Bioinformatics, 31(15), 2015, 2545–2552
doi: 10.1093/bioinformatics/btv188
Advance Access Publication Date: 2 April 2015
Original Paper

Systems biology

Using neighborhood cohesiveness to infer interactions between protein domains

Joan Segura[1,*], C. O. S. Sorzano[1], Jesus Cuenca-Alba[1], Patrick Aloy[2,3] and J. M. Carazo[1]

[1]GN7 of the National Institute for Bioinformatics (INB) and Biocomputing Unit, National Center of Biotechnology (CSIC), c/ Darwin no 3, Campus of Cantoblanco, 28049, Madrid, Spain, [2]Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), c/ Baldiri Reixac 10-12, 08028, Barcelona, Spain and [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010, Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on December 1, 2014; revised on March 12, 2015; accepted on March 28, 2015

Abstract

Motivation: In recent years, large-scale studies have been undertaken to describe, at least partially, protein-protein interaction maps, or interactomes, for a number of relevant organisms, including human. However, current interactomes provide a somehow limited picture of the molecular details involving protein interactions, mostly because essential experimental information, especially structural data, is lacking. Indeed, the gap between structural and interactomics information is enlarging and thus, for most interactions, key experimental information is missing. We elaborate on the observation that many interactions between proteins involve a pair of their constituent domains and, thus, the knowledge of how protein domains interact adds very significant information to any interactomic analysis.

Results: In this work, we describe a novel use of the neighborhood cohesiveness property to infer interactions between protein domains given a protein interaction network. We have shown that some clustering coefficients can be extended to measure a degree of cohesiveness between two sets of nodes within a network. Specifically, we used the meet/min coefficient to measure the proportion of interacting nodes between two sets of nodes and the fraction of common neighbors. This approach extends previous works where homolog coefficients were first defined around network nodes and later around edges. The proposed approach substantially increases both the number of predicted domain-domain interactions as well as its accuracy as compared with current methods.

Availability and implementation: http://dimero.cnb.csic.es

Contact: jsegura@cnb.csic.es

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

1 Introduction

Cellular processes are regulated by an intricate network of protein-protein interactions (PPIs) that form both transient as well as more permanent complexes. Different experimental techniques are used to determine PPIs (Berggard *et al.*, 2007). However, the molecular details behind the interaction itself are usually not detected and require other types of assays to be determined. In this context, it is well-known that protein function is supported by the underlying structure (Watson *et al.*, 2005) and, thus, solving the structure of protein complexes is essential for a detailed understanding of their

function. Still, the amount of interactomics data, and the speed at which they are produced, is substantially higher than the pace at which protein structures can be experimentally obtained. Consequently, the development of computational methods to predict some of the details behind sets of proposed PPIs is a must. Lacking experimental structural information about a certain complex, but having that information for its unbound constituent proteins, probably the best approaches fall in the area of protein binding site prediction (Segura *et al.*, 2011) or protein docking (Halperin *et al.*, 2002). Nonetheless, most of these methods are only suitable when the structure of proteins have been solved, at least partially. This fact has motivated us to focus on the prediction of some characteristics of sets of PPIs based on non-structural information. In this way, we have focused on the important observation that most proteins are composed of one or more domains connected by inter-domain regions (Apic *et al.*, 2001; Ekman *et al.*, 2005). Moreover, many protein interactions involve a pair of their constituent domains and, thus, how a particular pair of proteins interacts is strongly driven by their domains (Apic *et al.*, 2001; Gupta *et al.*, 2010; Itzhaki *et al.*, 2006) (the important case of intrinsically disordered proteins will be discussed in Section 3.3). Consequently, the knowledge of domain-domain interactions (DDIs) is of great value during the analysis of PPIs, helping, for instance, in the identification of potential binding sites (Desjarlais and Berg, 1992; Moya-Garcia and Ranea, 2013; Pawson and Nash, 2003).

The question whether PPIs are mediated via a limited set of domain pairs has been approached in different works. Several studies found that some domain pairs were overrepresented in large datasets of experimentally determined PPIs (Deng *et al.*, 2002; Gomez *et al.*, 2003; Liu *et al.*, 2005; Ng *et al.*, 2003; Nye *et al.*, 2005; Riley *et al.*, 2005; Sprinzak and Margalit, 2001). Therefore, the inferred domain pairs were shared by multiple interactions. Using a different type of analysis, Itzhaki *et al.* (2006) studied the relation between structurally derived DDIs and PPI networks of different organisms. Their statistical analysis proved that the number of PPIs attributed to DDIs was significantly larger than expected by random. This result supports the conjecture that PPIs may be driven by a limited catalogue of DDIs. However, the fraction of PPIs to which experimentally determined DDIs could be mapped back in 2006 ranged from 6 to 20% in different organisms. Indeed, in this study, we have observed that the gap between PPIs and experimentally solved DDIs nowadays remains in a similar proportion, so that current experimental data on DDIs can be mapped to <20% of the PPIs for most organisms (Section 3.3.2).

Most previous methods to infer DDIs rely on the assumption that an interaction between a pair of proteins involves at least a pair of their constituent domains (one from each interacting protein). One of the first methods was proposed by Sprinzak and Margalit (2001). In their work, the authors analyzed a particular set of PPIs looking for domain pairs that co-occurred more frequently in interactions than expected by chance. Later, Deng *et al.* (2002) used an expectation maximization algorithm to maximize a certain likelihood function over an observed interactome. Protein interactions were described in terms of DDI probabilities, and the expectation maximization algorithm searched for the DDI probabilities that maximized the likelihood function. This work was extended by Riley *et al.* (2005) in their domain pair excluding analysis approach. Riley *et al.* method calculates different likelihood scores negating interactions coming from particular domain pairs. In this way, the authors improved the performance of previous methods, predicting interaction between domains based on the differences of the likelihood values when a particular domain pair was excluded. Other approaches used a parsimony model to find the minimum set of DDIs that could explain the interactions present in a given PPI map (Chen *et al.*, 2011). Finally, several methods introduced other types of data, such as correlated mutation at protein interfaces (Jothi *et al.*, 2006; Kann *et al.*, 2007), gene ontology (GO) terms (Liu *et al.*, 2009) or co-evolutionary data (Pazos and Valencia, 2008).

In this work, we describe a new strategy to infer DDIs based on the topology of a protein interaction network. This approach exploits the neighborhood cohesiveness (NC) property of small-world networks (Watts and Strogatz, 1998), analyzing the interactions between two sets of nodes that contain proteins with a particular domain. In this way, we have extended previous definitions of NC, first calculated on a network node (Watts and Strogatz, 1998), later on a network edge (Goldberg and Roth, 2003) and in this work defined over two sets of nodes within the network. Particularly, we used the meet/min coefficient (Ravasz *et al.*, 2002) to measure the proportion of interacting nodes between two sets of nodes and the fraction of common neighbors. Although more sophisticated clustering coefficients (Hulovatyy *et al.*, 2014) could be used, we will prove that this simple cohesiveness measure already has enough discriminative power to differentiate between interacting and non-interacting domains. The performance has been tested using a novel DDI benchmark compiled using 3DID (Mosca *et al.*, 2014) and Negatome (Blohm *et al.*, 2014) data. We have compared our methodology against the correlated sequence signature (CSS) method (Sprinzak and Margalit, 2001) and DOMINE predicted data (Yellaboina *et al.*, 2011), showing that the proposed approach is competitive with previous methods. Finally, the approach and the PPI data used during the present development have been integrated together into a web platform termed DIMERO (Data Integration for MolEcuar stRucture mOdelling). DIMERO allows both the evaluation of interactions between domains as well as a direct access to source data in the form of the PPIs used for the evaluation. In this way, browsing these PPI data offers the possibility to find additional information about known interactions involving the domains of interest, including experimental data, information from prediction methods and scientific literature.

2 Materials and methods

2.1 The negative sample problem

The lack of experimentally determined non-interacting domain pairs makes the evaluation of DDI prediction methods a challenging problem. Typically, positive cases for testing DDI predictions are collected from crystal structures of protein complexes (Jothi *et al.*, 2006; Liu *et al.*, 2009). However, the negative sampling has been handled in different ways. Several works compared the distributions of scores between known interacting domains and random pairs selected from proteins of different cellular compartments (Ben-Hur and Noble, 2006). Other works measured the fold enrichment between the number of predicted interactions and the number of predictions known to be true (Liu *et al.*, 2009). Also, several authors used PPI datasets to evaluate their DDI predictions (Chen *et al.*, 2011; Sprinzak and Margalit, 2001); nonetheless, it should be noted that the negative sampling for PPI presents a similar problem. In this work, we followed a different approach, using the data collected in the Negatome database (Blohm *et al.*, 2014). This database is a repository of protein and domain pairs that are unlikely to engage in a physical interaction. The data are collected by means of data mining algorithms and filtered by manual curation, minimizing the number of false negative pairs of the resulting dataset.

2.2 Data framework

Four different databases were used for developing and benchmarking the here presented DDI scoring system. STRING DB (Franceschini *et al.*, 2013) was the source of the interactomic networks for different organisms. This database offers a large collection of experimentally obtained, data mined as well as predicted interactions between proteins. For the purpose of this work, only interactions that were annotated as 'binding' were considered. The resulting dataset consisted of 28 888 908 interactions involving 2 135 719 proteins from 1133 organisms. Supplementary Figure S1 in Supplementary Material shows that the distribution and relation between the number of proteins and interactions contained in the different networks. Then, for each protein involved in these interactions, its constituent domains were defined in terms of the Pfam classification (v27) (Finn *et al.*, 2014) and delineated using the HMMER algorithm (Johnson *et al.*, 2010). Finally, this protein domain information was annotated into the interactomic network nodes, setting up the framework for the assessment of domain interactions.

The DDI benchmark was constructed using two databases: 3DID (Mosca *et al.*, 2014) and Negatome DB (Blohm *et al.*, 2014). 3DID is a compilation of structurally solved domain interactions mined from the PDB (Rose *et al.*, 2013). The positive cases were extracted from this database selecting the domain interactions between different protein chains of binary complexes. Finally, the interactions between domains of the same family were rejected. The resulting dataset contained 1405 Pfam domain pairs; this set will be referred to as positive domain–domain interaction (PDDI). The negative samples of the benchmark were gathered from Negatome DB, using the 'Manual-Pfam' set, containing 1453 domain pairs which, in the context of this work, will be referred to as negative domain-domain interaction (NDDI).

2.3 From network topology to DDI scores

Our DDI assessment methodology exploits the NC property of small-world networks in a similar way to the approach proposed by Goldberg and Roth (2003). In their work, Goldberg and Roth defined four variants of a mutual cluster coefficient to measure NC around edges in a protein network, i.e. interactions. The authors proved that measurements of NC around network edges (protein interactions) had higher values than in pairs of proteins for which there was no evidence of interaction. Therefore, proteins sharing many neighbors in a PPI network were more likely to interact. Finally, these metrics were used to predict false interactions in protein networks derived from high-throughput experiments. Following a similar strategy, we have extended the meet/min coefficient (Ravasz *et al.*, 2002) defining two metrics that quantify a cohesiveness degree between two sets of nodes within a network. This approach extends previous works where a mutual cluster coefficient was defined, first around network nodes (Watts and Strogatz, 1998) and later on around edges (Goldberg and Roth, 2003). Furthermore, we prove that these measures can be used to infer DDIs when the selected sets include the network proteins that contain specific domains (Sections 3.1 and 3.2).

The main purpose of this work is the identification of potential interactions between domain families and its application to explore protein interactomic data. For that reason, the scores presented in this section have been expressed as a function of protein domains instead of using a graph-based description.

In this work, a protein interaction network for a particular organism, as the one shown in Figure 1A, is described as a 3-tuple (P, Δ, I), where P is the set of proteins in the network, Δ is the set of all domains in the proteins and, finally, $IP \times P$ is the set of PPIs in the network, being (P, I) an undirected graph.

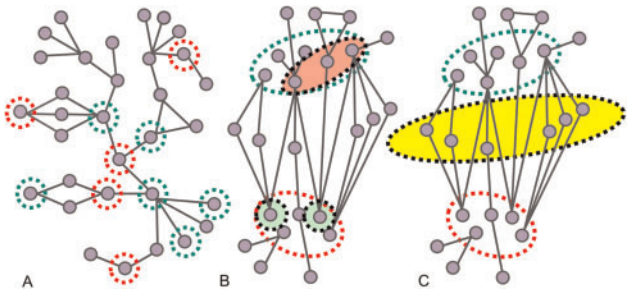

Fig. 1. Network cohesiveness between protein domains. (A) An exemplary PPI network. In dashed red lines, proteins that define the network region associated to domain d_a. In dashed green lines, proteins that define the network region associated to domain d_b. (B) The PPI network in (A) has been spatially rearranged while keeping the node connectivity, so that all proteins of the network region defined by d_a are placed within the dashed red line and those associated to d_b are within the dashed green line. In a solid red area delineated by a dashed black line, we show those directly interacting nodes of d_a contained in the network region of d_b, the opposite case is shown within solid green areas. In this example $c_{IP} = \max\{2/12, 3/12\}$ (C) Common neighbors nodes of the proteins in the network region defined by d_a, encircled in red, and d_b, encircled in green, are shown within the solid yellow area. In this example $c_{NP} = \max\{6/12, 6/12\}$ (Color version of this figure is available at *Bioinformatics* online.)

For a given protein $p \in P$, its neighborhood in the network is defined as the set of nodes that interact with p, thus

$$N(p) = \{q \in P; (p, q) \in I\} \tag{1}$$

Let $\delta \in \Delta$ be a protein domain, then the network region of δ is defined as the set of nodes (proteins) that contain the domain δ and formally is denoted as

$$R(\delta) = \{q \in P; \delta \in q\} \tag{2}$$

The network surroundings of a domain δ is defined by the nodes that interact with any protein of $R(\delta)$, thus

$$N(\delta) = \bigcup_{p \in R(\delta)} N(p) \tag{3}$$

The first proposed metric is the interacting nodes proportion, noted as $c_{IP}(\delta_n, \delta_m)$. It measures the degree of directly interacting nodes that exists between two domain regions in the network. This metric calculates the maximum fraction of elements in the network surrounding of one domain that intersects with the region of a second domain

$$c_{IP}(\delta_n, \delta_m) = \max\left\{\frac{|N(\delta_n) \cap R(\delta_m)|}{|N(\delta_n)|}, \frac{|N(\delta_m) \cap R(\delta_n)|}{|N(\delta_m)|}\right\} \tag{4}$$

Figure 1B shows that the interacting nodes between two sets of proteins defined by two domain regions.

The second metric is the common neighborhood proportion; it measures a degree of NC between two domain regions and it is calculated as the maximum fraction of common nodes between the domain surroundings

$$c_{NP}(\delta_n, \delta_m) = \max\left\{\frac{|N(\delta_n) \cap N(\delta_m)|}{|N(\delta_n)|}, \frac{|N(\delta_m) \cap N(\delta_n)|}{|N(\delta_m)|}\right\} \tag{5}$$

Figure 1C represents the common neighbors between two sets of proteins defined by two domain regions.

The metrics presented in this section follow the same principles as the meet/min coefficient (Ravasz *et al.*, 2002), a broadly used measure in the context of PPI networks (Goldberg and Roth, 2003; Li *et al.*, 2011). More sophisticated approaches (Hulovatyy *et al.*, 2014) could be adapted to calculate cohesiveness between sets of proteins, potentially leading to more accurate results. However, the main purpose of this work is to prove that basic measures of NC have enough discriminative power to differentiate between interacting and non-interacting domains.

2.4 STRING data for the assessment of DDIs

Domain interaction metrics presented in Section 2.3 are calculated per interactomic network. Therefore, these calculations have to be extended to multiple interactomic networks from multiple organisms so that, for each one of them, metric values for the particular domain pairs of that network are obtained. Then, the global score is computed as the average among all organisms of the values corresponding to each domain pair; thus

$$s_*(\delta_n, \delta_m) = \frac{1}{|\Theta|} \sum_{o \in \Theta} c_*^o(\delta_n, \delta_m) \quad (6)$$

where $c_*^o(\delta_n, \delta_m)$ is the interacting node proportion (Equation 4), or the common neighborhood proportion (Equation 5), between domains δ_n and δ_m in the interactome o. Note that although averaging information from multiple organisms certainly led to better performance, metrics c_{IP} (Equation 4) and c_{NP} (Equation 5) computed on single interactomes already have enough discriminative power in themselves (Section 3.2 and Supplementary Material, Section S5).

Final assessment was done using established statistical measures such as Precision, Recall, areas under precision-recall (AUPR) and receiver operating characteristic (AUROC) curves, the Mathew correlation coefficient (MCC) and the Mann-Whitney-Wilcoxon (MWW) test, that are presented in detail in Supplementary Material (Supplementary Section S2).

2.5 Combining the interacting and neighborhood proportion scores

We combined the interacting nodes proportion score s_{IP} and the common neighborhood proportion score s_{NP} (Section 2.4) into a single binary classifier named the combined proportion (CP). The CP classifier is based in the following strategy: a domain pair (δ, δ') is classified as interacting if $s_{IP}(\delta, \delta') > K_{IP}$ or $s_{NP}(\delta, \delta') > K_{NP}$, where K_{IP} and K_{NP} are two thresholds selected under a certain criteria. For example, in this work we have used those thresholds that maximises the MCC of s_{IP} and s_{NP}, respectively (Section 3.1).

This strategy, however, does not explicitly provide a score or statistical confidence measure. Consequently, in Supplementary Material (Supplementary Section S3) we present the conventions used in this work to describe confidence levels as high confident (HCP), medium confident (MCP) and low confident predictions (LCP).

2.6 The web application: DIMERO WS

DIMERO WS is a dedicated web platform that integrates the predicted DDI scores and the CP classifier (Sections 2.4 and 2.5) with the PPI data used for their calculations. The platform is composed of a collection of python and JavaScript packages. The server was programmed in python language using the PDB structural library of Biopython (Hamelryck and Manderick, 2003). Also, HMMER (Johnson *et al.*, 2010) was integrated in the server to calculate the domains of protein sequences based on Pfam classification (Finn *et al.*, 2014). The client interface was built using the ExtJS library from Sencha, providing a user-friendly environment where the information is distributed in different widgets. The web application is accessible at http://dimero.cnb.csic.es.

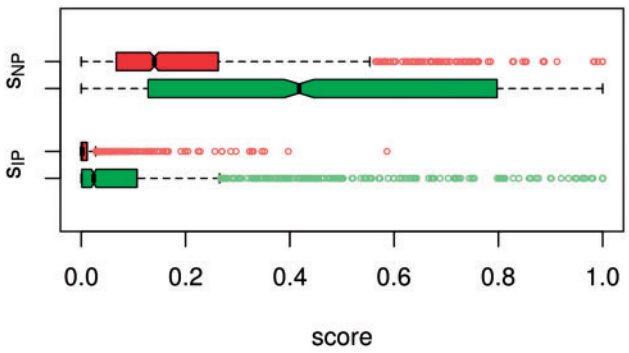

Fig. 2. Cohesiveness score distributions for PDDIs and for NDDIs. Score distribution whisker boxplots for several magnitudes are shown: interacting nodes proportion (s_{IP}) and common neighbors proportion (s_{NP}). In red, score distribution for NDDI pairs. In green, score distribution for PDDI cases

3 Results

3.1 Performance on the DDI benchmark

To assess the power of s_{IP} and s_{NP} scores (Section 2.4) inferring DDIs, we have compared their distributions on NDDI and PDDI datasets. To this end, PDDI and NDDI score distributions are represented using a boxplot schema (Fig. 2, PDDI scores in green and NDDI scores in red) for the scores s_{IP} (Fig. 2, s_{IP}) and s_{NP} (Fig. 2, s_{NP}). Figure 2 shows that, in general, score values on PDDI domain pairs tend to be larger than on the corresponding NDDI set. To confirm this observation, we computed the MWW (Supplementary Material, Section S2) test on PDDI and NDDI distributions for both scores. In both cases, the MWW test resulted in the rejection of the null hypothesis (P-value < 0.001) that score distributions for PDDI and NDDI are both equal.

Also, the performance of the scores was measured in terms of the precision, recall, AUPR, AUROC and MCC (Supplementary Material, Section S2). Supplementary Figure S2C in Supplementary Material shows that the precision-recall curves when the score threshold to consider a domain pair as interacting is decreased from 1 to 0. Table 1 shows that how the AUPR value for both scores is greater than 0.49, the AUPR value of a random classification for PDDI and NDDI benchmark. Also, the AUC value was better than the expected value of a random classifier (AUC value of 0.5). MCC measures the quality of a binary classification, in this case between interacting and non-interacting domains. All the methods achieved a MCC greater than 0 and, thus, their classification performance was better than random. The score s_{IP} achieved a maximum MCC of 0.41 when a pair of domains (δ, δ') was considered interacting if $s_{IP}(\delta, \delta') > 0.037$. For this threshold, the recall and precision were 44.2 and 83.8%, respectively; from 741 domain pairs predicted as interacting, 621 were true and 120 false. For the s_{NP} score, the best MCC was 0.43, this value was achieved when a domain pair was classified as interacting if $s_{NP}(\delta, \delta') > 0.45$. In this case, the precision and recall were 48.2 and 83.6%, respectively; from 811 domain pairs classified as interacting, 678 were true predictions and 133 false. Table 1 shows the precision and recall when the MCC achieved its maximum value.

Table 1. DDI prediction performance

Score	MCC	Recall (%)	Precision (%)	AUROC	AUPR
s_{IP}	0.41	44.2	83.8	0.71	0.75
s_{NP}	0.43	48.2	83.6	0.72	0.76
CP	0.53	67.7	81.3	*	*

Performance of the interacting nodes proportion score, neighborhood proportion score and CP classifier predicting DDIs for PDDI and NDDI datasets. First column: interacting nodes proportion (s_{IP}), common neighbors proportion (s_{NP}) and CP classifier. Note that recall, precision, MCC, AUROC and AUPR are defined in Supplementary Material (Supplementary Section S2). *These measures cannot be calculated on a binary classifier.

In terms of computational cost, computing the s_{IP} and s_{NP} scores took close to 20 and 30 min, respectively, using 256 Intel Xeon CPUs at 2 GHz.

To demonstrate whether the performance of the scores truly depends on the topology of the PPI networks, we applied the same evaluation process using random interactomes. To this end, we generated a random network for each organism in STRING employing the following strategy: networks were constructed using the original set of nodes and generating an equal number of random interactions as in the original interactome. Finally, s_{IP} and s_{NP} scores were computed using the random generated interactomes and evaluated on PDDI and NDDI domain pairs. Supplementary Figure S3 (Supplementary Material Section S4) shows that the distribution of MCC, AUROC and AUPR for the s_{IP} and s_{NP} scores inferring DDIs when this process is repeated 10^3 times. Indeed, the distributions of these measures are close to a random classification. For example, none of the tests led to an AUROC greater than 0.506 for the s_{IP} score and 0.512 for the s_{NP} score. Therefore, the topology of the PPI networks is an essential contribution for the predictive power of the proposed scores.

Finally, we used the thresholds that maximized the MCC of the s_{IP} and s_{NP} scores to set up the CP classifier (Section 2.5). Thus, a domain pair (δ, δ') was classified as interacting if $s_{IP}(\delta, \delta') > 0.037$ or $s_{NP}(\delta, \delta') > 0.45$. Table 1 shows that the performance of the CP classifier when is evaluated on PDDI and NDDI domain pairs. The CP classifier achieved a MCC of 0.53 with a recall and precision value of 67.7 and 81.3%, respectively, clearly showing that recall is increased, while maintaining a similar level of precision when compared with the s_{IP} or s_{NP} scores. Note that AUROC and AUPR cannot be calculated on a binary classifier.

3.2 Comparison with previous studies

We compared the performance of the s_{IP}, s_{NP} scores (Section 2.4) and the CP classifier (Section 2.5) inferring DDIs against two previous studies: The CSSs method proposed by Sprinzak and Margalit (2001), first, and the predictions stored in DOMINE database (Yellaboina *et al.*, 2011), second.

CSS approach was designed to find combinations of domain pairs that occur more frequently than random in particular PPI datasets. We evaluated the CSS method scoring the domain pairs of PDDI and NDDI among the different interactomes of STRING. To merge the CSS values computed on the different networks we used the same strategy as proposed in Section 2.4. Also, the same notation was adopted; thus, $c_{CSS}(\delta, \delta')$ denotes the CSS value of two domains calculated on a single interactome and $s_{CSS}(\delta, \delta')$ the average of the c_{CSS} values among different networks. Table 2 shows that the performance of the s_{CSS} score predicting DDIs on the PDDI and NDDI datasets. The proposed scores s_{IP}, s_{NP} (Sections 2.4) and the CP classifier (Section 2.5) achieved better results in terms of MCC, AUROC and AUPR than the s_{CSS} score (Tables 1 and 2).

Table 2. DDI prediction performance of CSS approach

Score	MCC	Recall (%)	Precision (%)	AUROC	AUPR
s_{CSS}	0.37	51.7	75.8	0.68	0.7

Performance of the CSSs method (Sprinzak and Margalit, 2001) predicting DDIs for the PDDI and NDDI datasets. Note that recall, precision, MCC, AUROC and AUPR are defined in Supplementary Material (Supplementary Section S2).

Table 3. Comparison with DOMINEDB data

Score	Recall (%)	Precision (%)	MCC
LC	17.2	86.2	0.25
MC	13.4	96.3	0.26
HC	12.1	96.7	0.25
CP	65.3	82.8	0.55

Performance of DOMINE predictions data and CP when predicting DDIs for the PDDI and NDDI cases included in Pfam v22 (Section 3.2). The first column defines the predicted DDIs: low confident or better DOMINE DDIs (LC), DDIs medium confident or better DOMINE DDIs (LC), high confident DOMINE DDIs (HC) and CP.

To evaluate the power of the c_{CSS}, c_{IP} and c_{NP} metrics inferring DDIs on single networks, we computed their values on the PDDI and NDDI datasets using the interactomes of STRING individually. Thus, the evaluation of the metrics was done for each network of STRING independently. Supplementary Figure S4 (Supplementary Material, Section S5) shows that the performance of the metrics predicting DDIs for the individual interactomes. The figure displays the relation between the number of interactions contained in the network and the MCC, AUROC and AUPR achieved by the metrics. In general, the performance of the three metrics improved the more interactions were contained in the networks. For all interactomes the best performance in terms of MCC, AUROC or AUPR was achieved by the c_{NP} metric except for zebrafish and saccharomyces where c_{CSS} achieved a better MCC. Finally, for all interactomes none of the metrics c_{IP}, c_{NP} or c_{CSS} achieved a better performance inferring DDIs than averaging their values over all networks; leading to the s_{IP}, s_{NP} or s_{CSS} scores (confront Supplementary Fig. S4; Tables 1 and 2).

The second comparison was between the CP classifier and DOMINE database. This database contains DDI predictions using seven different methods. The predictions are classified into low, medium and high confident, depending on the number of methods that confirm a particular DDI and the shared GO terms between the domains. Domain information in DOMINE is based on Pfam classification v22. Then, to build a proper and fare benchmark, we filtered PDDI and NDDI domain pairs selecting only those pairs for which both domains were defined in Pfam v22. The result was a set of 980 positive cases and 1115 negative cases. Then, for the resulting pairs, we checked whether they were present in DOMINE or not. If a domain pair was present in the database, then it was classified as interacting and scored in terms of low, medium or confident prediction as defined in DOMINE. On the other hand, those domain pairs that were not present in the database were classified as non-interacting.

Table 3 shows that the performance of DOMINE when the predicted interactions are defined by the low, medium or high confident domain pairs. In all cases the CP classifier achieved a better MCC

Table 4. Reduction of 'allowed' binding domain pairs in *Human* interactome

Threshold	Number of DDIs	Number of PPIs	Avg. number of DDI
All	44 750	23 758	5.4
HCP	5483	10 628	2.4
MCP	17 155	19 343	3.2
LCP	24 270	21 584	3.7

Reduction of the number of possible interacting domain pairs that can be mapped to Human interactions. The first column defines the threshold of CP used to accept a pair of interacting domains: all pairs are accepted (ALL), only HCPs, MCPs or better (MCP) and LCPs or better (LCP). The second column shows that the number of reported binding domain pairs for each threshold, while the third column indicates the number of interactions to which these reported binding domains can be mapped. Finally, the fourth column presents the average number of reported binding domains per interaction.

than DOMINE predictions and, thus, a more accurate classification. The medium and HCPs DOMINE had a better precision than the CP classifier but a lower recall value. When the thresholds (0.32 for s_{IP} and 0.85 for s_{NP}) for the CP classifier were set up to have a similar precision as the medium and HCPs of DOMINE (97%), then the recall of the CP classifier was 23% with a MCC value of 0.35, outperforming DOMINE predictions.

3.3 DDIs as mediators of PPIs

In this section, we want to analyze quantitatively the increment in the capacity to assign detailed molecular information at the DDI level to sets of PPIs, when using the method proposed in this work as opposed to not using it. The detailed derivations are presented in Supplementary Material (Supplementary Section S6), extracting here the main conclusion.

3.3.1 Human interactome analysis

Given a complete interactome, the number of all possible DDIs is large, creating ambiguities as to the precise molecular mechanisms involved. As an example, we have considered a subset of all human interactions compiled in Interactome3D database (Mosca *et al.*, 2013), formed by those interactions involving multi-domain proteins for which no molecular information is available. The resulting dataset contained 23 785 interactions. If we now calculate the number of all possible non-identical pairs of domains involved in these interactions, we obtain the number of 44 750 different domain pairs. In turn, the average number of domain pairs per interaction amounts to 5.4. We now considered CP classifier using the three thresholds defined in Supplementary Material (Supplementary Section S3): HCPs, MCPs and LCPs. Table 4 shows that how the total number of binding domains is reduced when domain pairs are accepted or rejected with the different thresholds and, thus, how the average number of 'allowed' domain combinations per interaction decreases, reducing ambiguities.

At this point, we would like to highlight two of the important limitations of this type of methodologies. First, when several copies of a particular domain are present in a protein, this type of prediction methods are not able to distinguish which of these copies is more likely to interact and all of them are scored the same. For example, in the dataset analyzed above, we found that 1263 proteins contained several copies of one domain. These proteins were involved in 5026 interactions; thus, nearly 25% of the analyzed set of interactions present ambiguities with respect to the precise domain involved in the interaction. Second, it has been estimated that

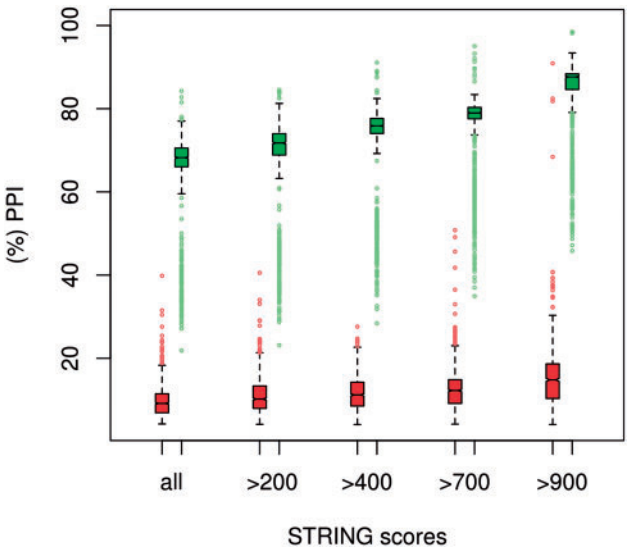

Fig. 3. Fraction of PPIs that can be complemented with predicted information on DDI. Vertical axis, distribution of the PPI fraction to which some DDIs can be mapped in the different organisms covered by STRING. Horizontal axis, score threshold used to filter PPIs in STRING. In red, 3DID data used as source of DDIs knowledge. In green, CP predictions used as source of DDIs knowledge

intrinsically disordered proteins are involved in 49% of human PPIs (Mosca *et al.*, 2012) and thus, binding regions for these interactions may not be contained in folded domains.

3.3.2 STRING interactions analysis

We were interested in analyzing what was the proportion of PPIs in STRING to which structurally solved DDIs could be mapped. To that end, we considered all DDIs between different protein chains stored in 3DID and then calculated the proportion for the different organisms in STRING. Figure 3 (red boxplots) shows these results using a boxplot diagram for different STRING scores. For all selected scores, the average proportion of PPIs to which DDIs could be mapped was <20%.

We explored the possibility of using the CP classifier (Section 2.5) to predict new DDIs and, consequently, increase the proportion of PPIs to which domain interactions could be mapped. Figure 3 (green boxplots) shows that the proportion of PPIs to which the new predicted set of DDIs can be mapped; in this case the average proportion is >60% for the selected scores. In general, DDI prediction offers the possibility to extend information derived from structural data and increase the range of PPI candidates.

3.4 Analysing interactions with DIMERO WS

In this section, we show how DIMERO WS can be used to analyse the interactions between different proteins. In this example, we have studied the interactions between the proteins of the apoptosis pathway defined in the KEGG database (Kanehisa *et al.*, 2014). Nodes in a KEGG pathway may represent several genes or proteins. In this study, we have considered all possible interactions when an edge between two nodes involves multiple protein pairs.

The apoptosis pathway contains 83 proteins and 173 interactions. We used DIMERO WS to evaluate the potential binding domains involved in the different interactions of the pathway. In this way, DIMERO WS computed the CP classifier for all possible domain binding pairs, classifying them as HPCs, MCPs and LCPs.

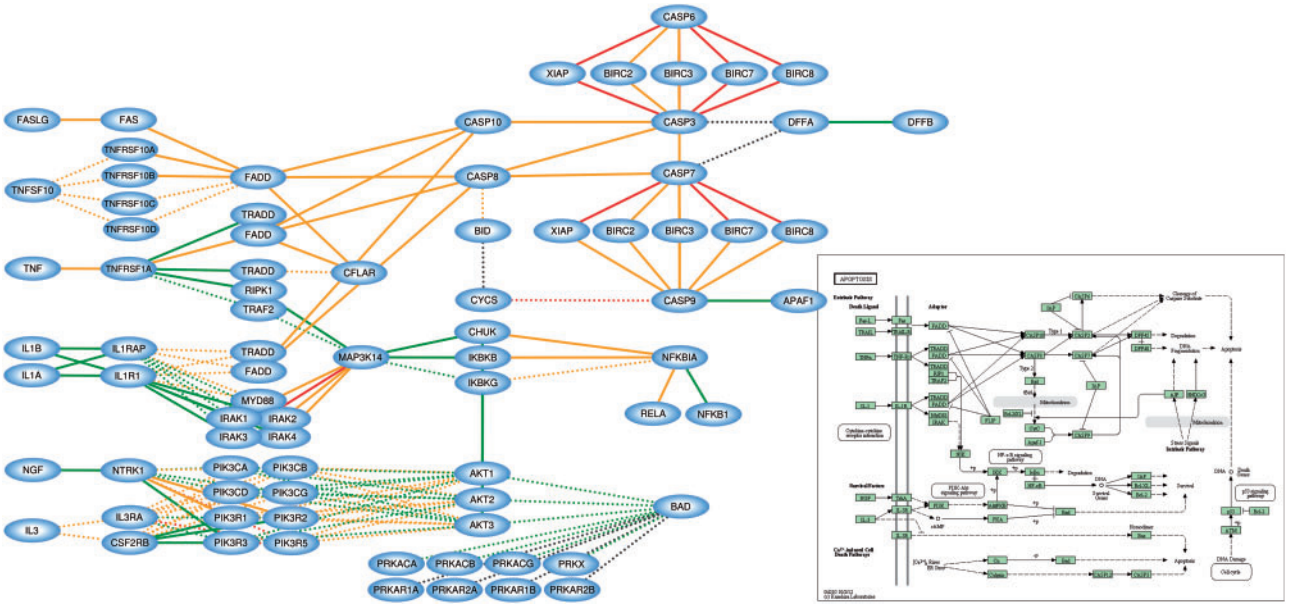

Fig. 4. Apoptosis pathway defined in KEGG database with additional information on DDIs mapped onto it. The nodes of the network represent proteins and the edges interactions between the proteins. The edge color indicates the best prediction for the CP classifier among all possible DDIs considering the proteins involved. Green lines indicate that at least one domain pair scored as a HCP, while orange lines correspond to MCPs, red lines to LCPs and grey lines mark the case for which the score was not significant. The line type of the edges indicates whether an experimentally solved structure is available for at least one of the possible domain pairs involved in the interaction (continuous lines) or not (discontinuous lines) (Color version of this figure is available at *Bioinformatics* online.)

Additionally, DIMERO WS linked the domain pairs with structural data when the structure of the interacting domains had been experimentally solved. Finally, Cytoscape (Smoot *et al.*, 2011) was used to represent the pathway at protein level and to annotate the edges with the CP predictions (Fig. 4). Protein interactions in the pathway were colored in terms of their best scored domain pair. In 62 interactions the method predicted at least one domain pair as HCP (green lines). In 76 interactions the method scored at least one of the domain pairs as MCP (orange lines). Finally, in 18 interactions the method scored at least one domain pair as LCP (red lines) and in 11 cases scores were not considered significant (grey lines). Thus, the CP classifier predicted binding domains for 90% of all possible interactions in the pathway.

Additionally, the type of line used to draw an edge referred to information regarding the availability (or lack of it) of experimentally determined structures. In this way, for 78 interactions of the apoptosis pathway, at least one of the possible binding domain pairs can be mapped to an experimentally solved structure. Interactions between the proteins of the pathway lead to 297 different binding domain pairs. When these pairs were evaluated with DIMERO WS, 29 pairs were classified as HCP, 96 as MCP and 85 as LCP. From the 297 different domain pairs, 22 could be mapped to solved structures and were involved in 78 interactions, as indicated in the previous paragraph. It is very interesting to note the behavior of the DDI prediction method on this set of experimentally solved structures. In this way, four of these pairs were classified as HCP, nine as MCP and nine as LCP. Thus, none of these pairs were considered to be below the significance level. An additional observation is that these 22 domain pairs can be classified in 7 homo-interactions and 15 hetero interactions, and that hetero-interactions were scored higher (four as HCP, eight as MCP and three as LCP) than homo-interactions (one as MCP and six as LCP). This different behavior is related to the fact that the neighborhood proportion measure cannot be used to compare a set of nodes with itself and, consequently, only the interacting node proportion can be used in these cases, while both scores (s_{IP} and s_{NP}) are used for hetero-interactions.

Finally, we analyzed some of the domain pairs classified as HCP and that were not mapped to solved structures. In this way, we found some interesting results involving kinase domains and phosphorylation events. For example, in the interaction between the protein kinase B family members (AKT1, AKT2 and AKT3) and the IKK complex (CHUK, IKBKB and IKBKG), the kinase-NEMO DDI was classified as HCP. Although, we cannot ensure this is a real interaction it is known that the NEMO domain of the IKBKG subunit is phosphorylated in other signaling processes (Palkowitsch *et al.*, 2008; Wu *et al.*, 2006). A similar result was found between the protein kinase B family and the BAD protein; but in this case there is experimental evidence supporting that AKT1 protein kinase indeed phosphorylates the Bcl-2_BAD domain of the BAD protein (Koh *et al.*, 2000), although no atomic structure of this complex has yet been obtained.

4 Conclusions

This work presents a novel application of NC property to infer interactions between protein domains. The approach is suitable for large scale prediction of detailed molecular information to be further considered in the understanding of interactomes. In this way, we are able to assign a significant value to the interaction of pairs of protein domains involved in a given interaction, filtering the less probable bindings. With this approach we are able to significantly extend current binding information coming from sets of experimentally determined structures, outperforming previously proposed methods.

Finally, a user-friendly web platform is available at http://dimero.cnb.csic.es, allowing the use to query about DDIs at the same that it provides direct access to the data used to calculate the interaction scores.

Funding

This work was supported by the Instituto de Salud Carlos III, project number PT13/0001/0009 funding the Spanish National Institute of Bioinformatics,

the Spanish Ministry of Economy and Competitiveness through grants AIC-A-2011-0638 and BIO2013-44647-R. Sorzano is recipient of a Ramón y Cajal fellowship.

Conflict of Interest: none declared.

References

Apic,G. *et al.* (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.

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

Berggard,T. *et al.* (2007) Methods for the detection and analysis of protein-protein interactions. *Proteomics*, **7**, 2833–2842.

Blohm,P. *et al.* (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.*, **42**, D396–D400.

Chen,C. *et al.* (2011) Inferring domain-domain interactions using an extended parsimony model. In *Systems Biology (ISB), 2011 IEEE International Conference on IEEE*, pp. 374–378.

Deng,M. *et al.* (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**, 1540–1548.

Desjarlais,J.R. and Berg,J.M. (1992) Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 7345–7349.

Ekman,D. *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 4372–4376.

Gomez,S.M. *et al.* (2003) Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, **19**, 1875–1881.

Gupta,S. *et al.* (2010) Unraveling the conundrum of seemingly discordant protein-protein interaction datasets. In: *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE Engineering in Medicine and Biology Society, pp. 783–786.

Halperin,I. *et al.* (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443.

Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.

Hulovatyy,Y. *et al.* (2014) Revealing missing parts of the interactome via link prediction. *PLoS One*, **9**, e90073.

Itzhaki,Z. *et al.* (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol.*, **7**, R125.

Johnson,L.S. *et al.* (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

Jothi,R. *et al.* (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J. Mol. Biol.*, **362**, 861–875.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

Kann,M.G. *et al.* (2007) Predicting protein domain interactions from coevolution of conserved regions. *Proteins*, **67**, 811–820.

Koh,H. *et al.* (2000) Inhibition of Akt and its anti-apoptotic activities by tumor necrosis factor-induced protein kinase C-related kinase 2 (PRK2) cleavage. *J. Biol. Chem.*, **275**, 34451–34458.

Li,S. *et al.* (2011) Network target for screening synergistic drug combinations with application to traditional Chinese medicine. *BMC Syst. Biol.*, **5** (Suppl. 1), S10.

Liu,M. *et al.* (2009) Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks. *Bioinformatics*, **25**, 2492–2499.

Liu,Y. *et al.* (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, **21**, 3279–3285.

Mosca,R. *et al.* (2012) The role of structural disorder in the rewiring of protein interactions through evolution. *Mol. Cell. Proteomics*, **11**, M111 014969.

Mosca,R. *et al.* (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.

Mosca,R. *et al.* (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **42**, D374–D379.

Moya-Garcia,A.A. and Ranea,J.A. (2013) Insights into polypharmacology from drug-domain associations. *Bioinformatics*, **29**, 1934–1937.

Ng,S.K. *et al.* (2003) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, **19**, 923–929.

Nye,T.M. *et al.* (2005) Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, **21**, 993–1001.

Palkowitsch,L. *et al.* (2008) Phosphorylation of serine 68 in the IkappaB kinase (IKK)-binding domain of NEMO interferes with the structure of the IKK complex and tumor necrosis factor-alpha-induced NF-kappaB activity. *J. Biol. Chem.*, **283**, 76–86.

Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Pazos,F. and Valencia,A. (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J.*, **27**, 2648–2655.

Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Riley,R. *et al.* (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**, R89.

Rose,P.W. *et al.* (2013) The RCSB protein data bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.

Segura,J. *et al.* (2011) Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinformatics*, **12**, 352.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.

Watson,J.D. *et al.* (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Wu,Z.H. *et al.* (2006) Molecular linkage between the kinase ATM and NF-kappaB signaling in response to genotoxic stimuli. *Science*, **311**, 1141–1146.

Yellaboina,S. *et al.* (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.*, **39**, D730–D735.