interfaces that can be domains or sequence fragments, that is, predictions are done at residue level. In Fig. 2, we show the results for the prediction of the binding regions between E3 ubiquitin-protein ligase (RING2_HUMAN) and Polycomb complex protein BMI-1 (BMI1_HUMAN) and the dimerisation of nucleoplasmin-2 (NPM2_HUMAN). In both examples, the region corresponds to the size of two interacting domains, but the sequence fragments involved in the interface are distributed along their sequence. Still, for the dimerisation of nucleoplasmin-2 (Fig. 2b), the prediction with iFrag helps to better define the actual interface, providing an accurate higher resolution than just the PFAM domains involved in the interaction (i.e., such as the prediction of DOMINE).

As mentioned above, iFrag can also predict smaller regions such as peptide fragments when the information on interactions is available. One of such examples is illustrated in the case of the study of the interaction between the amyloid beta peptide (amyloid beta peptide (Aβ), with length of 42 Aas) and clusterin (APOJ, also named CLUS_HUMAN, with length of 449 Aas) and serum albumin (ALBU_HUMAN, with length of 609 Aas). In this example, iFrag predicts several short-length regions with high score, two from clusterin (at the N-tail and the C-tail regions) and several from albumin (1 at the C-tail region and 3–4 distributed in the middle). The question is then whether these short peptides can indeed bind to the amyloid beta peptide and prevent its aggregation. This provides an opportunity to check the potential application of iFrag with an experimental validation. First, we extract the sequence fragments of albumin and clusterin with predicted high scores for their interactions with amyloid beta peptide. Then, we compare the sequences of these regions with an alignment, using ClustalW [27], and select the best patterns that include at least one fragment from each interactor (clusterin and albumin). After searching for common sequence patterns with the extracted sequences, only three regions show a potential good alignment. Supplementary Figure S1A shows the approach taken to select the best candidate sequences (two regions of albumin, in the middle and the C-tail regions, plus the N-tail region of clusterin). The three peptide fragments were synthetised, and we confirmed experimentally, using Bitan and Teplow protocol [28] and the Thioflavin T (ThT) aggregation assay, that at least one, the C-tail region of albumin, interferes with the aggregation of solubilised $A\beta_{40}$ peptide (see Supplementary Fig. S1B).

Discussion and Conclusion

In this work, we have presented iFrag, a computational approach to predict binding regions between two proteins based on minimal stretches with similar sequence of known interacting proteins. The predicted interfaces range from short fragments, composed by few residues, to complete domains or proteins. iFrag provides a user-friendly interface and a comprehensive results page. Traceability of sequence alignments and known PPIs allows the user to comprehend the results and devise new experiments that could be relevant for an interaction. For example, predictions on the interacting region could be useful to design experiments for disrupting an interaction or increasing the binding affinity or could be useful to combine with other computational techniques such as alanine-scanning methods to detect potential hot-spots in predicted binding sites. We have included one of these examples on the application to infer a potential peptide region binding the amyloid beta peptide. We have compared iFrag with other approaches that predict specific binding sites of two interacting proteins, such as DOMINE (or a similar approach, DIMERO [29], not included here because the evaluation yields the same results), PIPE-Sites, PPIPP, and SLIDER. Domain-based predictors such as DOMINE include large regions of the protein, while PIPE-Sites, PPIPP, and SLIDER exploit PPI networks to infer shorter fragments of sequences limited to the interface region.

We have proved that iFrag is versatile to be applied for the prediction of short and large regions with competing accuracy [i.e., iFrag obtains the larger AUC-ROC compared to other methods and similar percentages of positive predictive value (PPV) and applicability]. One of the problems faced by iFrag, as any other method of prediction of protein–protein interfaces based on sequence, is the value of the statistic measures used to evaluate them: we used classical statistical metrics such as accuracy, PPV, Matthew's correlation coefficient (MCC), F1, the area under the ROC curve (AUC-ROC), or the area under the precision curve (AUC-PR) as a measure of the quality of the prediction. However, results of these classical measures would discourage the potential use of ours and similar approaches, while these methods can still help experimentalist on the selection of potential binding regions specific of a PPI (i.e., see example above). This is because classical statistic metrics can lead to incorrect interpretation of the results when applied to contact map predictions. First, because they assume that individual predictions (residue–residue pairs) are independent, which is not true, once a residue–residue contact is predicted, its sequence neighbours should be affected and increase its probability as well. Second, this is also true for the structural neighbours of a residue, which is a property hidden in the sequence but still applies in the corresponding sequence prediction. Third, as the size of proteins varies, the chances of obtaining a correct prediction by random vary too; large proteins imply lower probabilities by random and vice versa [30]. The statistical metrics to evaluate the quality of these predictions are affected by the nature of the problem. Therefore, we have introduced in the supplementary material a new metric to evaluate and compare these heterogeneous approaches, proving that iFrag is comparable to PPiPP and confirming the best performance of BlastPDB in iFrag (see Supplementary Data). In our conclusion, iFrag is a useful tool that can help experimentalists to select the regions that may be involved in specific PPIs for further tests or synthesis of function-specific peptides.

iFrag computational approach can be used by a user-friendly web server that provides a platform to analyse the predictions presented on a comprehensive and intuitive web page with the results. Traceability of sequence alignments and known PPIs allows the user to comprehend the results and devise new experiments that could be relevant for an interaction. For example, predictions on the interacting region could be useful to design experiments for disrupting an interaction or increasing the binding affinity or could be useful to combine with other computational techniques such as alanine-scanning methods to detect potential hot-spots in predicted binding sites. As an example, we have proved a direct biological application of iFrag in helping to select a peptide that hinders the aggregation of the amyloid beta peptide.

Materials and Methods

Sources of PPIs

PPIs were integrated with BIANA [31], downloading the data from IntAct [32], DIP [33], BioGRID [34], HPRD [35], MINT [36], and MPact [37]. iFrag server also includes sequence annotation from Uniprot [21] and assigns domains from PFAM [22] using HMMER [38].

Minimal common fragments search and scoring (iFrag computational approach)

Query protein sequences are compared and aligned against all sequences in the BIANA database [31] (i.e., with all integrated PPIs) using BLAST [39]. The alignments provide a similarity measure based on the percentage of identical residues aligned, the percentage of the sequence covered by the alignment, and the E-value. We use BLAST to find short-fragment alignments by using low percentages of sequence similarity or high E-value thresholds selected by the user. To avoid unnecessary repetitions of BLAST searches, iFrag uses a local database of similarity measures already stored.

The resulting alignments from the BLAST searches are then used to score pair of residues to define putative interfaces. Sequence fragments of proteins aligned with the queries are paired if they belong to proteins with a known interaction (template interactions). We filter out template interactions reducing the set to pairs with less than 40% of sequence identity (in agreement with a previous work [40], close homologs

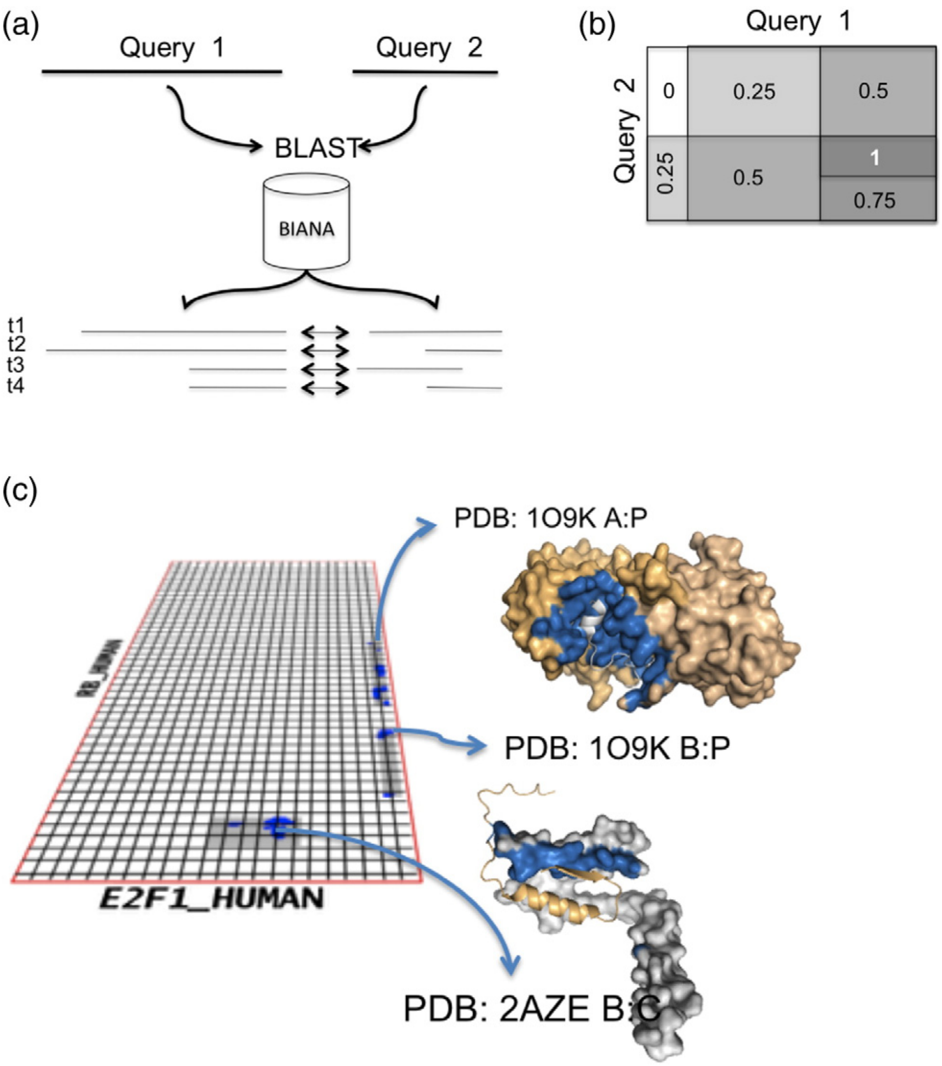


Fig. 1. iFrag outline of the method and the evaluation of a contact map. (a) Two query proteins are compared with sequences known to have reported interactions using BLAST [39]. Matches are grouped by paired sequence fragments that belong to known interacting proteins (template interactions). The set of template interactions is filtered to avoid redundancies, so that it does not contain any pair of template proteins with more than 40% of sequence identity. (b) The iFrag score is calculated as the proportion of matches from BLAST covering two residues, one in each protein, over the total number of known interactions. (c) Example of the contact map for the interaction between RB_HUMAN and E2F1_HUMAN used for evaluation. The interaction between retinoblastma (RB_HUMAN) and transcription factor E2F1 (E2F1_HUMAN) is produced by different interfaces. The contact map between them is represented in a grid. Grey areas of the contact map show the protein regions found in the PDB structures. Blue areas show specific residue–residue contacts. The complete contact map between these two proteins is composed by the union of all interface regions, which is obtained with the structures with PDB codes 2AZE (chains B and C) and 1O9K (chains, A, B, and P).

with 30–40% or higher sequence identity almost invariably interact the same way). The score of a pair of residues, one in each query protein, is calculated as the proportion of matches of template interactions over the total. The E-value of an alignment is used as criterion for including a template interaction. Sequence fragments of template interactions matching large parts of the query proteins are less informative than those covering short regions. Thus, the user can also restrict the set of template interactions by the percentage of the query sequences covered by the templates. An outline of the method is explained in Fig. 1.

Evaluation dataset on the human interactome

To evaluate the performance of iFrag and compare it to current state-of-the-art methods, we use a

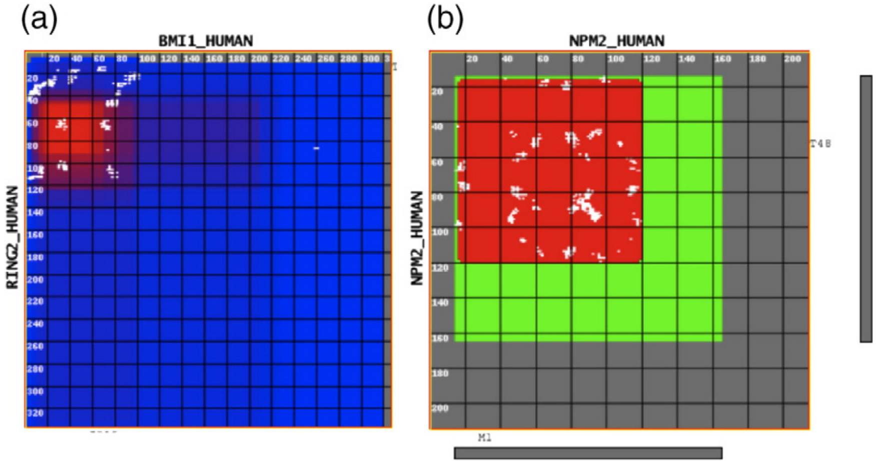


Fig. 2. iFrag server heat map examples. (a) Prediction on the interaction between RING2_HUMAN and BMI1_HUMAN; blue and red indicates low and high scores, respectively, while indicated in white are the real interface contacts as defined from the 3D structure of the protein complex (PDB code 2CKL [25]). (b) The homo-dimerisation of NPM2_HUMAN is done through its nucleoplasmin domain (Pfam PF03066). Red indicates high iFrag scores, green indicates the domain–domain interaction as predicted by DOMINE, and grey is a null prediction. Interface residue pairs as defined from the 3D structure of the dimer (PDB code 3T30) are showed in white [26].

set of validated interactions. The evaluation dataset is composed of a non-redundant set (less than 40% of sequence identity) of human protein complexes extracted from Uniprot database [21] for which the 3D structure is known and deposited in the PDB databank [41]. In order to avoid the possible bias produced by hub proteins [42], the sequence identity was calculated at the protein level, so that no protein in the dataset has more than 40% sequence identity with any other protein in the dataset. In this dataset, we use the whole sequence of the protein as defined in Uniprot and we exclude protein complexes that include functional complexes such as tandem affinity purification or co-sedimentation, that is, those which interactions between partners are not necessarily physical and direct. We consider only the biological assembly of structures obtained from the PDB database to exclude crystal contacts and non-biologically relevant complexes. One protein complex can be represented by more than one structure and we define the interface of each interaction as the set of contacting residues (we define that two residues are in contact if the distance between their carbon-β atoms is shorter than 12 Å).

iFrag web server

A server has been implemented to facilitate the use of iFrag. The input of iFrag is two sequences in FASTA format (query proteins). The user can modify the BLAST search conditions by specifying the maximum E-value threshold and coverage of the alignments with the query. The set of template interactions can also be selected by the method of experimental detection (e.g., by excluding co-complex methods).

Upon computing the predictions, iFrag outputs an interactive bidimensional heat map representing a scored contact map. The user can inspect the score of specific residue pairs and select regions of interest by browsing on the heat map. The output is complemented with sequence feature annotations described in Uniprot database [21] and matches with PFAM domains [22]. These are useful features that, in a real case scenario, can help the user to identify or discard regions of interest for the design of experiments. The alignments of the query proteins with the template interactions are also provided as part of the output. The server also shows the database and detection method of each template interaction. The traceability of sequence alignments with known PPIs allows the user to comprehend the results in a more rational way and help him decide on the relevance of the findings. The user can optionally select the potential region of binding and extract the multiple sequence alignment of the sequences found by BLAST in this region. This is retrieved in the form of a motif or sequence pattern for each of the interacting proteins, and it can be used to filter the fragments predicted as interface (see further in Experimental validation and Supplementary Fig. S1).

Comparison with other methods

Standard statistical metrics such as PPV, MCC, F_1-measure, AUC-ROC, and AUC-PR are used to assess and compare the performance of iFrag and other current approaches (see Supplementary Table S1 for details). The evaluation set described above is used to identify the interfaces of PPIs and define the true and false positives and negatives of each prediction in the form of a contact map (see example in Fig. 1c). To obtain iFrag predictions, we used only binary PPIs as templates (i.e., co-complex-derived interactions were excluded). We used an E-value threshold of 0.01 and none for sequence identity (0%), allowing for short fragments.

We have compared iFrag with PIPE-Sites [12], PPIPP [14], SLIDER [13], and DOMINE [11] using high confidence predictions between domains of PFAM (see Supplementary Table S2 for details). It is worth mentioning that some of these web servers might have been trained with proteins included in the validation dataset (e.g., PPIPP). This implies that some methods may perform better than expected and consequently better than iFrag. Still, most of them used the Docking Benchmark 4.0 [15] for training, while we selected a specific human interactome to try to reduce this potential overfitting. Additionally, we have also compared with the simplest approach based on sequence comparison (BLAST); the interface between two query proteins is based on their similarity to known structure complexes, that is, the same strategy as in homology modelling of protein complexes. For this test, we require a minimum of 30% identical residues aligned. This method is also available in iFrag (iFrag-BlastPDB, see above). Methods that were not accessible as servers, like a similar approach that uses functional motifs as alternative to BLAST [43], could not be considered in our comparison. We compared all methods with the random prediction of binding sites by means of the improvement ratio (i.e., the ratio of correct amino acid pair predictions over the total number of residue pairs available in the benchmark).

Experimental validation: Aβ_{40} aggregation assay

We predicted several peptides that could interact with the amyloid beta peptide Aβ_{40} using iFrag. Among them, we selected the C-tail peptide of albumin for experimental validation on a ThT aggregation assay. The experiment was defined as follows: lyophilised Aβ_{40} (Anaspec) was solubilised as previously described by Bitan and Teplow [28]. Briefly, 1 mg of Aβ_{40} was dissolved in 250 μL of MilliQ water, and pH was adjusted to $\geq$10.5 using 1 M NaOH solution. Then, 250 μL of 20 mM phosphate buffer (pH 7.4) was added. The preparation was placed for 1 min in a bath-type sonicator (Bioruptor, Diagenode) and immediately used for experiments. A 1-mM stock ThT (Sigma-Aldrich) solution was prepared by dissolving the dye in phosphate-buffered-saline solution (PBS). The solution was filtered through a 0.22-mm filter. Aβ_{40} peptide (10 μM) was incubated with or without albumin C-tail peptide (20 μM C-term;Peptide Synthesis Facility, UPF; AETFTFHADICTLSEKERQIKKQTALVELVKHKPK-amide) and 10 μM ThT in a Nunc-96-well flat bottom black polystrol microplate (Thermo Scientific) at 37 °C. ThT fluorescence was measured at 0, 24, and 48 h using excitation and emission wavelengths of 430 and 470 nm, respectively, using a multiplate reader fluorimeter (FLUOstar optima, BMG labtech).

Acknowledgements

This work was supported by the Spanish Ministry of Economy (MINECO) with FEDER BIO2014-57518-R, Fondo de Investigación Sanitaria PI13/00408, and Red HERACLES RD12/0042/0014, EG is supported by Beatriu de Pinós Fellowship from the Agency for Management of University and Research Grants (AGAUR) of Government of Catalunya.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2016.11.034.

Received 29 June 2016;
Received in revised form 27 November 2016;
Accepted 30 November 2016
Available online 10 December 2016

Keywords:
binding site prediction;
protein interactions;
interface prediction;
β-amyloid aggregation

Abbreviations used:
PPI, protein–protein interaction; PPV, positive predictive value; MCC, Matthew's correlation coefficient; AUC-ROC, area under the ROC curve; AUC-PR, area under the precision-recall curve; ThT, Thioflavin T.

References

[1] M. Boxem, Z. Maliga, N. Klitgord, N. Li, I. Lemmens, M. Mana, et al., A protein domain-based interactome network for *C. elegans* early embryogenesis, Cell 134 (2008) 534–545.

[2] J.G. Kiselar, P.A. Janmey, S.C. Almo, M.R. Chance, Visualizing the Ca^{2+}-dependent activation of gelsolin by using synchrotron footprinting, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 3942–3947.

[3] W.S. Valdar, J.M. Thornton, Conservation helps to identify biologically relevant crystal contacts, J. Mol. Biol. 313 (2001) 399–416.

[4] J. Hoskins, S. Lovell, T.L. Blundell, An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements, Protein Sci. 15 (2006) 1017–1029.

[5] I. Halperin, H. Wolfson, R. Nussinov, Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families, Proteins 63 (2006) 832–845.

[6] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, B.B.I.P.S. Oliva, BIANA interolog prediction server. A tool for protein–protein interaction inference, Nucleic Acids Res. 40 (2013) W147–W151.

[7] J. Fernandez-Recio, M. Totrov, C. Skorodumov, R. Abagyan, Optimal docking area: a new method for predicting protein–protein interaction sites, Proteins 58 (2005) 134–143.

[8] J. Espadaler, O. Romero-Isart, R. Jackson, B. Oliva, Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships, Bioinformatics 21 (2005) 3360–3368.

[9] A. Henschel, C. Winter, W.K. Kim, M. Schroeder, Using structural motif descriptors for sequence-based binding site prediction, BMC Bioinformatics 8 (2007) S5.

[10] J. Planas-Iglesias, M.A. Marin-Lopez, J. Bonet, J. Garcia-Garcia, B. Oliva, iLoops: a protein–protein interaction prediction server based on structural features, Bioinformatics 29 (2013) 2360–2362.

[11] S. Yellaboina, A. Tasneem, D.V. Zaykin, B. Raghavachari, R. Jothi, DOMINE: a comprehensive collection of known and predicted domain-domain interactions, Nucleic Acids Res. 39 (2011) D730–D735.

[12] A. Amos-Binks, C. Patulea, S. Pitre, A. Schoenrock, Y. Gui, J.R. Green, et al., Binding site prediction for protein–protein interactions and novel motif discovery using re-occurring polypeptide sequences, BMC Bioinformatics 12 (2011) 225.

[13] P. Boyen, D. Van Dyck, F. Neven, R.C. van Ham, A.D. van Dijk, SLIDER: a generic metaheuristic for the discovery of correlated motifs in protein–protein interaction networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (2011) 1344–1357.

[14] S. Ahmad, K. Mizuguchi, Partner-aware prediction of interacting residues in protein–protein complexes from sequence data, PLoS One 6 (2011), e29104. http://dx.doi.org/10.1371/journal.pone.0029104.

[15] H. Hwang, T. Vreven, J. Janin, Z. Weng, Protein–protein docking benchmark version 4.0, Proteins 78 (2010) 3111–3114.

[16] J. Segura, P.F. Jones, N. Fernandez-Fuentes, Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams, BMC Bioinformatics 12 (2011) 352.

[17] J. Segura, P.F. Jones, N.A. Fernandez-Fuentes, Holistic in silico approach to predict functional sites in protein structures, Bioinformatics 28 (2012) 1845–1850.

[18] D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution, Nat. Rev. Genet. 14 (2013) 249–261.

[19] M.J. Skwark, A. Abdel-Rehim, A. Eloffsson, PconsC: combination of direct information methods and alignments improves contact prediction, Bioinformatics 29 (2013) 1815–1816.

[20] R. Esmaielbeiki, K. Krawczyk, B. Knapp, J.C. Nebel, C.M. Deane, Progress and challenges in predicting protein interfaces, Brief. Bioinform. 17 (2016) 117–131.

[21] C. UniProt, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.

[22] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, et al., The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res. 44 (2016) D279–D285.

[23] N. Sahni, S. Yi, M. Taipale, J.I. Fuxman Bass, J. Coulombe-Huntington, F. Yang, et al., Widespread macromolecular interaction perturbations in human genetic disorders, Cell 161 (2015) 647–660.

[24] X. Yang, J. Coulombe-Huntington, S. Kang, G.M. Sheynkman, T. Hao, A. Richardson, et al., Widespread expansion of protein interaction capabilities by alternative splicing, Cell 164 (2016) 805–817.

[25] G. Buchwald, P. van der Stoop, O. Weichenrieder, A. Perrakis, M. van Lohuizen, T.K. Sixma, Structure and E3-ligase activity of the Ring–Ring complex of polycomb proteins Bmi1 and Ring1b, EMBO J. 25 (2006) 2465–2474.

[26] O. Platonova, I.V. Akey, J.F. Head, C.W. Akey, Crystal structure and function of human nucleoplasmin (npm2): a histone chaperone in oocytes and embryos, Biochemistry 50 (2011) 8078–8089.

[27] J.D. Thompson, H. D. G ., T.J. Gibson, et al., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.

[28] G. Bitan, D.B. Teplow, Preparation of aggregate-free, low molecular weight amyloid-beta for assembly and toxicity assays, Methods Mol. Biol. 299 (2005) 3–9.

[29] J. Segura, C.O. Sorzano, J. Cuenca-Alba, P. Aloy, J.M. Carazo, Using neighborhood cohesiveness to infer interactions between protein domains, Bioinformatics 31 (2015) 2545–2552.

[30] J. Martin, Benchmarking protein–protein interface predictions: why you should care about protein size, Proteins 82 (2014) 1444–1452.

[31] J. Garcia-Garcia, E. Guney, R. Aragues, J. Planas-Iglesias, B. Oliva, Biana: a software framework for compiling biological interactions and analyzing networks, BMC Bioinformatics 11 (2010) 56.

[32] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, et al., The IntAct molecular interaction database in 2012, Nucleic Acids Res. 40 (2012) D841–D846.

[33] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, D. Eisenberg, DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Res. 30 (2002) 303.

[34] C. Stark, B.J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M.S. Livstone, et al., The BioGRID interaction database: 2011 update, Nucleic Acids Res. 39 (2011) D698–D704.

[35] R. Goel, H.C. Harsha, A. Pandey, T.S. Prasad, Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis, Mol. BioSyst. 8 (2012) 453–463.

[36] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, et al., MINT, the molecular interaction database: 2012 update, Nucleic Acids Res. 40 (2012) D857–D861.

[37] U. Guldener, M. Munsterkotter, M. Oesterheld, P. Pagel, A. Ruepp, H.W. Mewes, et al., MPact: the MIPS protein interaction resource on yeast, Nucleic Acids Res. 34 (2006) D436–D441.

[38] R.D. Finn, J. Clements, W. Arndt, B.L. Miller, T.J. Wheeler, F. Schreiber, et al., HMMER web server: 2015 update, Nucleic Acids Res. 43 (2015) W30–W38.

[39] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[40] P. Aloy, H. Ceulemans, A. Stark, R.B. Russell, The relationship between sequence and interaction divergence in proteins, J. Mol. Biol. 332 (2003) 989–998.

[41] S. Velankar, G. van Ginkel, Y. Alhroub, G.M. Battle, J.M. Berrisford, M.J. Conroy, et al., PDBe: improved accessibility of macromolecular structure data from PDB and EMDB, Nucleic Acids Res. 44 (2016) D385–D395.

[42] J. Yu, M. Guo, C.J. Needham, Y. Huang, L. Cai, D.R. Westhead, Simple sequence-based kernels do not predict protein–protein interactions, Bioinformatics 26 (2010) 2610–2614.

[43] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. Gibson, et al., Systematic discovery of new recognition peptides mediating protein interaction networks, PLoS Biol. 3 (2005) e45.