Bioinformatics, 2017, 1–8
doi: 10.1093/bioinformatics/btx044
Advance Access Publication Date: 25 January 2017
Original Paper

OXFORD

Structural bioinformatics

ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model

Castrense Savojardo[1,2,†], Piero Fariselli[3,†], Pier Luigi Martelli[1,2,*] and Rita Casadio[1,2]

[1]Biocomputing Group, Department of Biological, Geological and Environmental Sciences (BiGeA), University of Bologna, Bologna 40126, Italy, [2]CIG, Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna 40127, Italy and [3]Department of Comparative Biomedicine and Food Science (BCA), University of Padova, Padova 35020, Italy

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Anna Tramontano

Received on July 27, 2016; revised on December 20, 2016; editorial decision on January 17, 2017; accepted on January 20, 2017

Abstract

Motivation: The identification of protein–protein interaction (PPI) sites is an important step towards the characterization of protein functional integration in the cell complexity. Experimental methods are costly and time-consuming and computational tools for predicting PPI sites can fill the gaps of PPI present knowledge.
Results: We present ISPRED4, an improved structure-based predictor of PPI sites on unbound monomer surfaces. ISPRED4 relies on machine-learning methods and it incorporates features extracted from protein sequence and structure. Cross-validation experiments are carried out on a new dataset that includes 151 high-resolution protein complexes and indicate that ISPRED4 achieves a per-residue Matthew Correlation Coefficient of 0.48 and an overall accuracy of 0.85. Benchmarking results show that ISPRED4 is one of the top-performing PPI site predictors developed so far.
Contact: gigi@biocomp.unibo.it
Availability and Implementation: ISPRED4 and datasets used in this study are available at http://ispred4.biocomp.unibo.it.

1 Introduction

In the context of the cell macromolecular crowding, protein–protein interactions (PPIs) are the major drivers of all the biological processes. The knowledge of the specific protein surface residues involved in PPI is then crucial for characterizing the role of each protein in complex metabolic and signaling pathways and in sustaining cell physiology. Furthermore, PPI is an invaluable source of information for elucidating complex disease networks, for improving protein-docking studies and for drug design (Sudha *et al.*, 2014).

Experimental determination of how proteins interact is expensive in terms of both time and costs. Theoretical methods can complement PPI experimental knowledge [for recent reviews see Xue *et al.* 2015; Esmaielbeiki *et al.* (2016), and references therein]. Routinely, PPI computational methods can (i) identify pairs of interacting proteins in complex interaction networks; (ii) identify pairwise residue contacts between two query proteins known to interact; (iii) predict on a target protein surface residues in interaction with one or more (not necessarily known) binding partner/s (predict PP interface/s). For this last type of prediction, several computational tools are available. Depending on the exploited information, they group into three different categories: template-, sequence-, and structure-based predictors (Aumentado-Armstrong *et al.*, 2015; Esmaielbeiki *et al.*, 2016)

Template-based approaches (Jordan *et al.*, 2012; Xue *et al.*, 2011; Zhang *et al.*, 2011) perform the prediction searching for similar structures (sequence homologues or structural neighbours).

Interface residues are then transferred from the template to the query after structural alignment. These methods typically achieve good performances (Esmaielbeiki *et al.*, 2016), but require the availability of reliable structural templates for a given query protein.

Sequence-based predictors (Chen and Li, 2010; Gallet *et al.*, 2000; Ofran and Rost, 2007; Res *et al.*, 2005; Yan *et al.*, 2004) extract descriptors derived only from the protein sequence. These descriptors, or features, are subsequently processed using machine-learning algorithms. Typical features include: evolutionary information in the form of sequence profiles or position-specific scoring matrices; residue conservation computed on multiple sequence alignments (MSAs); residue interface propensity; residue physico-chemical properties (Aumentado-Armstrong *et al.*, 2015). Although these approaches can potentially be applied to all known protein sequences, their prediction performance is limited by the amount of information that can be derived from the protein sequence alone (Esmaielbeiki *et al.*, 2016).

Structure-based predictors rely on protein sequence and structure as sources of information. They process different types of features, including different metrics of residue solvent exposure (e.g. relative accessibility, protrusion and depth indexes), surface curvature, electrostatic potentials, atomic B-factor and/or secondary structure (Bradford and Westhead, 2005; Chen and Zhou, 2005; Fariselli *et al.*, 2002; Jones and Thornton, 1997; Koike and Takagi, 2004; Li *et al.*, 2007, 2012; Liu *et al.*, 2009; Porollo and Meller, 2007; Savojardo *et al.*, 2012; Sikić *et al.*, 2009; Dong *et al.*, 2014). Methods in this category, although limited to proteins with known three-dimensional structure, achieve higher performance when compared to sequence-based methods (Aumentado-Armstrong *et al.*, 2015; Esmaielbeiki *et al.*, 2016).

Residue co-evolutionary analysis of protein complexes indicates that PPI prediction improves when correlated protein variations are taken into account (Burger and van Nimwegen, 2008; Pazos *et al.*, 1997; Weigt *et al.*, 2009). Recent computational advances in extracting protein coevolution information with statistical global models largely improve the prediction of protein 3D structure (Ekeberg *et al.*, 2013; Jones *et al.*, 2012; Kamisetty *et al.*, 2013; Marks *et al.*, 2011; Morcos *et al.*, 2011) and PPIs between given pairs of proteins (Hopf *et al.*, 2014; Ovchinnikov *et al.*, 2014). Global models of co-evolution extract co-evolution indexes from MSAs by adopting different approaches that include sparse inverse covariance estimation (Jones *et al.*, 2012), direct coupling analysis (Marks *et al.*, 2011; Morcos *et al.*, 2011) and pseudo-likelihood-based approaches (Ekeberg *et al.*, 2013; Hopf *et al.*, 2014; Kamisetty *et al.*, 2013; Ovchinnikov *et al.*, 2014). In the context of PPI, all methods based on residue co-evolution are partner-specific, and they require the knowledge of the interacting protein pairs (Burger and van Nimwegen, 2008; Hopf *et al.*, 2014; Ovchinnikov *et al.*, 2014; Weigt *et al.*, 2009).

Routinely, sequence- and structure-based approaches process input features with machine-learning algorithms. These comprise Neural Networks (Chen and Zhou, 2005; Fariselli *et al.*, 2002; Porollo and Meller, 2007), Support Vector Machines (SVMs) (Bradford and Westhead, 2005; Koike and Takagi, 2004), Random Forests (Li *et al.*, 2012; Sikić *et al.*, 2009), Hidden Markov Models (Liu *et al.*, 2009; Savojardo *et al.*, 2012) and Conditional Random Fields (Dong *et al.*, 2014; Li *et al.*, 2007).

We present ISPRED4, an improved structure-based predictor of PPI sites on unbound monomer surfaces in the absence of interaction partners. Building on top of our previously released ISPRED3 (Savojardo *et al.*, 2012), the proposed method exploits a richer set of 46 features derived from both sequence and structure, including local and global co-evolutionary analyses. Features are elaborated to predict interface residues on an input protein surface with a hybrid procedure. The method is based on the combination of SVMs and grammar-restrained conditional random fields (GRHCRF) (Fariselli *et al.*, 2009).

When benchmarked towards other available methods for the same task and on the same dataset, ISPRED4 scores better than the other state-of the-art approaches.

2 Materials and methods

2.1 Datasets

The dataset, adopted in this study to train/test ISPRED4, derives from the Docking Benchmark dataset version 5 (DBv5) (Vreven *et al.*, 2015). The original DBv5 dataset contains 230 protein complexes whose key feature is the availability of both bound and unbound structures of the interacting proteins. Unbound monomers are distributed over 420 different PDB entries. Starting from this initial set of entries, we retained only the subset of protein complexes that met the following criteria:

- Both bound and corresponding unbound structures were obtained by means of X-Ray crystallography.
- Interfaces calculated from the bound structure could be successfully mapped to unbound structures unambiguously.

After this filtering procedure, we retained 151 bound protein complexes. For each complex, the corresponding unbound structures were collected, leading to 314 different monomer chains. In what follows, this dataset is referred to as DBv5Sel and it is available at http://ispred4.biocomp.unibo.it.

Two alternative definitions of PPI sites on a protein surface are routinely adopted:

1. Any surface residue whose distance (all-atom or Cα–Cα) from one residue in the partner molecule is below a pre-defined threshold (in the range of 5–8 Å).
2. Any surface residue undergoing a reduction of its Accessible Surface Area (ASA) upon complex formation.

Several studies showed that the choice of interface definition has only a limited impact on the performance of a predictor (De Vries and Bonvin, 2008; Savojardo *et al.*, 2012).

In this paper, we adopt the second definition. In particular, surface residues are those with Relative Solvent Accessibility (RSA) $\geq 20\%$ in the unbound monomer and interacting residues are those undergoing a decrease of the Absolute Solvent Accessibility (ASA) ≥ 1 Å^2 in the corresponding complex.

Per-residue ASA values were computed with the DSSP program (Kabsch and Sander, 1983); RSA values are then obtained by normalizing with respect to the residue-specific maximal accessibility values as previously described (Rost and Sander, 1993). The ASA differences were computed comparing each unbound chain with the corresponding chain taken from the bound complex. When the same monomer is part of different complexes, the union of the interacting residues was considered.

The performance benchmark was performed on a blind test set derived from past CAPRI experiments. In particular, we considered the targets released in CAPRI rounds from 1 to 29 (in total 67 targets). We filtered out target chains sharing sequence identity $\geq 30\%$ with respect to any sequence of our training dataset. Furthermore, also inter-target redundancy was reduced to 30% sequence identity. The final CAPRI-Blind dataset comprises 22 different bound

structures including 29 protein chains. The above-described procedure was adopted to extract surface residues and PPI sites. Summary statistics of both DBv5Sel and CAPRI-Blind datasets are reported in Table 1.

2.2 Input feature encoding

Several feature descriptors were extracted and adopted to perform the interface/non-interface classification task. The complete feature set consists of 10 different groups of descriptors encoded in a 46-dimensional real vector for each input surface residue. Table 2 reports a summary of the different descriptor sets used in this study.

2.2.1 Evolutionary information

Evolutionary information for each position of the primary chain sequence was extracted in the form of a sequence profile. For a given protein chain sequence of length L, the PSI-BLAST (Altschul *et al.*, 1997) program was used to search the Uniprot Reference Cluster 90 database (Suzek *et al.*, 2015) for similar sequences and the corresponding 20-by-L profile matrix built during the search was extracted as additional PSI-BLAST output (option -ascii_pssm). The matrix computed by PSI-BLAST already incorporates a data-dependent pseudocount model (Altschul *et al.*, 1997; Tatusov *et al.*, 1994). Finally, for each surface residue i, a 20-dimensional vector v_i was computed by averaging sequence profile entries over the surface structural context of the residue i, i.e.

$$v_i = \frac{1}{|C(i)|}\sum_{k\in C(i)} P_k \quad (1)$$

where $C(i)$ is the surface context of the residue i, defined as the set of surface residues whose Cα–Cα distance from i is below 12Å and P_k is the 20-dimensional profile vector of residue k.

2.2.2 Residue conservation

Given the sequence profile derived from PSI-BLAST, a conservation score c_i was computed for each surface residue position i using the normalized Shannon's entropy as previously described by Sander and Schneider (1991):

$$c_i = -\frac{1}{\log K}\sum_{j=1}^{K} P_{ij} \times \log P_{ij} \quad (2)$$

where $K = 20$ (i.e. the 20 different residue types) and P_{ij} is the frequency of residue type j at position i extracted from the sequence profile.

Table 1. Summary statistics computed on the DBv5Sel and CAPRI-Blind datasets

	DBv5Sel	CAPRI-Blind
Number of bound structures	151	22
Number of unbound chains	314	29
Total number of residues	67,235	6,369
Total number of surface residues	39,046	3,613
Total number of PPI sites	8,469	868

Table 2. Sets of descriptors adopted in this study to encode surface residues

Descriptor	Program(s) used	Number of features
Sequence profile (PROF)	PSI-BLAST	20
Conservation score (CONS)	PSI-BLAST	1
Interface propensity (PROP)	In-house script	1
Residue properties (RPROP)	In-house script	10
Mutual Information (MI)	HHBlits	2
PSICOV	HHBlits	2
Depth indexes (DPX)	PSAIA	3
Protrusion indexes (CX)	PSAIA	4
Secondary structure (SS)	DSSP	3
Average B-Factor (BFACT)	In-house script	1
RSA difference (dRSA)	DSSP, SABLE	1

2.2.3 Residue interface propensity

The propensity p_k of each residue type to be in interaction sites was scored using the following log-ratio formula:

$$p_k = \log\frac{f_I(k)}{f_S(k)} \quad (3)$$

where $f_I(k)$ is the frequency of residue of type k in interaction sites and $f_S(k)$ is the frequency of residue type k in the surface. Both frequencies and propensities scores were computed on the training set and kept fixed when encoding the testing set, for each cross-validation iteration.

2.2.4 Residue physico-chemical properties

The 10 orthogonal properties introduced by Kidera *et al.* (1985) were used to represent the physico-chemical nature of each residue. As described by the authors, the 10 properties were derived with multivariate statistical analysis of a set of 188 different physical properties of naturally occurring amino acids. This allows representing each residue type with a small number of parameters that contain a sufficient amount of information (Kidera *et al.*, 1985). Each residue in the surface was represented according to its type with a 10-dimensional real vector.

2.2.5 Residue co-evolution scores

Different methods are available to extract residue co-evolutionary indexes starting from a MSA. In this work, we adopted sparse inverse covariance estimation as implemented in the PSICOV method (Jones *et al.*, 2012) as well as Mutual Information (MI).

MSAs were generated running the HHBlits aligner (Remmert *et al.*, 2011) against the UniprotKB database (we used the clustered Uniprot release 2016/02 at 20% sequence identity available at the HHSuite FTP site). Hence, PSICOV and MI co-evolutionary scores were computed from the alignments.

MI is defined as:

$$\mathrm{MI}_{ij} = \sum_{a=1}^{K}\sum_{b=1}^{K} f_{ij}(a,b) \times \log\frac{f_{ij}(a,b)}{f_i(a)\times f_j(b)} \quad (4)$$

where $K = 20$ (i.e. the 20 different residue types), $f_i(a)$ and $f_j(b)$ are the frequencies of residue type a and b at positions i and j in the MSA, respectively, and $f_{ij}(a,b)$ is the frequency of residue pair ab at positions ij.

Two different descriptors were computed for each surface residue, namely the average co-evolution of the residue with respect to (i) its surface context (co_i^c) and (ii) the rest of surface residues (co_i^r). More formally, for a given surface residue i:

$$co_i^c = \frac{1}{|C(i)|}\times\sum_{k\in C(i)} s(i,k) \quad (5)$$

$$co_i^r = \frac{1}{l_s - |C(i)|} \sum_{k \notin C(i)} s(i,k) \quad (6)$$

where *C(i)* is the surface context (see Section 2.2.1) of the residue *i*, l_s is the total number of surface residues and $s(i,k)$ is either the PSICOV or MI co-evolutionary score between residues *i* and *k*.

The rationale behind this approach is to characterize the extent of co-evolutionary correlations of a given surface residue on both its structural proximity and the rest of the protein surface. Similar co-evolutionary proximity-based indexes, based on MI, were also adopted in literature for catalytic site identification (Aguilar *et al.*, 2012; Buslje *et al.*, 2010).

2.2.6 Geometrical descriptors

Protrusion (Pintar *et al.*, 2002) and depth (Pintar *et al.*, 2003) indexes were computed for each surface residue using the Protein Structure And Interaction Analyzer (PSAIA) toolkit (Mihel *et al.*, 2008). In particular, four descriptors were used to account for protrusion (total average, side-chain average, minimum and maximum) and three descriptors were used for depth index (total average, side-chain average and maximum). Minimum depth was not considered since it is always zero for surface residues.

Furthermore, secondary structure assignment were obtained using the DSSP program (Kabsch and Sander, 1983) and mapped to three different classes: helix (H, G, I), strand (E, B) and coil (T, S). Finally, for each surface residue, we computed three descriptors, $f_H(i)$, $f_E(i)$ and $f_C(i)$, representing the frequency of helical, strand and coil residues in the surface context of residue *i*, respectively,

2.2.7 Residue average B-factor

The average B-factor b_i for the surface residue *i* was computed by averaging over individual B-factors of its atoms, as reported in the PDB file.

2.2.8 Difference between predicted and real RSA

The idea of using predicted solvent accessibility to discover PPI sites is originally due to Porollo and Meller (2007). In their work, authors showed that RSA predictions starting from protein primary sequences, obtained using their SABLE predictor (Adamczak *et al.*, 2004), tend to be biased towards the level of solvent exposure observed in protein complexes. As a consequence, the difference between the predicted and the real (i.e. taken from the unbound structure) RSA values (here referred to as dRSA) could be used as fingerprint of PPI sites (Porollo and Meller, 2007). Subsequently, this feature was also included in a previous version of ISPRED (Savojardo *et al.*, 2012). To capture this trend, in this work the dRSA descriptor was computed as follows:

$$dRSA_i = \frac{1}{|C(i)|} \sum_{k \in C(i)} rsa_p(k) - rsa_o(k) \quad (7)$$

where *C(i)* is the surface context (see above) of the residue *i*, $rsa_p(k)$ and $rsa_o(k)$ are, respectively, predicted and observed RSA value for the residue *k*. RSA predictions from sequence were obtained using the SABLE predictor (Adamczak *et al.*, 2004).

2.3 The ISPRED4 prediction algorithm

2.3.1 The SVM classifier

In this work, we adopt SVMs to predict whether a residue located at the surface of a query protein chain is part of a PPI interface or not. A SVM model was trained on protein chains in the DBv5Sel dataset whose surface residues were represented with the 46-dimensional feature vectors described in the previous section. The SVM model was trained/tested using the LIBSVM implementation (Chang *et al.*, 2011) with a radial basis function (RBF) kernel. The penalty factor C (which controls the trade-off between margin size and training error) and the RBF γ hyper-parameters were both optimized using a grid-search procedure. The optimal C and γ were selected in the sets $[2^{-5}, 2^{-3}, \ldots, 2^{13}]$ and $[2^{-11}, 2^{-9}, \ldots, 2^{5}]$, respectively.

To assess the contribution of the different features in predicting protein–protein interfaces, we also trained/tested several SVM models using only specific subsets of the 46 input features. In all cases, the same grid-search procedure was applied to optimize hyper-parameters C and γ.

2.3.2 Grammar-based correction

Standard classification approaches like SVMs treat residues as independent from each other's, classifying them as interacting (label I) or not interacting (label N). As a consequence, possible correlations between neighbouring residues on the protein surface sequence are neglected. On the contrary, sequence-labeling methods like CRFs or HMMs are able to capture correlations between neighbouring labels and to introduce global constraints on the labeling of the whole protein surface sequence (Li *et al.*, 2007; Liu *et al.*, 2009; Savojardo *et al.*, 2012).

In this paper, we combine the SVM classifier described in Section 2.3.1, with a Grammatical-Restrained Hidden Conditional Random Field (GRHCRF) model (Fariselli *et al.*, 2009). GRHCRF is a discriminative framework that has been developed to address bio-sequence labeling tasks (Indio *et al.*, 2013; Savojardo *et al.*, 2011, 2013). In analogy with HMMs, GRHCRFs can be represented with a Finite State Automaton (FSA). The structure of the FSA casts the grammar modeling the constraints of the specific labeling problem at hand. The main difference with respect to standard CRFs (Lafferty *et al.*, 2001) is the explicit inclusion of hidden variables represented by the states of the FSA.

More formally, the labelling task is to map an observation sequence $\mathbf{x} = (x_1, x_2, \ldots, x_L)$ (x_i is a feature-vector) into a label sequence $\mathbf{y} = (y_1, y_2, \ldots, y_L)$, where each y_i belongs to a finite alphabet **Y**. The labelling process is mediated by a layer of 'hidden' states $\mathbf{h} = (h_1, h_2, \ldots, h_L)$. Each state h_i is member of a finite set **H**. A one-to-many mapping is established between labels and states: each hidden state h belongs to the subset $\boldsymbol{H}_y$ ($\boldsymbol{H}_y \subset \mathbf{H}$) of states associated to a given label y. Furthermore, the set of allowed hidden-state transitions is controlled by means of a user-defined FSA. In other words, only state paths allowed by the FSA are taken into consideration by the model (Fariselli *et al.*, 2009).

A GRHCRF is a model of the conditional probability distribution of label sequences given observations, defined as

$$p(y|x) = \frac{Z(\mathbf{x}, \mathbf{y})}{Z(\mathbf{x})} \quad (8)$$

where $Z(x,y)$ and $Z(x)$ are *partition functions*, defined as follows:

$$Z(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{h}} \prod_{j=1}^{L} \Psi(h_j, h_{j-1}, y_j, \mathbf{x}) \quad (9)$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \sum_{\mathbf{h}} \prod_{j=1}^{L} \Psi(h_j, h_{j-1}, y_j, \mathbf{x}) \quad (10)$$

The partition function $Z(\boldsymbol{x}, \boldsymbol{y})$ is obtained by keeping fixed the label sequence and summing over all possible corresponding hidden-state sequences; $Z(\mathbf{x})$ is computed by summing over all possible label and state sequences (Fariselli *et al.*, 2009). $\Psi(h_j, h_{j-1}, y_j, \mathbf{x})$ are called

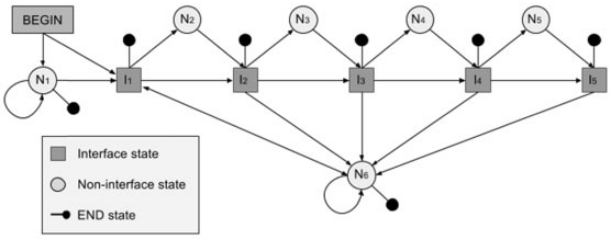

Fig. 1. The GRHCRF model adopted by ISPRED4

potential functions and score the transition of the model from state h_{j-1} to state h_j at time j, assuming the usual log-linear form (Fariselli *et al.*, 2009; Lafferty *et al.*, 2001):

$$\Psi(h_j, h_{j-1}, y_j, \mathbf{x}) = \exp\left(\left\langle \boldsymbol{\omega}_{h_j}, x_j \right\rangle + \lambda_{h_j,h_{j-1}}\right) \cdot 1_{\{h_{j-1}\to h_j \in FSA\}} \quad (11)$$

where the indicator function $1_{\{h_{j-1}\to h_j \in FSA\}}$ (which is 1 only when the condition in the brackets is met, otherwise it takes the value of 0), is used to enforce the grammatical constraints. The notation $\langle .,. \rangle$ defines the standard dot product, and $\boldsymbol{\omega}_{h_j}$ and $\lambda_{h_j,h_{j-1}}$ are model parameters, scoring, respectively, the compatibility between the state and the observation at position j and the transition from state h_{j-1} to state h_j. These parameters are optimized by maximizing the log-likelihood over training data. Once the model is trained, an input observation sequence is labelled efficiently using posterior Viterbi decoding (Fariselli *et al.*, 2009).

For the specific application described in this paper, the observations are defined as sequences of two-dimensional feature-vectors. The feature-vector components are the probability of being, or not being in an interaction site as computed by the SVM (using the *-b* option of LIBSVM).

The grammar for the GRHCRF model used here is depicted in Figure 1. Two kinds of states are defined: interaction states, labelled as Ix, and non-interaction states, labelled as Nx. The grammar models the proximity relationships of interface residues along the sequence. GRHCRF introduces correlations among the different SVM predictions by filtering out single isolated spots and by reinforcing locally coherent predictions.

2.4 Scoring measures

2.4.1 Performance evaluation at the residue level

Standard scoring measures were adopted to score the method at the level of binary residue classification. In what follows, let TP, FP, TN and FN be true positives, false positives, true negatives and false negatives, respectively (we consider the interaction site label as the positive class). The following scoring measures were adopted to score interface residue predictions:

Recall (true positive rate) of the positive class [Rec(I)], defined as:

$$Rec(I) = \frac{TP}{TP + FN} \quad (12)$$

Precision of the positive class [Pre(I)], defined as:

$$Pre(I) = \frac{TP}{TP + FP} \quad (13)$$

The F1-score of the positive class [F1(I)], defined as:

$$F1(I) = \frac{2 \times Rec(I) \times Pre(I)}{Rec(I) + Pre(I)} \quad (14)$$

The classification accuracy [Q_2], defined as:

$$Q_2 = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

The Matthews Correlation Coefficient [MCC], defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}. \quad (16)$$

2.4.2 Performance evaluation at the patch level

In order to better characterize predicted PPI sites, we also evaluated the performance in identifying interaction patches, i.e. clusters of neighbouring PPI sites identified on the protein surface.

More formally, let I^T and I^P be, respectively, the sets of true and predicted PPI sites in the protein. In each set, we built a graph by connecting pairs of residues whose Cα–Cα distance is below 12Å. We defined the interaction patches as the connected components of the graphs.

In analogy with the Segment OVerlap (SOV) measure, previously introduced to score secondary structure prediction at the segment level (Zemla *et al.*, 1999), we defined the Patch OVerlap (POV) index to measure the overlap between true and predicted interaction patches:

$$POV = \frac{1}{|I^T|} \sum_{\{(p,q):\ |p\cap q|\neq\varnothing\}} \left[\frac{|p \cap q| + \delta(p,q)}{|p| + |q| - |p \cap q|} \times |p|\right] \quad (17)$$

where $|p|$ indicate the cardinality of the patch and the quantity $\delta(p, q)$ is defined as:

$$\delta(p,q) = \min\{|p|/2, |q|/2, |p \cap q|, |p| + |q| - 2 \times |p \cap q|\}. \quad (18)$$

2.5 Cross-validation

Primary sequences were extracted for each unbound chain in the DBv5Sel dataset from the corresponding PDB entry. Sequences were pairwise aligned using the blastp program (Altschul *et al.*, 1997) and clusters of similar sequences were computed. Two chain sequences were put into the same cluster if blastp detected a pairwise sequence identity $\geq 25\%$ (no coverage threshold was adopted in order to also consider local sequence similarity). A 10-fold cross-validation split was then generated by randomly assigning each cluster to one of the different cross-validation subsets. By adopting this procedure, we ensured that even local sequence identity was confined into the same cross-validation subset, avoiding any possible training/testing bias.

3 Results

3.1 PPI site prediction with different descriptors

The discriminative power of the different descriptors used in this study is evaluated by training and testing our method on different combinations of descriptor sets. Table 3 lists the 10-fold cross-validation results on the DBv5Sel dataset for each descriptor set.

At this stage, the goal is to evaluate the effect of the different input features. Results are listed without applying the grammar correction described in Section 2.3.2, directly using the SVM classification outputs (see Section 2.3 for details),

The baseline predictor is defined as the one taking in input only the 20 sequence profile descriptors. The first two rows of Table 3 show that averaging of the sequence profile over the surface context (see Section 2.2.1) improves the performance. Indeed, when the

Table 3. Performance evaluation of different methods and feature combinations on the DBv5Sel dataset in 10-fold cross-validation

Input	MCC	Q_2	Rec(I)	Pre(I)	F1(I)
PROF (no context)*	0.27	0.80	0.16	0.69	0.26
PROF	0.33	0.82	0.24	0.70	0.36
PROF+PROP	0.33	0.82	0.24	0.71	0.36
PROF+CONS	0.34	0.82	0.24	0.71	0.36
PROF+BFACT	0.34	0.82	0.26	0.69	0.38
PROF+DPX	0.35	0.82	0.25	0.71	0.37
PROF+CX	0.35	0.82	0.26	0.71	0.38
PROF+MI	0.35	0.82	0.26	0.70	0.38
PROF+PSICOV	0.36	0.82	0.26	0.71	0.38
PROF+dRSA	0.36	0.82	0.26	0.71	0.38
PROF+RPROP	0.38	0.82	0.32	0.67	0.43
PROF+SS	0.38	0.83	0.30	0.71	0.42
SEQUENCE**	0.40	0.83	0.33	0.72	0.45
STRUCTURE***	0.43	0.83	0.36	0.72	0.48
ISPRED4 (SVM only)	0.46	0.84	0.41	0.72	0.52
ISPRED4 (SVM+GRHCRF)	0.48	0.84	0.39	0.78	0.52

Note: *PROF (no context) = the individual profile vector of each residue is used instead of averaging over the surface context.

**SEQUENCE = the subset of 34 sequence-based features comprising the baseline PROF as well as PROP, CONS, PSICOV and RPROP.

***STRUCTURE = the subset of 32 structure-based features comprising the baseline PROF as well as BFACT, DPX, CX, dRSA and SS.

individual profile vector of each surface residue is used (first row) instead of the average (second row), the MCC drops from 0.33 to 0.27. For this reason, in the following experiments, we adopted the average profile as the baseline feature. All other classifiers were hence generated adding to this baseline predictor one descriptor set at a time.

As reported in Table 3, all descriptors positively contribute to the prediction performance. With the exception of the residue interface propensity, all descriptors improve the MCC when compared with the baseline. The most informative descriptors are secondary structure and residue physico-chemical properties, both improving MCC by 5 points with respect to the baseline (from 0.33 to 0.38). The two different coevolution scores (MI and PSICOV) result in very similar performances, with a slight increase of MCC, when PSICOV is adopted. For this reason, PSICOV score is used in all the subsequent experiments. As expected, structure-based descriptors (secondary structure, dRSA, depth and protrusion indexes, B-factor) are on average more informative than sequence-based ones (residue properties, co-evolution and conservation scores, propensity).

To better highlight this trend, we also trained and tested two classifiers adding to the baseline, respectively, the subsets of sequence (i.e. 34 features comprising PROF, PROP, CONS, PSICOV and RPROP) and structural (i.e. 32 features comprising PROF, BFACT, DPX, CX, dRSA and SS) descriptors. Performances of these two classifiers are reported in Table 3, in rows SEQUENCE and STRUCTURE, respectively. As expected, the structural features perform better than the sequence ones. The combination of sequence and structure descriptors further improves over top-scoring individual feature sets.

3.2 Scoring ISPRED4 performances on the full descriptor set

In the last two rows of Table 3, we also list the cross-validation performances of ISPRED4 trained/tested using the full descriptor set. For the sake of comparison, we report scoring measures obtained with and without the GRHCRF-based grammar correction described as in Section 2.3.2.

When all the features are included, the prediction performance improves up to 0.46 of MCC (compare values in Table 3). This suggests that the contribution of each descriptor is non-redundant with respect to the others.

Furthermore, it is evident from the last row (Table 3) that the grammar correction further improves the prediction performance. In particular, precision increases up to 0.78, at the cost of only a slight decrease in recall, leading to an overall MCC value of 0.48. Given the size and distribution of interaction residues in our dataset, this result is among the best reported in literature (Aumentado-Armstrong *et al.*, 2015; Esmaielbeiki *et al.*, 2016).

The beneficial effect of the grammar correction is more evident when ISPRED4 is scored at the level of interaction patches. Indeed, when the grammar correction is enforced, the POV index measuring the overlap between predicted and real patches improves from 0.41 up to 0.57. This indicates that predictions after the GRHCRF-based correction are more similar to the real PPI patches.

3.3 Comparison with other predictors

We compared the ISPRED4 predictor (including both the SVM and the GRHCRF) with other available state-of-the-art approaches on both the DBv5Sel and the CAPRI-Blind datasets. The benchmarked methods are our previously released ISPRED3 (Savojardo *et al.* 2012), PredUS (Zhang *et al.*, 2011), PrISE (Jordan *et al.*, 2012), SPPIDER (Porollo and Meller, 2007), ProMate (Li *et al.*, 2008), cons-PPISP (Chen and Zhou, 2005) and metaPPI (Huang and Schroeder, 2008). All the methods are available as web servers and are among the best performing approaches developed so far for predicting PPI sites (Aumentado-Armstrong *et al.*, 2015; Xue *et al.* 2015).

Residue- and patch-level prediction performances evaluated on both datasets DBv5Sel and CAPRI-Blind are reported in Table 4, sorted by MCC values computed at the residue level.

It is worth noticing that only ISPRED4 is tested in real cross-validation (as described above), since we cannot control the overlap between the benchmark sets and the training sets adopted to develop the released web-servers.

Results listed in Table 4 indicate that ISPRED4 scores with the highest MCC and POV values on both testing and blind datasets. The true positive rate (recall) is lower than that of other predictors, indicating that ISPRED4 labels as 'interacting' less residues than other methods, however, with a higher likelihood to be correct. Indeed, ISPRED4 is endowed with the highest precision with respect to other predictors (Table 4).

4 Conclusion

The availability of accurate methods to identify PPI sites is crucial to characterize protein function and identify functionally important residues on protein surfaces. Moreover, PPI site predictions can be also incorporated into protein-docking methods to speed-up the search procedure in the conformational space (Zhou and Qin, 2007). In this paper, we present ISPRED4, an improved structure-based predictor of PPI sites. As a classification method, ISPRED4 adopts a combination of SVMs and probabilistic sequence labelling. We performed cross-validation experiments on a newly generated dataset derived from the Docking Benchmark v5 (Vreven *et al.*, 2015) and consisting of 151 high-resolution protein complexes.

Table 4. Residue- and patch-level performance of different methods on the DBv5Sel and CAPRI-Blind datasets

Method	DBv5Sel						CAPRI-Blind					
	MCC	Q_2	Rec(I)	Pre(I)	F1(I)	POV	MCC	Q_2	Rec(I)	Pre(I)	F1(I)	POV
ISPRED4	0.48	0.84	0.39	0.78	0.52	0.57	0.28	0.67	0.38	0.60	0.47	0.49
PredUS[(a)]	0.34	0.67	0.76	0.37	0.50	0.53	0.26	0.67	0.62	0.38	0.47	0.46
PrISE[(b)]	0.33	0.83	0.41	0.42	0.41	0.49	0.21	0.72	0.36	0.41	0.38	0.42
SPPIDER[(c)]	0.28	0.72	0.54	0.39	0.45	0.44	0.16	0.68	0.39	0.36	0.37	0.39
ProMate[(d)]	0.27	0.74	0.48	0.38	0.42	0.36	0.15	0.68	0.38	0.35	0.36	0.34
cons-PPISP[(e)]	0.23	0.77	0.27	0.46	0.34	0.35	0.08	0.72	0.17	0.33	0.22	0.22
ISPRED3[(f)]	0.16	0.47	0.80	0.26	0.39	0.38	0.05	0.45	0.68	0.26	0.38	0.35
metaPPI[(g)]	0.09	0.79	0.13	0.34	0.19	0.20	0.05	0.72	0.12	0.34	0.18	0.17

Note: (a) Zhang *et al*. (2011), (b) Jordan *et al*. (2012), (c) Porollo and Meller (2007), (d) Li *et al*. (2008), (e) Chen and Zhou (2005), (f) Savojardo *et al*. (2012) and (g) Huang and Schroeder (2008).

Furthermore, comparative benchmarks were also performed on a blind test set derived from previous CAPRI experiments.

We trained and tested ISPRED4 on unbound complexes, which is a more stringent evaluation than those usually carried out for this task. Nonetheless, when compared with other state-of-the-art approaches, ISPRED4 out-performs other top-performing methods, on both residue- and patch-level performance measures, scoring as one of the best tools developed so far for PPI site prediction.

Funding

This work was partially supported by: COST BMBS Action TD1101 and Action BM1405 (European Union RTD Framework Program, to R.C.); FARB UNIBO 2012 (to R.C.)

Conflict of Interest: none declared.

References

Adamczak,R. *et al*. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins: Struct. Funct. Genet*., **56**, 753–767.

Aguilar,D. *et al*. (2012) Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. *PLoS One*, **7**, e41430.

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.

Aumentado-Armstrong,T.T. *et al*. (2015) Algorithmic approaches to protein–protein interaction site prediction. *Algorithms Mol. Biol*., **10**, 1–7.

Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.

Burger,L. and van Nimwegen,E. (2008) Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol*., **4**, 165.

Buslje,C.M. *et al*. (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol*., **6**, e1000978.

Chang,C.C. *et al*. (2011) *LIBSVM. A Library for Support Vector Machines, ACM Trans. Intell. Syst. Technol*., **2**, 1–27.

Chen,H.L. and Zhou,H.X. (2005) Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins*, **61**, 21–35.

Chen,P. and Li,J. (2010) Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC Bioinformatics*, **11**, 402.

De Vries,S.J. and Bonvin,A.M. (2008) How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr. Protein Pept. Sci*., **9**, 394–406.

Dong,Z. *et al*. (2014) CRF-based models of protein surfaces improve protein–protein interaction site predictions. *BMC Bioinformatics*, **15**, 277.

Ekeberg,M. *et al*. (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, **87**, 012707.

Esmaielbeiki,R. *et al*. (2016) Progress and challenges in predicting protein interfaces. *Brief. Bioinform*., **17**, 117–131.

Fariselli,P. *et al*. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem*., **269**, 1356–1361.

Fariselli,P. *et al*. (2009) Grammatical-restrained hidden conditional random fields for bioinformatics applications. *Algorithms Mol. Biol*., **4**, 13.

Gallet,X. *et al*. (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol*., **302**, 917–926.

Hopf,T.A. *et al*. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, **3**, e03430.

Huang,B. and Schroeder,M. (2008) Using binding site to improve protein–protein docking. *Gene*, **1-2**, 14–21.

Indio,V. *et al*. (2013) The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields. *Bioinformatics*, **29**, 981–988.

Jones,D.T. *et al*. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Jones,S. and Thornton,J.M. (1997) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol*., **272**, 133–143.

Jordan,R.A. *et al*. (2012) Predicting protein–protein interface residues using local surface structural similarity. *BMC Bioinformatics*, **13**, 1–14.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kamisetty,H. *et al*. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence-and-structure-rich era. *PNAS*, **110**, 15674–15679.

Kidera,A. *et al*. (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem*., **4**, 23–55.

Koike,A. and Takagi,T. (2004) Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel*., **17**, 165–173.

Lafferty,J. *et al*. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the ICML01*, Williams College, Williamstown, MA, USA, pp. 282–289.

Li,M.H. *et al*. (2007) Protein–protein interaction site prediction based on conditional random fields. *Bioinformatics*, **23**, 597–604.

Li,B.Q. *et al*. (2012) Prediction of protein–protein interaction sites by random forest algorithm with mRMR and IFS. *PLoS One*, **77**, e43927.

Li,N. *et al*. (2008) Prediction of protein–protein binding site by using core interface residue and support vector machine. *BMC Bioinformatics*, **9**, 553.

Liu,B. *et al*. (2009) Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC Bioinformatics*, **10**, 381.

Marks,D.S. *et al*. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Mihel,J. *et al*. (2008) PSAIA – protein structure and interaction analyzer. *BMC Struct. Biol*., **8**, 21.

Morcos,F. *et al*. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, **108**, E1293–E1301.

Ofran,Y. and Rost,B. (2007) ISIS: interaction sites identified from sequence. *Bioinformatics*, **23**, 13–16.

Ovchinnikov,S. *et al*. (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, **3**, e02030.

Pazos,F. *et al*. (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol*., **271**, 511–523.

Pintar,A. *et al*. (2002) CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics*, **18**, 980–984.

Pintar,A. *et al*. (2003) DPX: for the analysis of protein core. *Bioinformatics*, **19**, 313–314.

Porollo,A. and Meller,J. (2007) Prediction-based fingerprints of protein–protein interactions. *Proteins: Struct. Funct. Genet*., **66**, 630–645.

Rost,B. and Sander,C. (1993) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.

Remmert,M. *et al*. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Res,I. *et al*. (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Savojardo,C. *et al*. (2011) Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization. *Bioinformatics*, **27**, 2224–2230.

Savojardo,C. *et al*. (2012) Machine-learning methods to predict protein interaction sites in folded proteins. In: Biganzoli, E. *et al*. (eds.) *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Lecture Notes in Computer Science. Vol. **7548**, p. 127–135.

Savojardo,C. *et al*. (2013) BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics*, **29**, 504–505.

Sikić,M. *et al*. (2009) Prediction of protein–protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput. Biol*., **5**, e1000278.

Sudha,G. *et al*. (2014) An overview of recent advances in structural bioinformatics of protein–protein interactions and a guide to their principles. *Prog. Biophys. Mol. Biol*., **116**, 141–150.

Suzek,B.E. *et al*. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Tatusov,R.L. *et al*. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *PNAS*, **91**, 12091–12095.

Vreven,T. *et al*. (2015) Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol*., **427**, 3031–3041.

Weigt,M. *et al*. (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS*, **106**, 67–72.

Xue,L.C. *et al*. (2011) HomPPI: a class of sequence homology based protein–protein interface prediction methods. *BMC Bioinformatics*, **12**, 244.

Xue,L.C. *et al*. (2015) Computational prediction of protein interfaces: a review of data driven methods. *FEBS Lett*., **589**, 3516–3526.

Yan,C. *et al*. (2004) A two-stage classifier for identification of protein–protein interface residues. *Bioinformatics*, **20**, i371–i378.

Zemla,A. *et al*. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins: Struct. Funct. Genet*., **34**, 220–223.

Zhang,Q.C. *et al*. (2011) PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res*., **39**, 283–287.

Zhou,H.X. and Qin,S. (2007) Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**, 2203–2209.