Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 670125, 9 pages
doi:10.1155/2010/670125

Methodology Report

Development of a Novel Bioinformatics Tool for In Silico Validation of Protein Interactions

Nicola Barbarini,[1] Luca Simonelli,[2] Alberto Azzalin,[3,4] Sergio Comincini,[4] and Riccardo Bellazzi[1]

[1] *Laboratory of Biomedical Informatics, Department of Computer Science and Systems, University of Pavia, 27100 Pavia, Italy*
[2] *Laboratory of Structural Biology, Institute for Research in Biomedicine, 6500 Bellinzona, Switzerland*
[3] *IGM-CNR and Policlinico San Matteo, University of Pavia, 27100 Pavia, Italy*
[4] *Laboratory of Functional Oncogenomics, Department of Genetics and Microbiology, University of Pavia, 27100 Pavia, Italy*

Correspondence should be addressed to Nicola Barbarini, nicola.barbarini@unipv.it

Received 19 October 2009; Revised 10 March 2010; Accepted 30 March 2010

Academic Editor: Rita Casadio

Protein interactions are crucial in most biological processes. Several in silico methods have been recently developed to predict them. This paper describes a bioinformatics method that combines sequence similarity and structural information to support experimental studies on protein interactions. Given a target protein, the approach selects the most likely interactors among the candidates revealed by experimental techniques, but not yet in vivo validated. The sequence and the structural information of the in vivo confirmed proteins and complexes are exploited to evaluate the candidate interactors. Finally, a score is calculated to suggest the most likely interactors of the target protein. As an example, we searched for GRB2 interactors. We ranked a set of 46 candidate interactors by the presented method. These candidates were then reduced to 21, through a score threshold chosen by means of a cross-validation strategy. Among them, the isoform 1 of MAPK14 was in silico confirmed as a GRB2 interactor. Finally, given a set of already confirmed interactors of GRB2, the accuracy and the precision of the approach were 75% and 86%, respectively. In conclusion, the proposed method can be conveniently exploited to select the proteins to be experimentally investigated within a set of potential interactors.

1. Introduction

Proteins rarely perform their biological functions independently, since they usually interact with each other. In fact, most of the biological activities require the creation of protein complexes. As a consequence, the different levels of complexity of the biological systems are not exclusively determined by the number of proteins of an organism, but also by the number of their interactions. Many experimental methods have been developed to study protein interactions, such as the two hybrid system in yeast, the affinity purification followed by mass spectrometry and the phage display libraries [1–4]. The use of these techniques led to the creation of many databases containing a great number of protein-protein interactions, such as the Database of Interacting Proteins (DIPs), the General Repository for Interaction Datasets (BioGRIDs), the Human Protein Reference Database (HPRD), and the Biomolecular Interaction Network Database (BIND) [5–8]. Because of the high amount of false positives and false negatives resulting from the application of these experimental approaches on a large scale, such data repositories must be cautiously used. Once a list of candidates is obtained, it is necessary to analyze in vivo every possible interactor by expensive, time-consuming, and labour-intensive experimental techniques in order to validate the in vitro experimental result. For this reason, in silico methods for the prediction of protein interactions are considered valid tools to reduce the number of candidates [9].

Two main computational approaches, based on sequence similarity and structural modelling, have been applied for the prediction of protein interactions. The former is based on

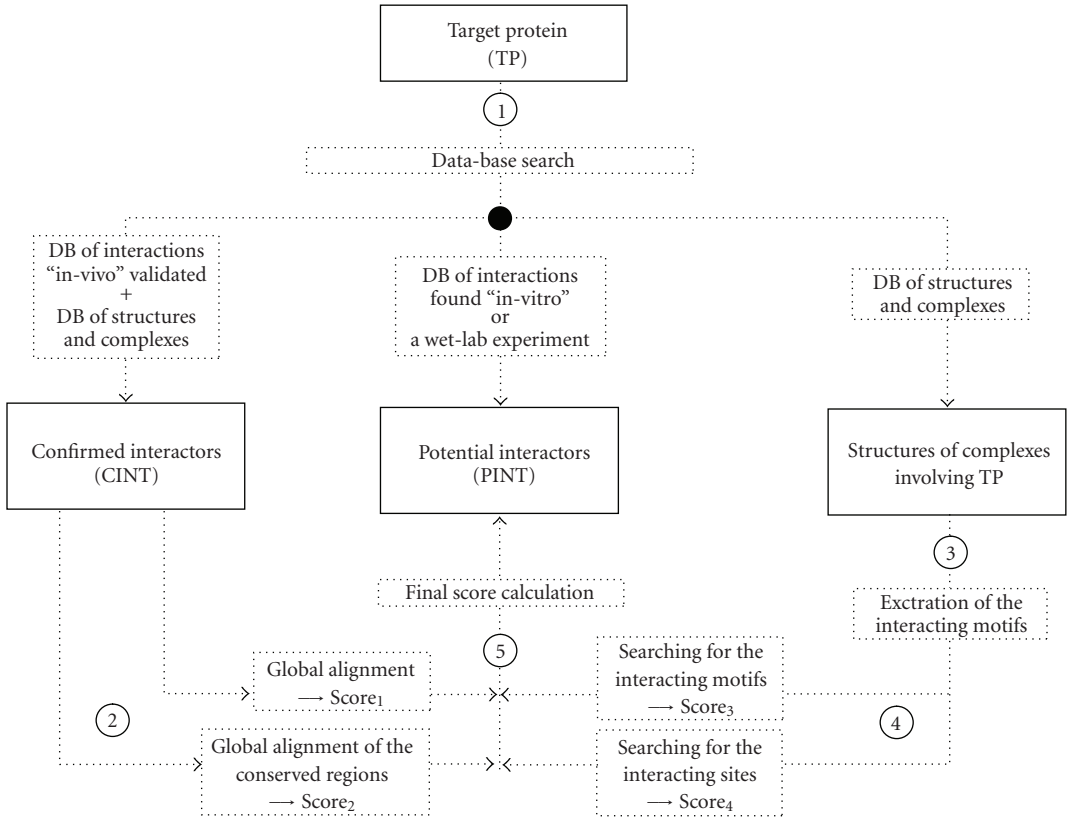

Figure 1: The Bioinformatics strategy for protein interaction prediction.

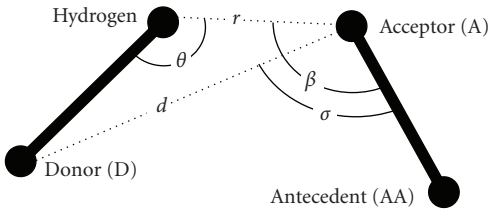

Figure 2: Hydrogen bond geometrical structure scheme. The distance (d) between a donor (D) and an acceptor (A) and the resulting angles σ, β, θ and σ are reported.

the selection of potential interacting partners on the basis of the sequence similarity with confirmed interactors [10–13]. This approach is based on the "homology modelling" principle: similar protein sequences, sharing similar structures, should also share similar interactants [14]. Among sequence-similarity based methods, one of the most widely used is the "mirror tree" approach, based on the assumption that interacting protein pairs are likely to evolve in a correlated fashion [15, 16]. The latter set of approaches is based on the properties of the three-dimensional structures of the proteins, generally referred to as docking methods; these methods exploit surface complementarity and electrostatics properties to predict reliable structural complexes [17–25].

Although both these approaches are based on strong theoretical and experimental data, they exhibit some limitations. The first approach might fail to predict protein interactions, since a protein complex might be subjected to a different selection pressure than each single constituting protein during evolution [26, 27]. For this reason, all the structural details of a protein interaction become important to determine the affinity and the specificity of protein interactions. Docking methods, in fact, analyze the interactions at a three-dimensional level and are therefore considered to be more accurate, also evaluating the biophysical parameters of the interaction sites [28]. However, docking methods are generally limited by the lack of the structures for the majority of the proteins and by incomplete bio-physical interaction knowledge [29, 30]. For these reasons, at present, an integration of the two approaches is essential to better predict putative interacting proteins [31–33].

The aim of this work is to design a knowledge-based tool that can integrate the two approaches described above. In particular, considering a target protein, we first select a set of the candidate interactors already obtained from other experimental results, but not yet in vivo validated. Then, we exploit the sequence and the structural information on in vivo confirmed interacting proteins and complexes, to finally select the most reliable partners of the target protein.

2. Materials and Methods

The proposed bioinformatics approach is summarized in Figure 1. The algorithm predicts the protein interactors of

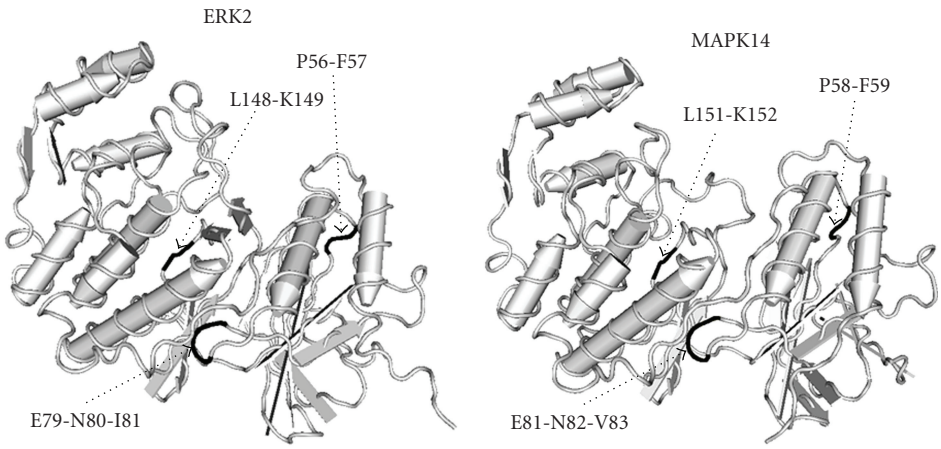

Figure 3: Three-dimensional structures of ERK2 (a) and MAPK14 (b) human proteins from PDB database (PDB files 2E14 and 1A9U, resp.). Critical residues and their positions within the two proteins are reported.

a specific target protein (TP) relying on the following information: (i) a list of potential interactors, already obtained from other experimental results, but not in vivo validated; (ii) in vivo confirmed interactors; and (iii) three-dimensional complex structures involving TP.

The approach follows five steps:

2.1. Database Search. Known structures complexes involving TP are searched in public available databases, specifically the Protein Data Bank (PDB) and the Protein Quaternary Structure (PQS). First, TP interactors are searched in the Human Protein Reference Database (HPRD) [7, available at http://www.hprd.org/]. Then, two collections of TP interactors are generated: confirmed and validated in vivo (CINT) and potential in vitro discovered (PINT). The CINT group also includes TP interacting chain sequences, extracted from PDB and PQS databases.

2.2. Alignment with the Confirmed Interactors: $Score_1$ and $Score_2$. PINT and CINT related sequences are globally aligned using the Needleman and Wunsch algorithm with the matrix BLOSUM50 and fixing the penalties for row and column gaps both equal to -8 [34]. The score of the best alignment of each $I \in$ PINT is the first element of the score function ($\text{Score}_1(I)$).

CINT aminoacid conservation with known protein structures is evaluated using the ConSurf tool [35]. PINT sequences are then aligned with the conserved regions of CINT members, using the Needleman and Wunsch algorithm, thus computing the second component of the score ($\text{Score}_2(I)$).

2.3. Extraction of the Interaction Motifs from the Protein Complexes. The aminoacids at the interface binding sites belonging to chains interacting with TP are retrieved on the basis of the information on known protein complexes. These interacting aminoacids are exploited to build a set of interacting motifs, which are then searched within the PINT members for the calculation of the third and fourth components of the score function ($\text{Score}_3(I)$ and $\text{Score}_4(I)$). These steps of the methods are described in details in the following sections.

2.3.1. Finding the Interacting Aminoacids by Looking at the Intermolecular Distances. Using the Cartesian coordinates of the complexes involving TP, reported in PDB files, we identify putative TP interacting proteins. For every interactor, we find the interacting aminoacids between TP and the interactor chains: these residues are defined as "centres of bond". In order to reduce the computational burden involved in the identification of the interacting residues, we primarily select the aminoacids at the protein interfaces. In detail, we first look for the amino acids that were less than 15 Å far from one residue of the target chain. Then, for each interactor, we consider as interface residues also those closer than 10 Å to the aminoacids already found. In this step, the distance between two aminoacids is calculated as the distance between the α carbon atoms. Once the interfaces are defined, we search for the interacting aminoacids, deriving from disulfur bridges, hydrogen bonds and salt bridges electrostatic interactions. The following aspects are taken into consideration:

(i) A cysteine sulfur bridge satisfies the following geometrical constraints [36]: the two sulfur atoms (SG, according to PDB nomenclature) must be 2–2.1 Å far and the distance between the β carbon of a cysteine and the sulfur atom of the other cysteine is set to 3–3.1 Å.

(ii) For hydrogen bonds, the distances between the acceptor and donor are computed. As shown in Figure 2, the distance (d) between a donor (D) and an acceptor (A) should be less than 3.5 Å and the angle σ should be less than $\pi/2$. Because the majority of

Table 1: List of the atoms considered as acceptors and donors. For both classes, the 3-letter codes of the amino acids, the symbol used in the PDB files of the considered atom, and the maximum number of hydrogen bonds of atom are reported.

Acceptors			Donors		
all	O	1	all	N	1
ASP	OD1	2	HIS	NE2	1
ASP	OD2	2	HIS	ND1	1
GLU	OE1	2	LYS	NZ	3
GLU	OE2	2	ASN	ND2	2
GLN	OE1	1	GLN	NE2	2
ASN	OD1	1	ARG	NE	1
SER	OG	1	ARG	NH1	2
THR	OG1	1	ARG	NH2	2
			TRP	NE1	1
			SER	OG	1
			THR	OG1	1
			TYR	OH	1

the PDB files do not contain the coordinates of the hydrogen atoms, we do not consider the other three geometrical features of the bond ($r < 2.5$ Å, $\beta > \pi/2$ and $\theta > \pi/2$) [37, 38]. Moreover, we define the maximum number of hydrogen bonds that an atom can form (as showed in Table 1), following the results of a statistical analysis reported by McDonald and Thornton [39]. If a residue overtook the maximum number of bonds, we considered only those with the lowest distances between acceptor and donor.

(iii) Salt bridges are electrostatic interactions between residues with opposite charges. The negatively charged atoms at physiological pH are OD1 and OD2 of asparagine, and OD1 and OE2 of glutamic acid. The positively charged ones are NH1, NH2 of arginine and NZ of lysine. For a salt bridge between two residues with opposite charges, the distance between the charged atoms is set to be less than 3.5 Å [36].

2.3.2. Enlargement of the Center of Bond. Because of the folding of the primary structure, two residues that are neighbours on the surface of the three-dimensional structure can be far apart in the protein sequence. This explains why the interacting amino acids are often spread in the protein linear sequence, so that we can find them completely isolated. Moreover, although the interacting amino acids are the most important components of the interaction, also the neighbouring residues may effectively contribute. For this reason, we enlarge the "centre of bond" considering the neighbouring residues with the same hydropathy of the center of bond and then adding the so called proximal amino acids. Hydropathy and proximal amino acids were computed as follows.

Table 2: Example of a center of bond enlargement. The columns show the amino acid one-letter code, the residues coordinates, the status of proximity with respect to the pattern chain in the complex, the status of hydropathy (1 hydrophobic, 0 hydrophilic) and the secondary structure (H = alpha chain; L = loop) of every amino acid around a center of bond. The center of bond is represented by the amino acids within the bold lines (i.e., S and N); the grey-highlighted rows are the results of the symmetrical enlargement due to hydropathy, while the amino acid reported in italic (V) are grouped because of its proximity to the opposite chain.

Amino acid	Position	Proximity	Hydropathy	Sec.Struct.
V	46	0	1	
F	47	0	1	
V	*48*	1	1	H
P	49	0	0	H
K	50	0	0	L
S	**51**	**0**	**0**	**L**
N	**52**	**0**	**0**	**L**
R	53	0	0	L
K	54	0	0	L
V	55	0	1	
I	56	0	1	

(i) We calculate the hydrophobic and the hydrophilic regions of the proteins, using the hydropathy scale of Kyte-Doolittle [40]. In particular, we set to 1 every amino acid with a positive value in the scale (hydrophobic residues) and to 0 those having negative values (hydrophilic residues).

(ii) Then, in addition to directly interacting residues, we also consider those that do not interact, but are closer than 4.2 Å to the corresponding ones of the TP chain, defined as "proximal aminoacids". To calculate the distances, we consider the most external atoms of the backbone of each amino acid.

Table 2 reported an example of the centre of bond enlargement process.

(1) If the center of bond (i.e., SN) is in a hydrophilic region, we perform a symmetrical enlargement, until a hydrophobic amino acid is found (i.e., PK and RK).

(2) Then the proximal amino acids adjacent to the result of the enlargement (i.e., PKSNRK) is added (i.e., V at position 48) As a result, it is possible to group two or more adjacent centers of bond. We define a set of one or more grouped centers of bond as a "binding site." The set of binding sites belonging to an interface between two chains is denoted as "interacting site".

2.3.3. Building the Interacting Motifs. For each binding site, we extract "interacting motif" through the analysis of secondary and tertiary structures of the proteins; these motifs are then searched within the PINT sequences.

To analyse the sequence motifs, we divide the aminoacids into six classes looking at their hydropathy and charge, as

Table 3: Amino acids classes considered with respect to their hydropathy and charge. Every row shows the identification class and the amino acids components.

Class	Amino Acids
I	ILE-VAL-LEU
II	PHE-CYS-MET-ALA
III	GLY-THR-SER-TRP-TYR-PRO
IV	HIS-GLN-ASN
V	GLU-ASP
VI	LYS-ARG

shown in Table 3. Every amino acid belonging to the center of bond is assumed to be invariant in the motif, while the others residues (except proline) belonging to the binding site are variable within their class, as defined in Table 3. In case of proline, its structure do not allow the movement of the bond between the α carbon and the nitrogen of the backbone, blocking the rotation of the chain; as a consequence, the substitution of the proline with any other aminoacid could influence the stability of the interaction. The secondary structures of the interactors, as extracted from the PDB files of the protein complexes, are also considered to compute of the structural motifs.

The obtained motifs of the example sequence (Table 2) is therefore

$$\begin{matrix} [I/V/L] & P & [K/R] & S & N & [R/K] & [K/R] \\ H & H & L & L & L & L & L \end{matrix} \quad (1)$$

where loops (L) and alpha chains (H) are reported.

Finally, we assign a score to every motif according to the following rules: every unchanged residue is scored 20, while every variable amino acid is scored as the ratio between 20 and the number of allowed variations.

Then, the final score of the motif (motif score, ScoreM) is computed by multiplying the single scores of the aminoacids. For example, the score of the sequence motif [IVL]P[KR]SN[RK][RK] is: $(20/3) \times 20 \times (20/2) \times 20 \times 20 \times (20/2) \times (20/2) = 5.33 \times 10^7$.

2.4. Searching for the Interacting Motifs in the Potential Interactors: $Score_3$ and $Score_4$. We search for both the sequence and structural motifs in all the predicted interactors. Because the secondary structure is unknown for the majority of the potential interactors, three of the most used tools for secondary structure prediction, that is, PREDATOR, NNPREDICT and NPS [41–43], are employed. All these algorithms assign to every amino acid a secondary structure as loop (L), alpha chain (H) and beta sheet (E). Therefore, we compute a score for every interactor I of the list PINT as

$$\text{Score}_3(I) = \max\left(\frac{\text{ScoreM}_I}{\text{length}(\text{sequence}(I))}\right), \quad (2)$$

where ScoreM_I is the list of motif scores of the interactor I.

Finally, assuming that a single motif is not sufficient to determine an interaction, it is also important to take into account how many complete interaction sites (sets of binding sites belonging to an interface between two chains) are present. For this reason, we define another score related to the subset of the motifs of the interactor I belonging to an interaction site

$$\text{Score}_4(I) = \sum\left(\frac{\text{Score}'^{\text{M}_I}}{\text{length}(\text{sequence}(I))}\right), \quad (3)$$

where $\text{Score}'\text{M}_I$ is the list of motif scores of the interactor I belonging to an interaction site that includes all the motifs.

2.5. Final Score Calculation. Once the four scores are computed, we calculate a normalised final score, which expresses a measure of the likelihood that a potential interactor is a real interactor of the target protein.

$$\text{Score}(I) = \sum_{j=1}^{4}\left(\frac{\text{Score}_j(I)}{\max\left(\text{Score}_j(I)\right)}\right). \quad (4)$$

3. Results

The proposed procedure was tested by using the growth factor receptor-bound protein 2 (GRB2) as the target protein. Moreover, a validation approach was applied (seeSection 3.6) to evaluate the accuracy and precision of the procedure and to choose a suitable score threshold to predict new interactors.

3.1. Database Search of GRB2 Interactors. We initially retrieved 21 known structures of protein complexes containing GRB2 main domains (i.e., SH2 and SH3).

The list of potential interactors (PINT) was then obtained by retrieving from HPRD database 46 in vitro discovered interactors. We also retrieved from the same database 141 interactors in vivo validated, which together with the sequences of the proteins extracted from the 21 GRB2 complexes, formed a list of 247 confirmed interactors (CINT).

3.2. Alignment with the Confirmed Interactors: $Score_1$ and $Score_2$. Each PINT member was aligned with CINT counterparts, evaluating Score_1 for the best alignments. Next, every potential interactor was aligned with the conserved regions of the confirmed ones, computed with ConSurf, thus obtaining the values of Score_2 (Table 4).

3.3. Extraction of the Interaction Motifs from the Protein Complexes. We extracted binding sites for every interface between GRB2 and the different chains of each of the 21 retrieved complexes. We found 190 different interaction motifs, with scores ranging from a minimum of 20 (single aminoacid motifs) to a maximum of 1.58×10^{11} (R[HQN]QQ[IVL][FCMA][IVL][KR][ED][IVL]E motif).

3.4. Searching for the Interacting Motifs in the Potential Interactors: $Score_3$ and $Score_4$. We searched for the interacting motifs in the PINT list and we selected, for each potential

Table 4: Overview of the scores used to rank the putative GRB2 interactors. NCBI protein accession number, $Score_1$, $Score_2$, the sequence and the structure configuration of the best motif, $Score_3$ and $Score_4$, were reported. The last column showed the final score assigned to each interactor.

Access number	$Score_1$	$Score_2$	Best motif of sequence	Best struct. motif	$Score_3$	Number int. sites	$Score_4$	Final score
NP_001306	344.7	2378	PPP[IVL]	LLLL	148.1	27	181.2	2.52
NP_003014	4	2177	PPP[IVL]	LLLL	95	12	168	2.33
NP_060910	−11.3	1774	PPP[IVL]P	LLLLL	2469.1	18	8.9	2.21
NP_004432	1688.3	3910	[ED]D[ED]	LLL	2	31	14.5	2.08
NP_004407	−43.3	2468	PPP[IVL]	LLLL	86	22	92.2	2.02
NP_003713	−40.3	2417	PPP[IVL]	LLLL	78.4	24	87	1.99
NP_005145	−52.3	3370	[ED]N[IVL]	LLL	1.2	34	15.1	1.98
NP_002030	−2.3	2269	PPP[IVL]	LLLL	76.8	21	83.3	1.95
NP_002511	−3.7	1220	[ED]ED[ED]	LLLL	136	20	151.8	1.9
NP_005148	29.3	3261	[ED]N[IVL]	LLL	1.2	24	4.9	1.88
NP_003311	−25	2204	[ED]ED[ED]	LLLL	71.3	26	81.6	1.86
NP_003362	1032.3	3399	[ED]D[ED]	LLL	2.4	24	7	1.86
NP_006566	798	3359	[ED]D[ED]	LLL	2.4	27	9.9	1.84
NP_149129	−24.7	3141	[HQN][KR]S[GTSWYP][GTSWYP]	HLLLL	18.7	11	20.1	1.82
NP_005556	93.7	2141	[ED]ED[ED]	LLLL	75	12	77	1.8
NP_036252	11.3	2472	PPP[IVL]	LLLL	83.5	26	94.1	1.76
NP_003806	516.3	3112	[IVL]N[IVL]	LLL	1	22	8.3	1.76
NP_001973	955.7	1960	[ED]ED[ED]	LLLL	29.8	31	40.6	1.75
NP_004439	1297.3	2468	[ED]D[ED]	LLL	1.6	32	11.7	1.73
NP_612401	−50.3	2223	PPP[IVL]	LLLL	75.4	11	77	1.73
NP_542179	−19.3	2370	PPP[IVL]	LLLL	91	27	103.9	1.72
NP_004680	−47.7	2296	RR[KR]	LLL	5.6	23	8.2	1.57
NP_002244	1179	1981	[HQN]Q[HQN]	LLL	0.6	30	6.7	1.55
NP_002960	297.7	2304	[ED]D[ED]	LLL	5.4	14	20.5	1.55
NP_000689	−36.3	2377	[KR]D[GTSWYP]	ELL	1	25	8.6	1.5
NP_005875	258.7	2345	RR[KR]	LLL	6.8	13	12.6	1.5
NP_114098	−10.7	2370	[ED]D[ED]	LLL	3	20	8.9	1.49
NP_036428	−14.7	2471	G[GTSWYP]F	LLL	2	18	4.8	1.48
NP_066189	237.7	2526	[GTSWYP]E[IVL]	LLE	0.7	9	1.1	1.48
NP_000869	−4.7	2251	RR[KR]	LLL	7.2	26	11.7	1.44
NP_005222	−22.3	2246	[ED]D[ED]	LLL	3.6	21	7.9	1.42
NP_055413	−20.7	2313	[KR]D[GTSWYP]	ELL	1.1	8	0.6	1.42
NP_065795	−33.3	2166	RR[KR]	LLL	7.8	18	6.1	1.35
NP_000732	106.7	1894	[IVL]N[HQN]	LLL	1.8	16	4.9	1.28
NP_002637	−144.7	186	PPP[IVL]	LLLL	32.6	31	40.8	1.27
NP_000675	347.7	1831	[ED]D[ED]	LLL	4.2	10	1.2	1.25
NP_001773	−28.7	1520	[ED]N[IVL]	LLL	3.7	13	22	1.23
NP_003170	−24.3	1300	[ED]N[IVL]	LLL	4.2	21	31.6	1.21
NP_006454	−18.7	1727	[GTSWYP]K[ED]	HLL	1.6	17	6.5	1.19
NP_000013	−26.7	1523	[GTSWYP]E[KR]	LLE	1.8	19	10.2	1.17
NP_057627	−28	1540	[GTSWYP]S[IVL]	LLL	1.2	13	4.2	1.12
NP_689901	219.3	1436	[GTSWYP]E[KR]	LLE	1.9	17	8.9	1.1
NP_004030	−34.3	1448	[GTSWYP]S[IVL]	LLL	1.3	9	1.3	1.09
NP_955359	376	1292	[ED]D[ED]	LLL	6.3	10	7.9	1.06
NP_542417	−207.7	−4169	[GTSWYP]RP[IVL]P	LLLLL	82.4	26	110.2	1.03
NP_002342	591	538	[KR]D[GTSWYP]	ELL	5.2	6	3.4	0.77

interactor, the one with the highest score. Table 4 highlighted the sequence, the structural configuration and the score ($Score_3$) of the motif with the highest ranking. By grouping the motifs belonging to the same interface between two chains, we found 76 interaction sites. Table 4 showed the number and the sum of the calculated scores only for the motifs belonging to an interaction site ($Score_4$).

3.5. Final Score Calculation. The final score of each of the 46 possible GRB2 interactors was computed as the normalized sum of all the scores, as illustrated in Table 4.

3.6. Validation of the Proposed Method. To test the methodology, a dataset made of 10 GRB2 true (confirmed) and 10 false (not confirmed, randomly chosen proteins) interactors was employed; the final scores were computed for each simulated interaction. To perform an unbiased comparison, we removed the positive interactors from the CINT sequences. Then, we applied a "Leave-one-out" cross-validation procedure by considering each time a different protein as the unique "test" case and the remaining as "training set". We then applied a simple classifier, obtained by computing the best threshold (Th) on the scores of the training set to maximize the Information Gain (IG):

$$\mathrm{IG} = -\sum_{c=1}^{2} p(c)\log_2 p(c) + \sum_{T\,\mathrm{Score}=1}^{2} p(T\mathrm{Score}) \sum_{c=1}^{2} p(cT\,\mathrm{Score})\log_2(cT\,\mathrm{Score}), \tag{5}$$

where c was the class of the protein (confirmed interactors/not interacting proteins) and TScore a binary variable such that

$$T\,\mathrm{Score} = 0, \quad \text{if Score} \le \mathrm{Th}, \qquad T\,\mathrm{Score} = 1, \quad \text{if Score} > \mathrm{Th}. \tag{6}$$

As a result, the mean accuracy of the Leave-one-out procedure was 75%, and the mean precision was 86%.

Finally, we calculated the median value of the thresholds obtained in the leave-one-out process (i.e., 1.63); this value was then used to select 21 proteins with the highest probability to be GRB2 interactors: among these, the isoform 1 of the mitogen-activated protein kinase 14 (MAPK14) had the highest probability score.

4. Discussion

We have developed a novel algorithm for prediction of protein-protein interactions that combines structure similarity and sequence conservation of protein complex interfaces.

The performance of the algorithm was tested on the ability to predict GRB2-interacting proteins. GRB2 is a small adapting protein composed of a SH2 and two SH3 domains. This protein plays a very important role in the process of signal transduction, as a mediator between the growth factors receptors at the cellular membrane level and the cytosolic RAS proteins. In particular, the mitosis promoting signal, stimulated by the epidermal growth factor (EGF), requires the tyrosine kinase activity, originated by specific trans membrane receptors (EGRFs). Starting from these receptors, the activation of RAS consists in a cascade of protein interactions that involves the GRB2/SOS-1 complex. Different sites of auto phosphorylation in the C-terminal region of EGFRs are binding sites for the SH2 domain of GRB2, while its SH3 domains mediate the recruitment of the exchange factor SOS-l, inducing the subsequent activation of RAS proteins. Then, RAS lead to a cascade that ends with the nuclear translocation of phosphorylated MAP kinase, which then activates transcription factors [44].

After our bioinformatics prediction, we ranked a set of 46 potential GRB2 interactors according to their scores, which we assumed to be putative interactors with the target protein. Resorting to a score threshold chosen by means of a cross-validation strategy, we further screened 21 of the most probable interactors. Among these, MAPK14 had the highest probability score of interaction with GRB2. MAPKs are a group of serine/threonine kinases, activated in response to many extracellular stimuli and mediating different signal transduction pathways. Four different MAPKs families were identified in mammalian cells: extracellular signal-regulated kinase (ERK), c-Jun N-terminal kinase/stress-activated protein kinase (JNK/SAPK), ERK5/big MAP kinase 1 (BMK1) and MAPK p38. In particular, MAPK p38 proteins are involved in growth regulation, cellular differentiation, apoptosis, cellular response to inflammation and stress [45]. This subfamily is composed of four members (α, β, γ and δ) and MAPK14 is the isoform α, that together with β, is ubiquitary expressed [46]. To further confirm our in silico prediction, the MAPK14-GRB2 interaction was previously in vitro observed using the GST pull-down technique [47]. In particular, this study hypothesized that in platelets p38α bind to the SH2 domain of GRB2 in response to stimulation mediated by activation of FcγRIIA (CD32) receptor. This association could act by carrying the cytosolic GRB2, with its complexed proteins, towards specific subcellular topologies, driving the complex to specific substrates [47].

At a structural level, the best motif found for MAPK14-GRB2 interaction was PPP[IVL]; this motif was also identified as an interacting site, because in the in silico extracted complex (1GBQ), it allowed the SH3 domain to bind to SOS-1 protein.

However, global alignments between GRB2 and MAPK14 protein sequences did not directly reveal a contribution of the above mentioned motif: differently, three motifs, extracted from complexes in which the SH2 domain of GRB2 was involved, were aligned (i.e., PF, [ED]N[IVL] and [IVL]K). These motifs were localized in Pro58-Phe59, Glu81-Asn82-Val83, and Leu151-Lys152 in MAPK14 protein tertiary structure; as reported in Figure 3, a similar structural topology of the residues Pro56-Phe57, Glu79-Asn80-Ile81, and Leu148-Lys149 was highlighted for ERK2, a tyrosine kinase which has been in vivo confirmed to interact with GRB2 [48].

For this reason, it was possible to predict that MAPK14, as well as ERK2, can interact with the SH2 domain of GRB2, probably through the above mentioned amino acids.

In summary, the method herein proposed is a first step to the definition of a bioinformatics tool to support experimental studies on protein interactions. According to the validation procedure performed, the accuracy and precision of this method were 75% and 86%, respectively. These results might suggest that the proposed bioinformatics approach can be effectively applied to preliminary screen a wide set of protein interactants, such as those deriving from two hybrids systems, to select those to be primarily investigated.

Currently, the main limitations of our method are the small number of complexes with known structures and the relatively poor knowledge on confirmed interactors.

To overcome these limits, we are working on future refinements of the method, in particular on exploiting the available bioinformatics and database knowledge to define different levels of prediction.

Acknowledgments

This study was supported by University of Pavia (FAR, Fondo Ateneo Ricerca). N. Barbarini and R. Bellazzi are supported by the FIRB-MIUR ITALBIONET project and by the SUMMIT project, funded by the European Commission.

References

[1] A. Valencia and F. Pazos, "Computational methods for the prediction of protein interactions," *Current Opinion in Structural Biology*, vol. 12, no. 3, pp. 368–373, 2002.

[2] M. Ferrer and S. C. Harrison, "Peptide ligands to human immunodeficiency virus type 1 gp120 identified from phage display libraries," *Journal of Virology*, vol. 73, no. 7, pp. 5795–5802, 1999.

[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[4] A. C. Gavin, M. Bösche, R. Krause, et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[5] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.

[6] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, pp. D535–D539, 2006.

[7] S. Peri, J. D. Navarro, T. Z. Kristiansen, et al., "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, pp. D497–D501, 2004.

[8] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.

[9] L. Salwinski and D. Eisenberg, "Computational methods of analysis of protein-protein interactions," *Current Opinion in Structural Biology*, vol. 13, no. 3, pp. 377–382, 2003.

[10] J. Espadaler, O. Romero-Isart, R. M. Jackson, and B. Oliva, "Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships," *Bioinformatics*, vol. 21, no. 16, pp. 3360–3368, 2005.

[11] Y. Ofran and B. Rost, "Predicted protein-protein interaction sites from local sequence information," *FEBS Letters*, vol. 544, no. 1–3, pp. 236–239, 2003.

[12] J. L. Chung, W. Wang, and P. E. Bourne, "Exploiting sequence and structure homologs to identify protein-protein binding sites," *Proteins*, vol. 62, no. 3, pp. 630–640, 2006.

[13] H. X. Ta and L. Holm, "Evaluation of different domain-based methods in protein interaction prediction," *Biochemical and Biophysical Research Communications*, vol. 390, no. 3, pp. 357–362, 2009.

[14] Y. Zhang, "Progress and challenges in protein structure prediction," *Current Opinion in Structural Biology*, vol. 18, no. 3, pp. 342–348, 2008.

[15] R. Jothi, P. F. Cherukuri, A. Tasneem, and T. M. Przytycka, "Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions," *Journal of Molecular Biology*, vol. 362, no. 4, pp. 861–875, 2006.

[16] M. G. Kann, B. A. Shoemaker, A. R. Panchenko, and T. M. Przytycka, "Correlated evolution of interacting proteins: looking behind the mirrortree," *Journal of Molecular Biology*, vol. 385, no. 1, pp. 91–98, 2009.

[17] Y. Inbar, D. Schneidman-Duhovny, I. Halperin, A. Oron, R. Nussinov, and H. J. Wolfson, "Approaching the CAPRI challenge with an efficient geometry-based docking," *Proteins*, vol. 60, no. 2, pp. 217–223, 2005.

[18] H. Neuvirth, R. Raz, and G. Schreiber, "ProMate: a structure based prediction program to identify the location of protein-protein binding sites," *Journal of Molecular Biology*, vol. 338, no. 1, pp. 181–199, 2004.

[19] W. A. McLaughlin, T. Hou, and W. Wang, "Prediction of binding sites of peptide recognition domains: an application on Grb2 and SAP SH2 domains," *Journal of Molecular Biology*, vol. 357, no. 4, pp. 1322–1334, 2006.

[20] P. Fariselli, F. Pazos, A. Valencia, and R. Casadio, "Prediction of protein-protein interaction sites in heterocomplexes with neural networks," *European Journal of Biochemistry*, vol. 269, no. 5, pp. 1356–1361, 2002.

[21] J. Fernández-Recio, M. Totrov, and R. Abagyan, "Soft protein-protein docking in internal coordinates," *Protein Science*, vol. 11, no. 2, pp. 280–291, 2002.

[22] J. Fernández-Recio, M. Totrov, and R. Abagyan, "Identification of protein-protein interaction sites from docking energy landscapes," *Journal of Molecular Biology*, vol. 335, no. 3, pp. 843–865, 2004.

[23] J. Fernández-Recio, R. Abagyan, and M. Totrov, "Improving CAPRI predictions: optimized desolvation for rigid-body docking," *Proteins*, vol. 60, no. 2, pp. 308–313, 2005.

[24] M. D. Daily, D. Masica, A. Sivasubramanian, S. Somarouthu, and J. J. Gray, "CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock," *Proteins*, vol. 60, no. 2, pp. 181–186, 2005.

[25] A. D. J. van Dijk, S. J. de Vries, C. Dominguez, H. Chen, H. X. Zhou, and A. M. J. J. Bonvin, "Data-driven docking: HADDOCK'S adventures in CAPRI," *Proteins*, vol. 60, no. 2, pp. 232–238, 2005.

[26] R. E. Valas, S. Yang, and P. E. Bourne, "Nothing about protein structure classification makes sense except in the light of evolution," *Current Opinion in Structural Biology*, vol. 19, no. 3, pp. 329–334, 2009.

[27] B. D. Bruce, "The paradox of plastid transit peptides: conservation of function despite divergence in primary structure," *Biochimica et Biophysica Acta*, vol. 1541, no. 1-2, pp. 2–21, 2001.

[28] N. Andrusier, E. Mashiach, R. Nussinov, and H. J. Wolfson, "Principles of flexible protein-protein docking," *Proteins*, vol. 73, no. 2, pp. 271–289, 2008.

[29] C. Pons, S. Grosdidier, A. Solernou, L. Pérez-Cano, and J. Fernández-Recio, "Present and future chanllenges and limitations in protein-protein clocking," *Proteins*, vol. 78, no. 1, pp. 95–108, 2010.

[30] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress, "Progress and challenges in predicting protein-protein interaction sites," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 233–246, 2009.

[31] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright, "Computational prediction of protein-protein interactions," *Molecular Biotechnology*, vol. 38, no. 1, pp. 1–17, 2008.

[32] S. J. de Vries and A. M. J. J. Bonvin, "How proteins get in touch: interface prediction in the study of biomolecular complexes," *Current Protein and Peptide Science*, vol. 9, no. 4, pp. 394–406, 2008.

[33] O. Keskin, A. Gursoy, B. Ma, and R. Nussinov, "Principles of protein-protein interactions: what are the preferred ways for proteins to interact?" *Chemical Reviews*, vol. 108, no. 4, pp. 1225–1244, 2008.

[34] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[35] M. Landau, I. Mayrose, Y. Rosenberg, et al., "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information," *Nucleic Acids Research*, vol. 33, pp. 299–302, 2005.

[36] M. Jabłoński, A. Kaczmarek, and A. J. Sadlej, "Estimates of the energy of intramolecular hydrogen bonds," *Journal of Physical Chemistry A*, vol. 110, no. 37, pp. 10890–10898, 2006.

[37] D. Xu, C. J. Tsai, and R. Nussinov, "Hydrogen bonds and salt bridges across protein-protein interfaces," *Protein Engineering*, vol. 10, no. 9, pp. 999–1012, 1997.

[38] G. A. Petsko and D. Ringe, *Protein Structure and Function*, chapter 1, New Science Press; Oxford University Press, Oxford, UK, 2004.

[39] I. K. McDonald and J. M. Thornton, "Satisfying hydrogen bonding potential in proteins," *Journal of Molecular Biology*, vol. 238, no. 5, pp. 777–793, 1994.

[40] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[41] D. Frishman and P. Argos, "Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence," *Protein Engineering*, vol. 9, no. 2, pp. 133–142, 1996.

[42] D. G. Kneller, F. E. Cohen, and R. Langridge, "Improvements in protein secondary structure prediction by an enhanced neural network," *Journal of Molecular Biology*, vol. 214, no. 1, pp. 171–182, 1990.

[43] C. Combet, C. Blanchet, C. Geourjon, and G. Deléage, "NPS@: network protein sequence analysis," *Trends in Biochemical Sciences*, vol. 25, no. 3, pp. 147–150, 2000.

[44] A. M. Tari and G. Lopez-Berestein, "GRB2: a pivotal protein in signal transduction," *Seminars in Oncology*, vol. 28, no. 5, pp. 142–147, 2001.

[45] K. Ono and J. Han, "The p38 signal transduction pathway activation and function," *Cellular Signalling*, vol. 12, no. 1, pp. 1–13, 2000.

[46] R. L. Eckert, T. Efimova, S. Balasubramanian, J. F. Crish, F. Bone, and S. Dashti, "p38 mitogen-activated protein kinases on the body surface. A function for p38δ," *Journal of Investigative Dermatology*, vol. 120, no. 5, pp. 823–828, 2003.

[47] A. Robinson, J. Gibbins, B. Rodríguez-Liñares, et al., "Characterization of Grb2-binding proteins in human platelets activated by FcγRIIA cross-linking," *Blood*, vol. 88, no. 2, pp. 522–530, 1996.

[48] M. Lopez-Ilasaca, P. Crespo, P. G. Pellici, J. S. Gutkind, and R. Wetzker, "Linkage of g protein-coupled receptors to the MAPK signaling pathway through Pl 3-kinase γ," *Science*, vol. 275, no. 5298, pp. 394–397, 1997.