Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces

A. Selim Aytuna, Attila Gursoy* and Ozlem Keskin*

Koc University, Center for Computational Biology and Bioinformatics, College of Engineering, Rumelifeneri Yolu 34450 Sariyer, Istanbul, Turkey

Received on February 2, 2005; revised on April 1, 2005; accepted on April 7, 2005
Advance Access publication April 26, 2005

ABSTRACT
Motivation: Elucidation of the full network of protein–protein interactions is crucial for understanding of the principles of biological systems and processes. Thus, there is a need for in silico methods for predicting interactions. We present a novel algorithm for automated prediction of protein–protein interactions that employs a unique bottom-up approach combining structure and sequence conservation in protein interfaces.

Results: Running the algorithm on a template dataset of 67 interfaces and a sequentially non-redundant dataset of 6170 protein structures, 62,816 potential interactions are predicted. These interactions are compared with the ones in two publicly available interaction databases (Database of Interacting Proteins and Biomolecular Interaction Network Database) and also the Protein Data Bank. A significant number of predictions are verified in these databases. The unverified ones may correspond to (1) interactions that are not covered in these databases but known in literature, (2) unknown interactions that actually occur in nature and (3) interactions that do not occur naturally but may possibly be realized synthetically in laboratory conditions. Some unverified interactions, supported significantly with studies found in the literature, are discussed.

Availability: http://gordion.hpc.eng.ku.edu.tr/prism
Contact: agursoy@ku.edu.tr; okeskin@ku.edu.tr

1 INTRODUCTION
Proteins rarely act in isolation; different levels of complexity of biological systems arise not only from the number of the proteins (genes) of the organism but also from the combinatorial interactions among them (Valencia and Pazos, 2002; Ferrer and Harrison, 1999). One of the primary objectives of the post-genomic era is the elucidation of the interactions in cellular systems. The detailed knowledge of the full network of protein–protein interactions, i.e. the distribution and the number of interactions as well as the presence of key nodes in these networks, is expected to provide new insights into the structures and properties of biological systems. Thus, bioinformatics and computational approaches are becoming increasingly important venues as large amount of data become available. Despite the ongoing effort to decipher the complex nature of protein interactions, they are not still entirely understood (Kortemme and Baker, 2004; Chakrabarti and Janin, 2002; LoConte et al., 1999; Jones and Thornton, 1997). The broad recognition of importance of characterizing the set of all protein interactions in a cell has rendered itself in the development of various experimental and computational techniques. These attempts shed light on both the global features and the specifics of the interactions for different types of interactions.

Various experimental methods have been developed to identify protein–protein interactions in various organisms. These involve (1) the traditional top-down proteomic approach where the experiments have been individually designed to identify and validate a small number of specifically targeted interactions or (2) the bottom-up genomic approach, the recently developed high-throughput experiments designed to probe all the potential interactions within an entire genome exhaustively. The latter approach makes use of high throughput mass spectrometry (Gavin et al., 2002), the yeast two-hybrid system (Ito et al., 2001) and phage display libraries (Ferrer and Harrison, 1999; Wu et al., 1999). These methods have so far yielded a considerable amount of data on protein–protein associations and their relative binding strengths. However, many false positives and false negatives identified in these high-throughput experiments highlight the need for caution when interpreting their results. Still, binary interaction results of these experiments are extremely invaluable to interpret protein–protein interactions and construct protein–protein networks (Salwinski and Eisenberg, 2003; Lu et al., 2002). Experimentally, verified interactions have been compiled in various large scale protein–protein interaction datasets (Gavin et al., 2002; Ito et al., 2001; Xenarios et al., 2002; Badger et al., 2003).

Computational methods can address protein–protein interactions at different levels. They may focus on in-depth analysis or carry out a broad scale analysis across large datasets. Through genomic and protein sequence analysis, they may infer whether proteins do interact (Valencia and Pazos, 2002; Marcotte et al., 1999; Salwinski and Eisenberg, 2003; Lu et al., 2002). Or, through structural analysis of proteins and their complexes, they may provide interaction details, essential for understanding processes at the microscopic level (Kortemme and Baker, 2004; Salwinski and Eisenberg, 2003; Chakrabarti and Janin, 2002; LoConte et al., 1999; Jones and Thornton, 1997). Methods using genomic and protein sequence data include analysis of presence or absence of genes in related species, conservation of gene neighborhood, gene fusion events, similarity of phylogenetic trees, correlated mutations on protein surfaces and co-occurrence of sequence domains (Valencia and Pazos, 2002; Salwinski and Eisenberg, 2003). Methods making use of structural data, usually strive to identify functional protein interfaces and rely on considerations of the solvent accessible surface area buried upon association (Janin, 1997), free energy changes...
upon alanine-scanning mutations (Thorn and Bogan, 2001), in silico two-hybrid systems (Pazos and Valencia, 2002), scoring functions based on statistical potentials (Ponsting et al., 2000), physicochemical and geometric properties of the surface, such as electrostatics, hydrophobicity, amino acid composition, shape complementarity and planarity (Jones and Thornton, 1997; Keskin et al., 2004) and evolutionary conservation (Pazos et al., 1997; Lichtenarge and Sowa, 2002; Keskin et al., 2005).

Computational and experimental methods concentrate on the protein–protein interaction problem from different aspects. Therefore, no single method can adequately discover the interactome fully. Converging toward an ideal solution will involve unification of different methods that take up the problem from different, innovative perspectives. This will provide a more complete picture of living cells, leading to a better understanding of biological processes.

1.1 Protein interfaces

Proteins associate through binding sites. These sites are believed to contribute to the biomolecular recognition and binding of proteins by providing specific chemical and physical properties necessary for these processes. There have been many studies on the protein–protein interaction and binding regions. These studies aim to provide a deeper insight of the nature and mechanism of protein recognition (Jones and Thornton, 1996). It has been found that binding regions bury usually surface areas <2000 Å², and these sites usually form a single patch, in contrast to larger multipatch interfaces (Chakraborti and Jamin, 2002). In a recent study, it has been shown that different protein folds may combinatorially assemble to yield similar local interface motifs (Keskin et al., 2004).

Alanine scanning mutagenesis is a very powerful method to analyze the contributions of individual amino acids to protein–protein binding by systematic replacement of protein interface residues by alanine and by measuring the drop in the resultant binding free energy. These experiments show that each residue at protein–protein interfaces does not contribute to the binding free energy equally. Rather, there are only small sets of hotspot residues at interfaces that contribute significantly to binding free energy of the interaction (Clackson and Wells, 1995). Many subsequent studies suggest that the presence of a few hotspots may be a general characteristic of most protein–protein interfaces (Thorn and Bogan, 2001). These generally polar residues are found to be highly correlated with the structurally conserved residues through evolution to optimize function, structure and stability of the protein complexes and enhance feasibility of protein–protein associations (Keskin et al., 2005).

Many of the residues on interfaces that are critical for binding are likely to be evolutionarily conserved. This is because the pace of evolution at interfaces is slower than the rest of the protein (Fraser et al., 2002). The cause of this slower pace of evolution at interfaces can be explained the phenomena of co-evolution, in which substitutions in one protein result in selection pressure for reciprocal changes in interacting partners (Pazos et al., 1997; Fraser et al., 2002). If mutations accumulated during the evolution of an interacting partner is not compensated by correlated mutations in the other partner, the interface, consequently the interaction, is likely to be disrupted. The alanine scanning mutagenesis method is actually based on this principle. Supportive arguments for co-evolution at protein–protein interfaces have been documented in two different studies. In the first one, corresponding phylogenetic trees of interacting proteins were argued to display, in certain cases, a greater degree of similarity than do non-interacting proteins, due to co-evolution (Auerbach et al., 2003; Fryxell, 1996). In the second one, evolutionarily convergent binding sites were found to correspond to the energetically most favorable states (Kortemme and Baker, 2004).

Through time, differences in paces of evolution result in accumulation of similar interfaces across different complexes, accomplishing different functions. In a way, evolution has reused ‘good’ favorable interface structural scaffolds and adapted them to different functions (Keskin et al., 2004).

In this paper, we present a novel, efficient algorithm to predict potential protein–protein interactions and complexes. We start with a set of structurally known protein interfaces, then seek for pairs of proteins that share structure and evolutionarily conserved residues (hotspot) similarity to our known interface dataset. A list of potentially interacting protein pairs is obtained as a final result. Some of these interacting pairs are verified in the Biomolecular Interaction Network Database (BIND) (Dandekar et al., 1998) and Database of Interacting Proteins (DIP) (Xenarios et al., 2002) and the Protein Data Bank (PDB) (Berman et al., 2000) itself.

The approach and the implementation of the algorithm are elaborated in Section 2. Discussions on prediction results and some case studies are presented in Section 3.

2 SYSTEMS AND METHODS

The rationale of our protein–protein prediction algorithm is that, if any two structures contain particular regions on their surfaces that resemble the complementary partners of a known interface, they ‘possibly interact’, through these regions. In other words, if A is known to interact with B, a shares similarity with the binding site of A, b shares similarity with the binding site of B, then we predict that a interacts with b. This resemblance indicates the ability of these structures to structurally and evolutionarily complement each other along an interface, as chains of any template interface do. Figures 1 and 2 show the top level pseudocode and schematic outline of our algorithm, respectively.

The algorithm requires a ‘template’ dataset, i.e. the representative dataset of ‘available’ interfaces; and a ‘target’ dataset, to seek every potential binary interaction between its members. The template dataset handles structure and sequence conservation by combining two previously generated datasets: the structurally non-redundant dataset of protein–protein interfaces extracted from the PDB and the set of conserved residues on these interfaces (computational hotspots). The target dataset is a sequentially non-redundant set of all protein complexes and chains in the PDB.

2.1 The template interface dataset

Keskin et al. (2004) describe a method for finding a structurally and sequentially non-redundant subset of all existing interfaces formed between two protein chains in dimers, trimers or higher complexes of proteins in PDB. They apply this method to get a set of 103 clusters of structurally related interfaces and their representatives.

In generation of this dataset, first, all existing interfaces formed between two protein chains in dimers, trimers or higher complexes of proteins were extracted from the PDB. Interfaces were defined as the set of residues representing a region through which two polypeptide chains bind to each other through non-covalent interactions. This set consisted of contacting residues between the chains (interacting residues), and those that are in their vicinity with a certain distance threshold (neighboring residues), representing the scaffold of the interface. Two residues from the opposite chains were marked as interacting, if there was at least a pair of atoms, one from each residue, at a distance smaller than the sum of their van der Waals radii plus a threshold of 0.5 Å. If the C-o of a non-interacting residue lay at a distance of at most 6.0 Å from a C-o of an already assigned interface residue in the same chain,
along the multiply aligned substructures. If a residue matched identically on and enhanced it with sequence conservation data, which they call computationally conserved residues, were highly correlated with polar residue hotspots, a procedure resulted in 67 interfaces that contained at least one hotspot. This situation, as with the neighboring residue, was clustered with respect to their structural similarities. The dataset is available at http://gordion.hpc.eng.ku.edu.tr/prism

Ma et al. (2003) discovered that particular residues are conserved on structurally similar interfaces, to an extent that suffices distinguishing between binding sites and exposed protein surfaces. Moreover, they found that these conserved residues, were highly correlated with polar residue hotspots, residues that bear more importance than others in defining affinity and stability of an interaction.

The proceeding work of Keskin et al. (2005), describes a method to find structurally conserved residues on clusters of structurally related interfaces. They have applied this method on the resulting dataset of Keskin et al. (2004) and enhanced it with sequence conservation data, which they call computational hotspots. In their method, they structurally aligned members of a given non-redundant interface cluster along their spatially recurring substructural motifs. Then, they considered the frequencies of identically matched residues along the multiply aligned substructures. If a residue matched identically on >50% of the multiply aligned structures, it qualified as a hotspot.

This procedure resulted in 67 interfaces that contained at least one hotspot. The final set contained members as diverse as enzymes, antibodies, viral capsids, etc. We import this dataset as our template interface dataset. We assume that this non-redundant dataset both structurally and evolutionarily (through computational hotspots) represents a subset of available interfaces in the PDB. The complete list of these 67 interfaces can be accessed through the URL: http://gordion.hpc.eng.ku.edu.tr/prism

2.2 The target dataset

The target dataset consists of the list of monomers and complexes that will be compared with the template dataset for structural and evolutionary similarities. Our algorithm predicts interactions by identifying pairs of proteins that may potentially interact in this dataset. The dataset is generated in two steps.

The first step involves the extraction of a non-homologous set of proteins obtained by applying a sequence identity filter of 50% to all existing protein structures in PDB [online service is available at http://www.pdb.org (Li et al., 2002)]. This preliminary list contains 5427 proteins, as of January 27, 2004.

This dataset is then expanded in the second step by splitting multimeric proteins into their constituent chains. But to avoid disturbing the non-redundant nature of the dataset, pairwise sequence alignments are carried out before splitting (by invoking FASTA (Pearson and Lipman, 1988)) and identical partner chains within the complexes are removed (i.e. homodimers) by grouping chains into sets and choosing a representative for each of them.

After these processes, the target dataset becomes a non-homogous subset of all the polypeptide chains and complexes existing in PDB. The polypeptide chains may be in the form of monomers or in the form of isolated constituent chains of multimeric complexes. As of January 27, 2004; the target dataset consists of 6170 structures. Of these structures, 1981 are multimeric and 4189 are monomeric. Of the monomeric structures, 2483 are derived from complexes.

2.3 The algorithm

To find every possible binary interaction between pairs of structures in the target dataset, we need a method to measure the similarity between partners of these representative interfaces and surfaces of target proteins. Accordingly, we extract the surfaces of target proteins and perform successive structural alignments between these surfaces and the partner chains of interfaces in template interface dataset, in an all-against-all manner. This enables us to measure the similarity of a target structure to a template interface partner. If the surfaces of two target proteins (A and B) contain regions ‘similar’ to complementary partner chains of a template interface, we say A and B may interact through these ‘similar’ regions. Figures 1 and 2 show the top level pseudocode and the schematic flow of our algorithm, respectively.

The algorithm starts by extracting the surfaces of target structures by invoking NACCESS program (Hubbard and Thornton, 1993). Along with the atomic accessible surface, NACCESS calculates the relative surface accessibility (RSA) of residues. Jones and Thornton (1997) argue that residues, whose RSAs (percentage of accessibility compared with the accessibility of that residue type X in an extended ALA-X-ALA tripeptide) are >5%, can be considered to be on the surface. We adopt the same criterion to qualify surface residues.

The algorithm then checks whether particular regions on the target surfaces resemble the complementary partners of representative interfaces in the template dataset. This necessitates a defined way to measure the structural and evolutionary similarities between a target surface and a representative interface partner. But before the similarities can be measured, the structures need to be structurally aligned. First, each representative interface picked from the template dataset is split into its constituent partners. Since the template dataset comprises only two-chain interfaces, this process always results in two partners per interface. These individual partners are then structurally aligned with the target surface, by invoking MULTIPROT (Shatsky et al., 2004). MULTIPROT detects common geometrical cores between given protein structures in a sequence-order-independent way. This feature makes MULTIPROT a favorable selection for the task, since protein surfaces and protein–protein interfaces have sequence discontinuity. MULTIPROT returns 10 best substructural matches resulting from every possible alignment. Each substructure corresponds to different regions on the surface, bearing
ively, we obtain an entire list of two-chain interfaces existing in the PDB, generated in Keskin et al. (2004). (2004).

The remaining 5188 are between two complex structures. Most of them are between a monomeric protein and a complex structure. Of these, 31 980 interactions are between the monomeric structures, and 25 448 of them are between a monomeric protein and a complex structure. The remaining 5188 are between two complex structures. Most of these interactions are heterodimers; only 284 are homodimers (100% sequence identity between partners). This number contains predictions with partners having identical sequences, within the same sequence identity removal phase (Section 2.2). Table 1 displays a selected set of predictions with high scores. The first 4 characters in columns 1, 2 and 5 are PDB representations of proteins and the following characters are PDB chain identifiers. In columns 1 and 2, multiple chains are enclosed in curly brackets, to indicate that the chains are identical and the prediction applies to all of them. In column 5, these two characters indicate the chains of the structures between which the template interface exists. Column 3 specifies if the interaction is verified in B (BIND), D (DIP) and P (PDB) databases, whereas an empty entry means an unverified interaction. Column 4 is the similarity score of the prediction. Columns 6

### Table 1. A selected set of verified and unverified predictions

<table>
<thead>
<tr>
<th>Left partner</th>
<th>Right partner</th>
<th>Verified Dbase</th>
<th>Prediction score</th>
<th>Template Left function</th>
<th>Right function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1cov1</td>
<td>1h8tC</td>
<td></td>
<td>4.192</td>
<td>1cov1</td>
<td>Echovirus 11 coat protein</td>
</tr>
<tr>
<td>1dzi</td>
<td>1ncgC</td>
<td></td>
<td>3.867</td>
<td>1cov1</td>
<td>Poliovirus receptor</td>
</tr>
<tr>
<td>1hk9[ AECG]</td>
<td>1jjo[EF]</td>
<td>P</td>
<td>3.453</td>
<td>1as4AB</td>
<td>Plasma serine protease inhibitor</td>
</tr>
<tr>
<td>2ac2[AB]</td>
<td>1e7w[AB]</td>
<td>D,B,P</td>
<td>2.873</td>
<td>1e92AC</td>
<td>Tropinone reductase-II</td>
</tr>
<tr>
<td>2csvE</td>
<td>1lw6l</td>
<td>P</td>
<td>2.749</td>
<td>2snIEI</td>
<td>Subtilisin BPN</td>
</tr>
<tr>
<td>1mho</td>
<td>1psb[AB]</td>
<td>D,B,P</td>
<td>2.484</td>
<td>1mr8AB</td>
<td>S-100 protein, Beta chain</td>
</tr>
<tr>
<td>1h99</td>
<td>1jb</td>
<td></td>
<td>2.469</td>
<td>1sbwAI</td>
<td>Beta-trypsin</td>
</tr>
<tr>
<td>2mn[ABC]</td>
<td>1dg6</td>
<td>D,P</td>
<td>2.225</td>
<td>1cdab</td>
<td>TNF</td>
</tr>
<tr>
<td>1fxKc</td>
<td>1jm7B</td>
<td></td>
<td>2.110</td>
<td>1jm7AB</td>
<td>Prefoldin</td>
</tr>
<tr>
<td>1kgk[AB]</td>
<td>2eb0[ABC]</td>
<td>P</td>
<td>2.088</td>
<td>1cosAB</td>
<td>Vimentin</td>
</tr>
<tr>
<td>1kb9K</td>
<td>1m8v</td>
<td></td>
<td>2.077</td>
<td>1hezCE</td>
<td>Light chain (VI) of Fv- fragment</td>
</tr>
<tr>
<td>1sk1</td>
<td>1msQ[A..Z12]</td>
<td></td>
<td>2.074</td>
<td>1sk12</td>
<td>Putative Smurf Sm-like protein</td>
</tr>
<tr>
<td>1lsd[AB]</td>
<td>1tc7</td>
<td></td>
<td>2.036</td>
<td>1geAC</td>
<td>RAD50 ATPase</td>
</tr>
<tr>
<td>1mso[AC]</td>
<td>1mso[BD]</td>
<td>P</td>
<td>1.981</td>
<td>6rlxAB</td>
<td>Insulin like growth factor A-chain</td>
</tr>
<tr>
<td>1lxm[AB]</td>
<td>1k75[AB]</td>
<td></td>
<td>1.953</td>
<td>1fuuAB</td>
<td>Sporulation response regulatory protein</td>
</tr>
<tr>
<td>1iesB</td>
<td>1ecm[AB]</td>
<td></td>
<td>1.952</td>
<td>1iesAB</td>
<td>Ferritin</td>
</tr>
<tr>
<td>1ju5C</td>
<td>1uff</td>
<td>B</td>
<td>1.947</td>
<td>1azeAB</td>
<td>Intersectin 2</td>
</tr>
<tr>
<td>1osh</td>
<td>1m6eE</td>
<td></td>
<td>1.930</td>
<td>1m6eDE</td>
<td>Bile acid receptor</td>
</tr>
</tbody>
</table>

The characters B, D and P in verified column corresponds to verification in BIND, DIP and PDB databases.

different levels of structural similarity to the interface partner. Among these alignments, the algorithm seeks the most favorable alignment that maximizes our similarity scoring function. The similarity scoring function is defined as \( f_{\text{similarity}} \), where \( f_{\text{evolution}} \) and \( f_{\text{structure}} \) are evolutionary and structural similarity scoring functions, respectively. The coefficient \( \alpha \), represents the relative importance of evolutionary similarity to structural similarity. The first function reflects the number of identically matched hotspots, the second function reflects the size and quality of the alignment along the target–template alignment. We assume that hotspots bear greater importance in defining an interface than geometrical complementarity. Therefore we select \( \alpha = 0.6 \). The condition prior to alignment restrains that interface partner size be at least 0.7 times the target surface size. (Size of a structure is defined as the number of residues it contains.) This condition keeps relatively small interfaces out of computations. Such relatively small interfaces are likely to align perfectly with target surfaces and yield high similarity scores, causing biased and unselective results.

After the completion of successive structural alignments, a similarity list for each interface partner is obtained. If the similarity lists of corresponding partners of a template interface contain \( N \) and \( M \) target structures, respectively, we obtain \( N \times M \) predictions for that interface. A prediction is uniquely represented by \((a,b,c)\) triplets, where \(a\) and \(b\) are predicted targets and \(c\) is the template interface via which the interaction was predicted. The extent of favorableness of the predicted interaction (prediction score) is quantified by simply the sum of the similarity scores of the target pairs.

These predicted interactions are finally verified for existence in two publicly available interaction databases, BIND, DIP and of course, PDB itself. Structures in our target dataset are referenced by PDB codes. However, entries in the interaction databases have their own referencing nomenclature. Therefore, there is a need to identify cross references of targets in the respective interaction databases. This is performed by finding homologous sequences in the interaction databases using FASTA and alignments yielding expectation values \( \geq 10^3 \) are considered homologous. Notice that this process may result in more than one homolog per database. Once this ‘translation’ is done, predicted interactions are checked for existence in the domains of interaction databases. In the case of PDB, the prediction is checked for its presence in the entire list of two-chain interfaces existing in the PDB, generated in Keskin et al. (2004).

### 2.4 Implementation

Both prediction and verification algorithms were implemented in Python Language, due to its powerful attributes regarding Bioinformatics related tasks. Both algorithms take a fairly long time for completion, i.e. on a Linux machine with 2.4 GHz Pentium processor and 1GB memory, the prediction algorithm needs about a week and the verification algorithm needs about a month. This limitation necessitates parallelization for more reasonable response times. Parallelized version of the both algorithms have proven to achieve almost linear speed ups, prediction algorithm was observed to perform 29.39 times faster at a 32 node Beowulf cluster.

### 3 RESULTS AND DISCUSSION

Prediction results contain various interaction pairs, some of which are verified in BIND and DIP interaction databases as well as PDB. Starting from 67 template interfaces we found 62 616 pairwise interactions among the 6170 target proteins. Of these, 31 980 interactions are between the monomeric structures, and 25 448 of them are between a monomeric protein and a complex structure. The remaining 5188 are between two complex structures. Most of these predictions are heterodimers; only 284 are homodimers (100% sequence identity between partners). This number contains predictions with partners having identical sequences, within the same complex. But we would expect these to be low in number, after the 50% sequence identity removal phase (Section 2.2). Table 1 shows a selected set of predictions with high scores. The first 4 columns in Table 1 display the results of predictions with high scores. The first 4 characters in columns 1, 2 and 5 are PDB representations of proteins and the following characters are PDB chain identifiers. In columns 1 and 2, multiple chains are enclosed in curly brackets, to indicate that the chains are identical and the prediction applies to all of them. In column 5, these two characters indicate the chains of the structures between which the template interface exists. Column 3 specifies if the interaction is verified in B (BIND), D (DIP) and P (PDB) databases, whereas an empty entry means an unverified interaction. Column 4 is the similarity score of the prediction. Columns 6
3.1 Biological evidence of some predicted binary protein interactions: case studies

In this section, we discuss two examples in detail. Neither of the cases has been verified in DIP/BIND or in PDB, but the literature search strongly suggests that such interactions exist.

3.1.1 Vitamin D binding protein–parathyroid hormone

In this case, residues 383–411 in vitamin D binding protein (DBP) (PDB reference: D chain of 1kxp, SWISSPROT reference: VTDB_HUMAN) were observed to bind to the residues 1–27 of parathyroid hormone (PTH) (PDB reference: A or B chain of 1et1, SWISSPROT reference: PTHY_HUMAN). The prediction had a score of 2.011. The potentially docked structure of the complex is shown in Figure 3.

PTH regulates calcium and phosphorus levels in blood by inducing transport of an inactive form of vitamin D (calcidiol) from liver to kidney and its conversion into active form (calcitriol) in proximal tubules. Calcitriol, in turn, is transported to small intestine, where it acts to raise the calcium level through increased intestinal absorption of calcium. Like all forms of vitamin D, calcidiol binds to DBP prior to transportation by blood to the kidney. In the kidney, the cellular uptake of DBP–calcidiol complex and PTH are both mediated by an endocytic receptor protein termed megalin, in proximal tubules. Under the regulation of PTH calcitriol is also synthesized in the proximal tubules (Christensen and Birn, 2001; Bikle, 2004, http://www.endotext.org/parathyroid/parathyroid3/parathyroid3.htm). Although an interaction has not been reported in literature, during megalin-mediated uptake, PTH may be interacting with the DBP–calcidiol complex through DBP while exerting its regulatory action on calcitriol synthesis. We believe that this prediction may provide new insights into vitamin D metabolism studies.

3.1.2 BRCA1–RAD50 ATPase

In this case, residues 2846–2882 in BRCA1 (PDB reference: A chain of 1miu, SWISSPROT reference: BRC2_MOUSE) were observed to bind to the residues 395–434 in RAD50 ATPase (PDB reference: A or B chain of 1l8d, SWISSPROT reference: RA50_PYRFU). This prediction had a score of 1.989. The potentially docked structure of the complex is shown in Figure 4.

BRCA1 protein, as a tumor suppressor, plays an important role in maintaining genomic stability. Through the several functional domains it contains, BRCA1 has the ability to interact with numerous proteins and to form complexes. It has been reported that disruption of the potential of BRCA1 to form complexes with RAD50 (via inherited mutations or epigenetic mechanisms in sporadic cancers) leads to loss of DNA repair ability. This is on account of some proteins among the binding partners being responsible for the recognition and repair of DNA, such as the DNA damage repair protein RAD50. RAD50 repairs DNA double-strand breaks by end joining (non-homologous recombination) and meiosis-specific double-strand break formation. It is an essential protein for cell growth and viability (Jhanwar-Uniyal, 2003; Deng and Brodie, 2000).

3.2 Summary of the verified interactions

We predict 62,616 binary interactions starting from 6,170 target proteins. Reasonable amount of these predictions were verified in interaction databases. Table 2 displays the number of verified interactions out of cross referenced interactions for three interaction databases. The results display a good balance of verified and unverified predictions. The higher verification ratio for the PDB database (1094 out of 1497) is because the template interfaces used in prediction have been derived from the PDB database. However, not all the cross referenced interactions have been verified, because the structurally conserved hot spot residues (evolutionary data) have dominant effect in the evaluation of similarity between interface templates and PDB interactions. The verified interactions prove the reliability of our algorithm, whereas unverified ones may correspond to unobserved interactions that actually occur in nature or may be synthetically realized in laboratory conditions. We believe these unverified predictions may have important implications regarding drug design.
Prediction of protein–protein interactions

Table 2. Number of verified predictions

<table>
<thead>
<tr>
<th>Interaction database</th>
<th>Unique verifications</th>
<th>Interaction database size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIP</td>
<td>597 of 4107</td>
<td>43 892</td>
</tr>
<tr>
<td>BIND</td>
<td>431 of 1739</td>
<td>31 243</td>
</tr>
<tr>
<td>PDB</td>
<td>1094 of 1497</td>
<td>21 686</td>
</tr>
</tbody>
</table>

4 CONCLUSION

As large amount of protein structure data become available, predict-ive methods to detect and characterize protein–protein interactions are becoming increasingly important in systems biology. Such knowl-edge aid to researchers in identifying nodes in biochemical or signaling pathways that cause disorders, and designing drugs that exert their therapeutic action on these nodes, instead of modulating the complete set of functions connected with the pathway. An ability to predict possible interaction partners of proteins through identification of their binding sites can provide valuable information on the interaction networks and pathways. In the light of this trend, we have developed a novel algorithm for automated prediction of protein–protein interactions that employs a bottom-up approach combining structure and sequence conservation in protein interfaces. Starting from a previously extracted non-redundant dataset that represents structurally available interfaces in protein–protein interactions, we devise a method to measure the similarity between partners of these representative interfaces and surfaces of target proteins. The algorithm resulted in some 60,000 predictions, some of which were verified in interaction databases and redundant dataset of interface dataset of Keskin et al. (2004). These verified interactions favor the reliability of our approach, whereas unverified ones may point to undiscovered interactions.

ACKNOWLEDGEMENT

We thank Maxim Shatsky for his assistance on using MULTIPROT.

REFERENCES


2855