A neural network method to improve prediction of protein-protein interaction sites in heterocomplexes

CONFERENCE PAPER · OCTOBER 2003

DOI: 10.1109/NNSP.2003.1318002 · Source: IEEE Xplore

6 AUTHORS, INCLUDING:

Piero Fariselli
University of Bologna
158 PUBLICATIONS **5,077** CITATIONS
SEE PROFILE

Michele Finelli
Biodec
12 PUBLICATIONS **167** CITATIONS
SEE PROFILE

Pier Luigi Martelli
University of Bologna
108 PUBLICATIONS **2,328** CITATIONS
SEE PROFILE

2003 IEEE XIII Workshop on
Neural Networks for Signal Processing

A NEURAL NETWORK METHOD TO IMPROVE PREDICTION OF PROTEIN-PROTEIN INTERACTION SITES IN HETEROCOMPLEXES

Piero Fariselli[1], Andrea Zauli[2], Ivan Rossi[1,2], Michele Finelli[2], Pier Luigi Martelli[1] and Rita Casadio[1]
1. Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy;
Fax: +39 051 242576;
Email: Rita.Casadio@unibo.it.
2. BioDec srl, via Fanin 48, 40127 Bologna, Italy
Email: info@biodec.com

Abstract. In this paper we describe an algorithm, based on neural networks that adds to the previously published results (ISPRED, www.biocomp.unibo.it) and increases the predictive performance of protein-protein interaction sites in protein structures. The goal is to reduce the number of spurious assignment and developing knowledge based computational approach to focus on clusters of predicted residues on the protein surface. The algorithm is based on neural networks and can be used to highlight putative interacting patches with high reliability, as indicated when tested on known complexes in the PDB. When a smoothing algorithm correlates the network outputs, the accuracy in identifying the interaction patches increases from 73% up 76%. The reliability of the prediction is also increased by the application the smoothing procedure.

INTRODUCTION

In the post genomic era, an effort is made to tackle a new challenge: the integration and exploitation of the theoretical and experimental knowledge base, generated by the various genome projects. The specific aim is to creating predictive models of cells, biochemical processes and ultimately complete organisms. A formidable task is therefore understanding how genetic information results in the concerted action of gene products to generate function. Genome-wide analysis of protein-protein interactions is performed with different experimental approaches, to understand which protein complexes are modules of putative protein networks, at the basis of cell functioning ([1-3] and references therein).

Large scale-studies identify hundreds of potentially interacting proteins or complexes in different organisms albeit with still an unknown rate of false positives, depending on the method adopted [4, 5]. While more experimental work is in progress, an alternative and complimentary approach is the attempt to extract information on putative protein complexes using the most detailed source of protein-protein interaction: the PDB data base of protein structures known with atomic resolution [6-8]

0-7803-8178-5/03/$17.00 © 2003 IEEE

Recently we developed a method based on neural networks that generalizes on known interacting surfaces [9]. The method (ISPRED, available at www.biocomp.unibo.it) correctly predicts, with a cross validation procedure, 73% of the residues involved in protein interaction in a set comprising 226 non redundant chains of functional hetero complexes known with atomic resolution.

In this paper we describe an algorithm, based on neural networks, that adds to the previously published results (ISPRED) and increases the predictive performance of protein-protein interaction sites in protein structures up to 76%. The goal is to reduce the number of spurious assignment and developing a knowledge-based computational approach to focus on clusters of predicted residues on the protein surface. The algorithm can be used to highlight putative interacting patches with high reliability, as indicated when tested on known complexes in the PDB.

SYSTEM AND METHODS

Data Base

The data set for training and testing is the same previously used [9], which was selected from the SPIN database (http://trantor.bioc.columbia.edu/cgi-bin/SPIN/). We ended up with 226 interacting protein chains, whose PDB codes are: 1aby_A, 1aby_B, 1acy_H, 1acy_L, 1ad0_A, 1ad0_B, 1ad9_H, 1ad9_L, 1agb_A, 1agb_B, 1agr_A, 1agr_E, 1ahw_A, 1ahw_B, 1aif_H, 1aif_L, 1ais_A, 1ais_B, 1aj7_H, 1aj7_L, 1all_A, 1all_B, 1aok_A, 1aok_B, 1aqd_A, 1aqd_B, 1aqk_H, 1aqk_L, 1atn_A, 1atn_D, 1aui_A, 1aui_B, 1axi_A, 1axi_B, 1axs_H, 1axs_L, 1baf_H, 1baf_L, 1bbd_H, 1bbd_L, 1bfv_H, 1bfv_L, 1bpl_A, 1bpl_B, 1brl_A, 1brl_B, 1cau_A, 1cau_B, 1cbv_H, 1cbv_L, 1cgs_H, 1cgs_L, 1clo_H, 1clo_L, 1cly_H, 1cly_L, 1clz_H, 1clz_L, 1cpc_A, 1cpc_B, 1dba_H, 1dba_L, 1dfb_H, 1dfb_L, 1eap_A, 1eap_B, 1ebd_A, 1ebd_C, 1efu_A, 1efu_B, 1efv_A, 1efv_B, 1fai_H, 1fai_L, 1fdh_A, 1fdh_G, 1fgv_H, 1fgv_L, 1fig_H, 1fig_L, 1fin_A, 1fin_B, 1flr_H, 1flr_L, 1for_H, 1for_L, 1fpt_H, 1fpt_L, 1frg_H, 1frg_L, 1frv_A, 1frv_B, 1fvc_A, 1fvc_B, 1ggb_H, 1ggb_L, 1ghf_H, 1ghf_L, 1gig_H, 1gig_L, 1gla_F, 1gla_G, 1gua_A, 1gua_B, 1hbh_A, 1hbh_B, 1hds_A, 1hds_B, 1hoc_A, 1hoc_B, 1hyx_H, 1hyx_L, 1ibc_A, 1ibc_B, 1ibg_H, 1ibg_L, 1iea_A, 1iea_B, 1igi_H, 1igi_L, 1igm_H, 1igm_L, 1igt_A, 1igt_B, 1ihf_A, 1ihf_B, 1ikf_H, 1ikf_L, 1ind_H, 1ind_L, 1jck_A, 1jck_B, 1kel_H, 1kel_L, 1kno_A, 1kno_B, 1lgb_A, 1lgb_C, 1lia_A, 1lia_B, 1mam_H, 1mam_L, 1mco_H, 1mco_L, 1mcp_H, 1mcp_L, 1mel_A, 1mel_L, 1mf2_H, 1mf2_L, 1mfb_H, 1mfb_L, 1mhc_A, 1mhc_B, 1mhl_A, 1mhl_C, 1mio_A, 1mio_B, 1mlb_A, 1mlb_B, 1mpa_H, 1mpa_L, 1mrc_H, 1mrc_L, 1ngp_H, 1ngp_L, 1nld_H, 1nld_L, 1npo_A, 1npo_C, 1opg_H, 1opg_L, 1out_A, 1out_B, 1phn_A, 1phn_B, 1psk_H, 1psk_L, 1rbl_A, 1rbl_M, 1rlb_A, 1rlb_E, 1rmf_H, 1rmf_L, 1sct_A, 1sct_B, 1scu_A, 1scu_B, 1seb_A, 1seb_D, 1spg_A, 1spg_B, 1tcr_A, 1tcr_B, 1tmc_A, 1tmc_B, 1ttp_A, 1ttp_B, 1vge_H, 1vge_L, 1vol_A, 1vol_B, 1yec_H, 1yec_L, 1yrn_A, 1yrn_B, 1yuh_H, 1yuh_L, 2btf_A, 2btf_P, 2dhb_A, 2dhb_B, 2fbj_H, 2fbj_L, 2fgw_H, 2fgw_L, 2pcb_A, 2pcb_B, 2pcc_A, 2pcc_B, 2req_A, 2req_B, 7fab_H, 7fab_L, 8fab_A, 8fab_B. This list contains only hetero-complexes (homodimers are not included) with the exclusion of complexes involving proteases, membrane peptides, small proteins, and coiled coils (as defined in the SCOP classification [10]).

Surface and Contact Definition

In order do simplify the patch detection, each protein amino acid is represented by collapsing the entire residue into its carbon alpha (CA). In this way only the three CA coordinates are recorded for each residue. The contacts between the proteins are computed using the CA atom distances between the two chains. According to this procedure, the protein surface is then the collection of the CA coordinates belonging to the exposed residues. Solvent exposure is separately computed for each chain, using the DSSP program [11]. Each complex is split into different files containing only the coordinates of a single chain. After a thorough inspection, for defining a residue exposed or buried, we selected as a threshold cut-off 16% of the relative solvent accessibility [9].

The patches relative to the protein-protein interaction sites are defined for each protein chain using a CA distance cut-off of 1.2 nm. This threshold value is selected after comparison with the patches obtained using an all-atom representation. By this, the number of residue involved in protein-protein interaction sites is about 40% of the whole set of residues contained in the selected database (31910 residues).

Evolutionary Information and Sequence Profile Construction

It is well known that the introduction of the evolutionary information improves significantly the prediction accuracy [9, 12]. This is usually done by using sequence profiles, instead single sequence in the training and testing phases of a trainable method. Given a protein, its sequence profile is a two dimensional array whose first index refers to a specific position in the protein sequence, while the second index (which runs over the sequence alphabet) refers to the frequency of a given residue in that sequence position. In this paper we use two different protein sequence profiles. The first one is directly obtained from the HSSP database, which was computed with the MAXHOM multiple sequence alignment program [13]. The second is computed by processing the PSIBLAST output [14]. In this case for each query sequence we scanned the non redundant sequence data base for the possible target sequences using three PSIBLAST rounds with an e-value set equal at 0.001. Then for each sequence we compile a multiple alignment obtained by the PSIBLAST pair-wise sequence alignments. Finally, we derive a sequence profile for each protein p in our dataset. For sake of clarity here we define Π_p to be the sequence profile of the protein p, and $\Pi_p(i,a)$ is the frequency of the amino acidic residue a in the sequence position i of that profile (obtained either from the HSSP files or by our PSIBLAST-base procedure). This imply that $\Sigma_a [\Pi_p(i,a)] = 1$ always holds.

Neural Networks

The neural network used is a classical feed-forward system with one output node, eight hidden neurons and 220 input units. The one output node defines the probability of being in an interaction site given the current input. Then during training, the desired output is set to one if the current input residue pertains to an interaction site, zero otherwise. The back-propagation algorithm is used to update the neuron junctions, and an early-stopping procedure is used as termination criterion [9].

The neural network is fed using an 11 residue-long window. This window is centered on the surface residue to be predicted with 10 nearest neighbors in the surface patch. The residues included in the input window are close in space, not necessarily contiguous in the sequence and represent a rough approximation of the local surface. Each residue in the input window is encoded as a vector of 20 elements, whose values are taken from the corresponding frequencies in the multiple sequence alignment of the protein as extracted from the profiles $\{\Pi_p\}$.
Training was accomplished on the HSSP profiles, whereas testing was performed on both types of sequence profiles.

In the sequel we represent the 11*20 neural-network-input for the residue i for the protein p with $X_p(i)$, while the corresponding neural network output as $O(i)$). The input $X_p(i)$ is then built starting from the computation of the Euclidean distance among all the surface residues as

$$d(i,j) = [\, \mathbf{r}(CA_i) - \mathbf{r}(CA_j) \,]^{1/2} \qquad (1)$$

where $\mathbf{r}(CA_i)$ are the coordinates of the CA atom belonging to the residue i. In this way for each residue we derive the list of all the protein amino acids sorted by the increasing distance. Since we want only the ten nearest neighbors (and itself) we define the rank function as

$$R_i(j) = k \qquad for\ j=0 \ldots 10 \qquad (2)$$

which returns the sequence position k of the j-*th* surface neighbor of i taken from the sorted distance list. Of course for $j=0$ the function returns i itself ($R_i(0) = i$) since no other residue is closer. It is now easy to see that for a given protein p when the residue i is presented to the network, the k-*th* input neuron assumes the value

$$X_p(i,k) = \Pi_p [\, R_i(\, \mathbf{int}\, (k/20)\,),\ k\ \mathbf{mod}\ 20\,] \qquad (3)$$

where the ***int*** and the ***mod*** functions stand for integer part and the remainder of the division $k/20$, respectively.

Patch Smoothing

The major novelty of this paper is the introduction of smoothing functions to improve the accuracy of the predicted surface patches. Actually, since the neural network predictions are uncorrelated, we force a synergic cooperation among close surface residues by averaging the network output values over the neighbor list ($N=10$). This is done recalculating the probability as

$$F(i) = \Sigma_{j=0..N}\, w(i,j) O(R_i(j)) \,/\, [\Sigma_{j=0..N}\, w(i,j)] \qquad (4)$$

where $w(i,j)$ is a weight associated to the neighbor j of i and $O(k)$ is the network output. We tested several weighting schemes, and among them the three best performing are reported. Surprisingly, they are also the simplest. The first and the most obvious is the uniform (*Uniform*)

$$w^U(i,j) = 1 \qquad for\ all\ j \tag{5}$$

the second is the exponential decreasing with the Euclidean distance between the residue and its neighbors (*Exp*)

$$w^E(i,j) = exp\ [-d(i\ ,\ R_i(j))] \tag{6}$$

The third is the decreasing with the inverse of the Euclidean distance between the residue and its neighbors (*Inv*), or

$$w^I(i,j) = 1/\ [d(i\ ,\ R_i(j))(1-\delta(0,j))+\delta(0,j)] \tag{7}$$

where for compactness we use the Kronecker delta (δ), so $w(i,j)$ is one when we refer to the current residue (j=0) otherwise it decreases with the inverse of the distance.

Measure of Accuracy

Several measures of accuracy are routinely used. For sake of completeness here we review the ones adopted in this paper. The efficiency of the predictor is scored using the statistical indexes defined in the following. The accuracy is:

$$Q2=P/N \tag{8}$$

where P is the total number of correctly predicted residues and N is the total number of residues (ranges from 1 to 0).
The correlation coefficient C is defined as:

$$C(s) = [\ p(s)n(s)-u(s)o(s)\ )]\ /\ D \tag{9}$$

where D is the normalization factor

$$D =[(p(s)+u(s))(p(s)+o(s))(n(s)+u(s))(n(s)+o(s))]^{1/2} \tag{10}$$

for each class s, $p(s)$ and $n(s)$ are respectively the total number of correct predictions and correctly rejected assignments while $u(s)$ and $o(s)$ are the numbers of under and over predictions (ranges from 1 to -1).
The accuracy for each discriminated structure s is evaluated as:

$$Q(s)=p(s)/[\ p(s)+u(s)] \tag{11}$$

where $p(s)$ and $u(s)$ are the same as in equation 10 (ranges from 1 to 0). The probability of correct predictions $P(s)$ is computed as:

$$P(s)=p(s)/[p(s)+o(s)] \tag{12}$$

where $p(s)$ and $o(s)$ are the same as in (10) (ranged from 1 to 0).

Finally it is usually very important assigning a reliability score the each network prediction. With one output neural network this is obtained computing

$$Rel(i)=20*abs(O(i)-0.5) \quad (13)$$

This measure alone does not guarantee that the a prediction with a high reliability score is also correct, but it can be used when there is a correlation between *Rel(i)* and the measures of accuracy reported above.

RESULTS AND DISCUSSION

The neural network predictor (NN) tested with a three fold cross-validation procedure, scores similarly to what previously reported ([9], Table 1). Noteworthy, when we apply the smoothing algorithm the accuracy increases significantly up to 3-4 percentage points, with respect to the neural network system alone (Q2 in table 1). Since both the P(I) and the correlation coefficient C increase, this highlights the fact that the interaction patches are better located using a filtering algorithm. Essentially all the three algorithms improve the NN performance, although *Inv* and *Uniform* perform slightly better than *Exp* on average (Q2 and C in Table 1). On the contrary, *Exp* algorithm achieves a somewhat better coverage than the other two on the interaction site class as indicated by the index Q(I).

TABLE 1: PERFORMANCE ON THE DATA SET USING HSSP FILES.

Method	Q2	P(I)	Q(I)	P(N)	Q(N)	C
NN	0.72	0.73	0.56	0.73	0.85	0.43
Uniform	0.76	0.75	0.58	0.74	0.87	0.48
Exp	0.75	0.75	0.61	0.75	0.85	0.48
Inv	0.76	0.75	0.59	0.75	0.87	0.50

I=interaction site; N = Non interaction site

In order to speed up the calculation (this is the usual implementation on web servers), the sequence profiles are computed using the PSIBLAST program suite [14]. The effect of the application of these new generated profiles is reported in Table 2. In this case, the NN accuracy is lower than that obtained before, when the HSSP profiles were used. This effect is possibly due to the difference in the alignment methods (global and multiple in the case of HSSP, local and pairwise in the case of PSIBLAST). When the smoothing algorithms are applied, obviously one should consider the starting performances. Noteworthy, it can be noticed that the application of the filtering procedures recovers the performance drop and the average accuracy is still two percentage points higher than that achieved by the NN method alone.

TABLE 2: PERFORMANCE ON THE DATA SET USING THE PSIBLAST DERIVED PROFILES.

Method	Q2	P(I)	Q(I)	P(N)	Q(N)	C
NN	0.71	0.72	0.55	0.72	0.84	0.41
Uniform	0.74	0.74	0.58	0.74	0.85	0.45
Exp	0.72	0.74	0.61	0.75	0.83	0.45
Inv	0.74	0.74	0.59	0.75	0.85	0.46

I=interaction site; N = Non interaction site

Another relevant point worth highlighting is the relationship between the reliability index and the method accuracy (see measure of accuracy). In Figure 1, the strong correlation between the two measures is highlighted. This is very important, since when the predictor is applied to proteins whose interaction patches are unknown, the reliability score gives a measure of the likelihood of the method correctness. In our case we can state that when the reliability of NN predictions are higher than 6 the probability of getting a correct answer is 0.83. At the same reliability level, we can see that using *Uniform* or *Exp* algorithms in top of the neural network the probability of giving a correct answer increases up to 0.90. Figure 1 shows also that all the smoothing algorithms outperform the basic method (NN).

In conclusion we can state that the accuracy achieved using a smoothing algorithm (mainly *Inv* or *Uniform*) on the top of a neural network system, is a step forward in the direction of the *interactome* assignment.

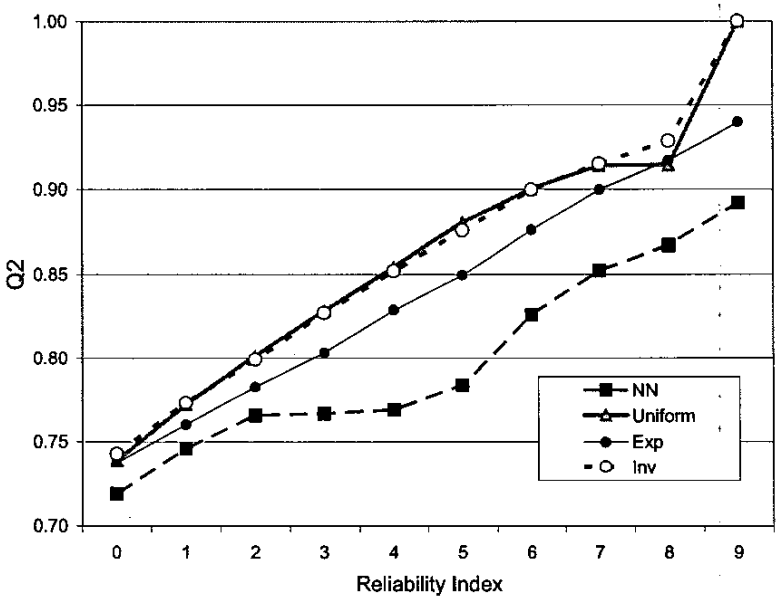

Figure 1: Q2 scores of the Neural network (*NN*) and smoothing algorithms (*Uniform, Exp, Inv*) as a function of the reliability index (*Rel*) of the prediction.

ACKNOWLEDGEMENTS

This work was partially supported by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) for the project "Hydrolases from Thermophiles: Structure, Function and Homologous and Heterologous Expression" and for the project "Development and implementation of algorithms for predicting protein structure", a project on Molecular Genetics and one on Functional Genomics, both of the Italian Centro Nazionale delle Ricerche (CNR), and a PNR 2001-2003 (FIRB art.8) project on Postgenomics, delivered to RC. PLM was the recipient of a fellowship from the Italian Center for National Researches (CNR) devoted to a target project of Molecular Genetics (Law No 449-1997). We thank F. Pazos and A. Valencia for the database construction and scientific discussions.

REFERENCES

[1] SH Strogatz "Exploring complex networks" **Nature** 410, 268-76, 2001

[2] AC Gavin AC, G Superti-Furga G "Protein complexes and proteome organization from yeast to man" **Curr Opin Chem Biol** 7, 21-7, 2003

[3] AC Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, JM Rick, AM Michon, CM Cruciat, M Remor, C Hofert, M Schelder, M Brajenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, B Huhse, C Leutwein, MA Heurtier, RR Copley, A Edelmann, E Querfurth, V Rybin, G Drewes, M Raida, T Bouwmeester, P Bork, B Seraphin, B Kuster, G Neubauer, G Superti-Furga "Functional organization of the yeast proteome by systematic analysis of protein complexes" .**Nature** 415, 141-7, 2002

[4] B Schwikowski, P Uetz, S Fields "A network of protein-protein interactions in yeast".**Nat Biotechnol** 18, 1257-61, .2000

[5] A Valencia, F Pazos "Prediction of protein-protein interactions from evolutionary information" **Methods Biochem Anal** 44, 411-26, 2003

[6] HX Zhou, Y Shan "Prediction of protein interaction sites from sequence profile and residue neighbor list" **Proteins** 44, 336-43, 2001

[7] P Aloy, RB Russell "Interrogating protein interaction networks through structural biology" **Proc Natl Acad Sci U S A.** 99, 5896-901, 2002

[8] A Sali, R Glaeser, T Earnest, W Baumeister "From words to literature in structural proteomics" **Nature** 422, 216-25, 2003

[9] P.Fariselli, F.Pazos, A.Valencia, R.Casadio "Prediction of protein-protein interaction sites in heterocomplexes with neural networks" **Eur J Biochem** 269, 1356-1361, 2002

[10] AG Murzin, SE Brenner, T Hubbard, C. Chotia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures" **J. Mol. Biol** 247, 536-540 1995.

[11] W Kabsch, C Sander "Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features". **Biopolymers** 22, 2577-2637 1983.

[12] B Rost, C Sander "Conservation and prediction of solvent accessibility in protein families" **Proteins: Struct. Func. Genet.**, 20, 216-226 1994.

[13] Dodge C, Schneider R, Sander C "The HSSP database of protein structure-sequence alignments and family profiles" **Nucleic Acids Res**. 26, 313-315 1998.

[14] SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, DJ Lipman "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs" **Nucleic Acid Res.**, 25, 3389-3402 1997.