3D-interologs: A protein-protein interacting evolution database across multiple species

Jinn-Moon Yang[1,2,3], Yung-Chiang Chen[1], and Yu-Shu Lo[1]

[1]*Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan*
[2]*Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*
[3]*Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*
moon@faculty.nctu.edu.tw

Abstract

The 3D-interologs database records the evolution of protein-protein interactions database across multiple species. Based on "3D-domain interologs" and a new scoring function, we infer 173,294 protein-protein interactions by using 1,895 three-dimensional (3D) structure heterodimers to search the UniProt database (4,826,134 protein sequences). For a protein-protein interaction, the 3D-interologs database shows functional annotations (e.g. Gene Ontology annotations), interacting domains and binding models (e.g. hydrogen-bond interactions and conserved residues). Additionally, this database provides couple-conserved residues and the interacting evolution by exploring the interologs across multiple species. Experimental results reveal that the proposed scoring function obtains good agreement for the binding affinity of 275 mutated residues from the ASEdb. The precision and recall of our method are 0.52 and 0.34, respectively, by using 563 non-redundant heterodimers to search on the Integr8 database (549 completely deciphered genomes). The proposed method can infer many of the interactions that would not have been identified from sequence similarity alone. The 3D-interologs database comprises 15,024 species and 283,980 protein-protein interactions, including 173,294 interactions (61%) discovered from 3D-domain interologs and 110,686 interactions (39%) summarized from the IntAct database. The 3D-interologs database is available at http://3D-interologs.life.nctu.edu.tw.

1. Introduction

A major challenge of postgenomic biology is to understand the networks of interacting genes, proteins and small molecules that produce biological functions. The large number of protein interactions [1-3], generated by large-scale experimental methods [4-6] and computational methods [7-13], provides opportunities and challenges in annotating protein functions, protein-protein interactions (PPI) and domain-domain interactions (DDI), and in modeling the cellular signaling and regulatory networks. An approach based on evolutionary cross-species comparisons, such as PathBLAST [14, 15] and interologs (i.e. interactions are conserved across species [9, 16]), is a valuable framework for addressing these issues. However, these methods often cannot respond how a protein interacts with another one across multiple species.

Protein Data Bank (PDB) [17] stores three-dimensional (3D) structure complexes, from which physical interacting domains can be identified to study the DDI and the PPI using comparative modeling [11, 18]. Some DDI databases, such as 3did [19], iPfam [20], and DAPID [21], have recently been derived from PDB. Additionally, some methods have utilized template-based methods (i.e. comparative modeling [11] and fold recognition [18]), which search a 3D-complex library to identify homologous templates of a pair of query protein sequences, in order to predict the protein-protein interactions by accessing interface preference, and score query pair protein sequences according to how they fit the known template structures. In this way, it is difficulty to apply these methods, which should evaluate all possible protein pairs (18,000,000) in one species if it has 6000 proteins [22, 23], to understand the networks, PPIs and DDIs across multiple species.

To address these issues, we provide the 3D-interologs database for protein-protein interacting evolution across multiple species by enhancing the "3D-domain interologs" proposed in our previous works [13, 21]. The 3D-domain interologs is defined as "Domain a (in chain A) interacts with domain b (in chain B) in a known 3D complex, meaning that their inferring protein pair A' (containing domain a) and B' (containing domain b) in the same species would be likely to interact with each other if both protein pairs are homologous". Based on this concept, protein sequence databases can be searched to predict protein-protein interactions across multiple

978-0-7695-3739-9/09 $25.00 © 2009 IEEE
DOI 10.1109/IJCBS.2009.79

species by querying a 3D-dimer complex. The 3D-interologs database can indicate interacting domains and contact residues in order to visualize molecular details of a protein-protein interaction. Additionally, this database can provide couple-conserved residues and evolutionary clues of a query sequence and its partners by examining the interologs across multiple species.

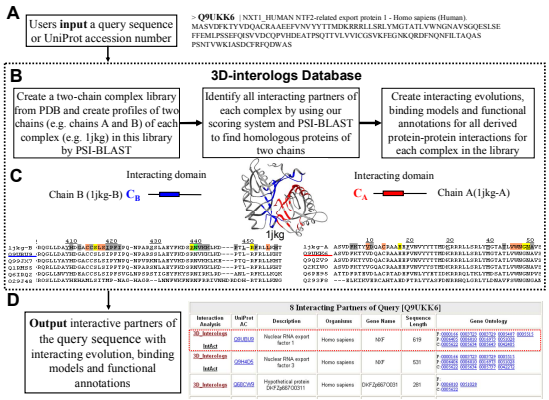

Figure 1. Overview of the 3D-interologs database for protein-protein interacting evolution, protein functions annotations and binding models across multiple species.

2.Methods and Materials

Figure 1 illustrates the overview of the 3D-interologs database. The 3D-interologs allows users to input the UniProt accession number (UniProt AC [24]) or the sequence with FASTA format of the query protein (Figure 1A). When the input is a sequence, 3D-interologs uses BLAST to identify the hit interacting proteins. We identified protein-protein interactions in 3D-interologs database through structure complexes and a new scoring function using the following steps (Figure 1B). First, a 3D-dimer template library comprising 1,895 heterodimers (i.e. 3,790 sequences) was selected from the PDB released in Feb 24, 2006. Duplicate complexes, defined by sequence identity of above 98%, were removed from the library. Dimers containing chains shorter than 30 residues were also excluded [18, 25]. Interacting domains and contact residues of two chains were identified for each complex in the 3D-dimer library. Contact residues, in which any heavy atoms should be within a threshold distance of 4.5 Å to any heavy atoms of another chain, were regarded as the core parts of the 3D-interacting domains in a complex. Each domain was required to have at least 5 contact residues and more than 25 interacting contacted-residue pairs to ensure that the interface between two domains was reasonably extensive. After the interacting domains were determined, its SCOP domains [26] were identified, and its template profiles were constructed by PSI-BLAST. PSI-BLAST was adopted to search the domain sequences against the UniRef90 database [24], in which the sequence identity < 90% of each other and the number of iteration was set to 3.

After 3D-dimer template library and template profiles were built, the 3D-interologs inferred candidates of interacting proteins. To identify the interacting-protein candidates against protein sequences in the UniProt version 11.3 (containing 4,826,134 protein sequences), the chain profile was used as the initial position-specific score matrix (PSSM) of PSI-BLAST in each template consisting of two chains (e.g. C_A and C_B, Figure 1C). The number of iterations was set to 1. Therefore, this search procedure can be considered as a profile-to-sequence alignment. A pairing-protein sequence (e.g. S1 and S2) was considered as a protein-protein interaction candidate if the sequence identity exceeded 30% and the aligned contact residue ratio (*CR*) was greater than 0.5 for both alignments (i.e. S1 aligning to C_A and S2 aligning to C_B). For each interacting candidate, the scoring function was applied to calculate the interacting score and the *Z*-value, which indicates the statistic significance of the interacting score. An interacting candidate was regarded as a protein-protein interaction if its *Z*-value was above 3.0 and it ranked in the Top 25 in one species. The candidate rank was considered in one species to reduce the ill-effect of the out-paralogs that arose from a duplication event before the speciation [27]. These inferred interacting protein pairs were collected in the database.

Finally, for the hit interacting partner derived from 3D-domain interologs, this database provides functional annotations (e.g. UniProt AC, organism, descriptions, and Gene Ontology (GO) annotations [28], Figure 1D), and the visualization of the binding models and interaction evolutions (Figure 1C) between the query protein and its partners. We then constructed two multiple sequence alignments of the query protein and its interacting partner (Figure 1C) across multiple species. Here, the interacting-protein pair with the highest Z-score in a species was chosen as interologs for constructing multiple sequence alignments using a star alignment. The chains (e.g. Chains A and B, Figure 1C) of the hit structure template were considered as the centers, and all selected interacting-protein pairs across species were aligned to respective chains of the template by PSI-BLAST. The 3D-interologs database annotates the important contact residues in the interface according to the following formats: hydrogen-bond residues (green); conserved residues (orange), conserved residues with hydrogen bonds (yellow) and other (gray).

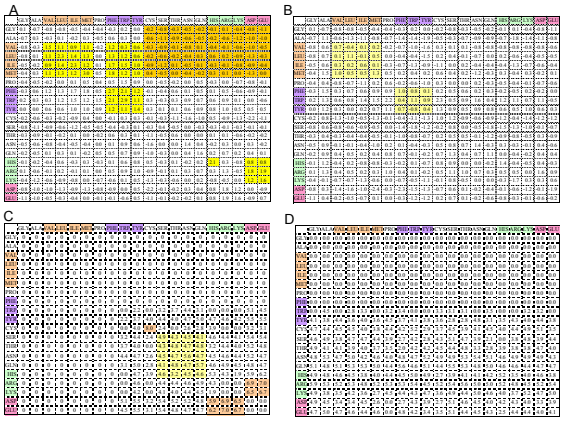

Figure 2. Protein-protein interaction scoring matrices: (A) sidechain-sidechain van-der Waals scoring matrix; (B) sidechain-backbone van-der Waals scoring matrix; (C) sidechain-sidechain special-bond scoring matrix; (D) sidechain-backbone special-bond matrix scoring. The sidechain-sidechain scoring matrices are symmetric and sidechain-backbone scoring matrices are nonsymmetric. For sidechain-sidechain van-der Waals scoring matrix, the scores are high (yellow blocks) if large-aliphatic residues (i.e. Val, Leu, Ile, and Met) interact to large-aliphatic residues or aromatic residues (i.e. Phe, Tyr, and Trp) interact to aromatic residue. In contrast, the scores are low (orange blocks) when nonpolar residues interact to polar residues. For sidechain-sidechain special-bond scoring matrix, the scores are high when an interacting resides (i.e. Cys to Cys) form a disulfide bond or basic residues (i.e. Arg, Lys, and His) interact to acidic residues (Asp and Glu). The scoring values are zero if nonpolar residues interact to other residues.

2.1 Scoring Function and Matrices

We have recently proposed a scoring function to determine the reliability of a protein-protein interaction [13]. This study enhances this scoring by dividing the template consensus score into the template similar score and the couple-conserved residue score. Based on this scoring function, the 3D-interologs database can provide the interacting evolution across multiple species and the statistic significance (*Z*-value), the binding models and functional annotations between the query protein and its interacting partners. The scoring function is defined as

$$E_{tot} = E_{vdw} + E_{SF} + E_{sim} + wE_{cons} \qquad (1)$$

where E_{vdw} and E_{SF} are the interacting van der Waals energy and the special interacting bond energy (i.e. hydrogen-bond energy, electrostatic energy and disulfide-bond energy), respectively; and E_{sim} is the template similar score; and the E_{cons} is couple-conserved residue score. The term w is constant weight (Here, w=3.0). The E_{vdw} and E_{SF} are given as

$$E_{vdw} = \sum_{i,j}^{CP} (Vss_{ij} + Vsb_{ij} + Vsb_{ji})$$

$$E_{SF} = \sum_{i,j}^{CP} (Tss_{ij} + Tsb_{ij} + Tsb_{ji})$$

where CP denotes the number of the aligned-contact residues of proteins A and B aligned to a hit template; Vss_{ij} and Vsb_{ij} (Vsb_{ji}) are the sidechain-sidechain and sidechain-backbone van der Waals energies between residues i (in protein A) and j (in protein B), respectively. Tss_{ij} and Tsb_{ij} (Tsb_{ji}) are the sidechain-sidechain and sidechain-backbone special interacting energies between i and j, respectively, if the pair residues i and j form the special bonds (i.e. hydrogen bond, salt bridge, or disulfide bond) in the template structure. The van der Waals energies (Vss_{ij}, Vsb_{ij}, and Vsb_{ji}) and special interacting energies (Tss_{ij}, Tsb_{ij}, and Tsb_{ji}) were calculated from the four knowledge-based scoring matrices (Figure 2), namely sidechain-sidechain (Figure 2A) and sidechain-backbone van der Waals scoring matrices (Figure 2B); and sidechain-sidechain (Figure 2C) and sidechain-backbone special-bond scoring matrices (Figure 2D).

These four knowledge-based matrices, which were derived using a general mathematical structure [29] from a nonredundant set of 621 3D-dimer complexes proposed by Glaser *et al.* [30], are the key components of the 3D-interologs database for predicting protein-protein interactions. This dataset is composed of 217 heterodimers and 404 homodimers and the sequence identity is less than 30% to each other. The entry (S_{ij}), which is the interacting score for a contact residue i, j pair ($1 \le i, j \le 20$), of a scoring matrix is defined as $S_{ij} = \ln \frac{q_{ij}}{e_{ij}}$, where q_{ij} and e_{ij} are the observed probability and the expected probability, respectively, of the occurrence of each i, j pair. For sidechain-sidechain van-der Waals scoring matrix, the scores are high (yellow blocks) if large-aliphatic residues (i.e. Val, Leu, Ile, and Met) interact to large-aliphatic residues or aromatic residues (i.e. Phe, Tyr, and Trp) interact to aromatic residue. In contrast, the scores are low (orange blocks) when nonpolar residues interact to polar residues. The top two highest scores are 3.0 (Met. interacting to Met) and 2.9 (Trp interacting to Trp).

The value of E_{sim} was calculated from the BLOSUM62 matrix [29] based on the alignments of two chains of the template to the query protein and its interacting partner, respectively. The couple-conserved residue score was determined from two profiles of the template and is given by

$$E_{cons} = \sum_{i,j}^{CP} (\max(0, (M_{ip} - K_{ii}) + (M_{jp'} - K_{jj})) \qquad (2)$$

where CP is the number of contact residue pairs; M_{ip} is the score in the PSSM for residue type i at position p in Protein A; $M_{jp'}$ is the score in the PSSM for

residue type j at position p' in Protein B, and K_{ii} and K_{jj} are the diagonal scores of BLOSUM62 for residue types i and j, respectively.

To evaluate statistical significance (Z-value) of the interacting score of a protein-protein interaction candidate, we randomly generated 10,000 interfaces by mutating 60% contact residues for each heterodimer in 3D-dimer template library. The selected residue was substituted with another amino acid residue according to the probability derived from these 621 complexes [30]. The mean and standard deviation for each 3D-dimer were determined from these 10,000 random interfaces which are assuming to form a normal distribution. Based on the mean and standard deviation, the Z-value of a protein-protein candidate predicted by this template can be calculated.

2.2 Inputs and Outputs

The 3D-interologs database server is easy-to-use. Users input the UniProt AC or the FASTA format of the query protein (Figure 1A). The server generally returns a list of interacting partners with functional annotations (e.g. the gene name, the protein description and GO annotations) (Figure 1D) and provides the visualization of the binding model and contact residues between the query protein and its partner by aligning them to respective template sequences and structures. Additionally, the 3D-interologs system indicates the interacting evolution analysis by using multiple sequence alignments of the interologs across multiple species (Figure 1C). The significant contact residues in the interface are indicated. If Java is installed in the user's browser, then the output shows the structures, and users can dynamically view the binding model, interacting domains and important residues in the browser.

3. Results and Discussions

3.1 Database

The 3D-interologs database currently contains 15,124 species and 283,980 protein-protein interactions, including 173,294 interactions (61%) derived from our method based on 3D-domain interologs and 110,686 interactions (39%) summarized from the IntAct database [3]. For the hit interacting partner derived from 3D-domain interologs, this database provides functional annotations (e.g. UniProt AC, organism, descriptions, and Gene Ontology (GO) annotations [28], Figure 1D), and the visualization of the binding models and interaction evolutions (Figure 1C) between the query protein and its partners. On the other hand, the 3D-interologs database presents only the functional annotations of the hit protein-protein interaction if this interaction was summarized from the IntAct database.

Among 15,124 species in the 3D-interologs database, Table 1 shows 19 species commonly used in molecular research projects, such as *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Escherichia coli*. To analyze couple-conserved residues and interface evolutions for providing evolutionary clues, the 15,124 species were divided into 10 taxonomic groups [31] (Table 2), namely mammalia, vertebrata, metazoa, invertebrata, fungi, plant, bacteria, archaea, viruses, and others.

Table 1. Statistics of 3D-interologs database on 19 species commonly used in research projects. The total number of species is 15124.

Species	3D-Doamin Interologs	IntAct
Mus musculus	8,876	2,634
Homo sapiens	8,639	18,716
Danio rerio	4,564	0
Xenopus laevis	4,057	58
Rattus norvegicus	3,685	958
Bos taurus	3,549	174
Drosophila melanogaster	2,644	25,036
Arabidopsis thaliana	2,418	2,111
Caenorhabditis elegans	1,433	4,684
Saccharomyces cerevisiae	443	36,821
Escherichia coli	426	14,007
Schizosaccharomyces pombe	371	341
Dictyostelium discoideum	284	84
Zea mays	219	0
Oryza sativa	193	69
Takifugu rubripes	191	0
Chlamydomonas reinhardtii	122	14
Plasmodium falciparum	68	2,707
Pneumocystis carinii	23	0
other species	131,083	2,272
Total	173,294	110,686

Table 2. Statistics of 3D-interologs database on 10 taxonomic groups

Taxonomic group name	Number of species	Number of interactions derived from 3D-domain interologs
Mammalia	642	33075
Vertebrata	2747	26028
Metazoa	2712	20278
Invertebrata	180	6767
Fungi	473	11148
Plant	1325	17321
Bacteria	1544	48089
Archaea	106	1866
Viruses	5385	8596
other	10	126
Total	15,124	173,294

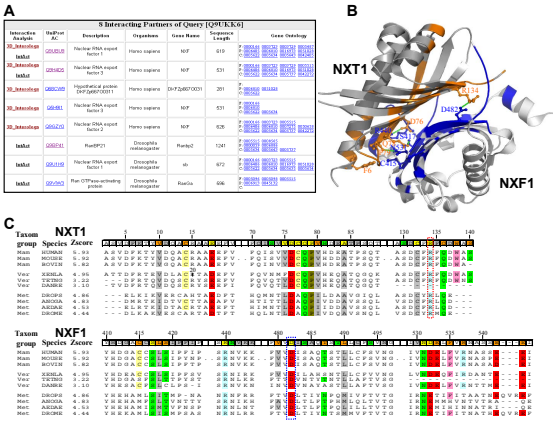

Figure 3. The 3D-interologs database search results of using human NXT1 (UniProt accession number Q9UKK6) as query. (A) Eight interacting partners of NXT1 are found in the 3D-Interologs. For each interacting partner, this server provides UniProt accession number, protein description, organism and Gene Ontology annotation. (B) Detailed interactions between the query and its interacting partner (UniProt accession number Q9UBU9) are indicated via the structure template which consists of NXT1 (PDB entry 1jkg-A) and NXF1 (PDB entry 1jkg-B). The contact residues of NXT1 (query side) and NXF1 (partner side) are colored by red and blue, respectively. The contact residues forming hydrogen bonds (green and dash) are given the atom details. (C) The interacting evolution analysis by using multiple sequence alignments of hit interacting partners of the query across multiple species. The 3D-interologs yields 10 interologs of the query template structure. The contacted residues are marked in template structure based on their interacting characteristics, including hydrogen-bond residues (green); conserved residues (orange); both (yellow), and others (gray). The couple-conserved contact positions are colored in the multiple alignments according to the physical-chemical property of amino acid residues. Twenty amino acid types are classified into 7 groups, namely polar positive (His,Arg, and Lys, blue); polar negative (Asp and Glu, red); polar neutral (Ser, Thr, Asn and Gln, green); cystein (yellow); non-polar aliphatic (Ala, Val, Leu, Ile and Met, gray); non-polar aromatic (Phe, Tyr and Trp, pink); and others: (Gly and Pro, brown).

3.2 Example Analysis

Figure 3 show the search results using the human protein NXT1 (UniProt AC Q9UKK6) [32] as the query sequence. The NXT1, which is a nucleocytoplasmic transport factor and shuttles between the nucleus and cytoplasm, accumulates at the nuclear pore complexes. For this query, 3D-interologs database yielded 8 hit interacting partners (Figure 3A), comprising 5 partners derived from 3D-domain interologs and 5 partners from 3D-domain interologs and 5 partners from the IntACT database. Thus, two partners were present in both databases. Among these 8 hits, 3 partners (i.e. Uniprot AC Q68CW9, Q5H9I1 and Q9GZY0) were not recorded in IntAct database. The Q68CW9, which is part of the protein NXF1 (UniProt AC Q9UBU9), consists of the UBA-like domain and the NTF-like domain, which is responsible for association with the protein NXT1 [33]. The sequence of the protein Q5H9I1 is the same as that of the protein Q9H4D5 (i.e. nuclear RNA export factor 3), which binds to NXT1 [34]. The protein Q9GZY0 (nuclear RNA export factor 2) binds protein NXT1 to export mRNA cargoes from nucleus into cytosol [35].

The protein NXT1 interacts with the protein NXF1 to form a compact heterodimers and an interacting β surface, which is lined with hydrophobic and hydrophilic residues (Figure 3B). Twenty hydrogen bonds or electrostatic interactions are formed in this compact interface. The salt bridge formed by NXT1 Arg134 and NXF1 Asp482 is especially important in the interface [36]. The interacting evolution analysis built by 10 interologs reveals that two residues (Arg134 and Asp482) are conserved in all species (Figure 3C). Additionally, some interacting residues forming the hydrogen bonds are also couple-conserved, for example NXT1 Asp76 and NXF1 Arg440; NXT1 Gln78 and NXF1 Ser417; NXT1 Pro79 and NXF1 Asn531 [36]. The evolution of interaction is valuable to reflect both couple-conserved and critical residues in the binding site.

Conversely, some positions, which are not conserved in all species but conserved in an individual taxonomic group, are important for observing the co-evolution across multiple species. The interacting residue pair (NXT1 Phe6 and NXF1 Cys415) in mammalia and vertebrata is different from that in metazoan (NXF1 Cys415→Met and NXT1 Phe→Leu variant). The van-der Waals potential (1.3 in the sidechain-sidechain van-der Waals scoring matrix, Figure 2A) between Leu and Met is much larger than the potential (−0.1 in this matrix) between Cys and Phe. This co-evolution favors the formation of the hydrophobic interaction in metazoan.

3.3 Binding Affinity Prediction

The enhanced scoring function was first evaluated and compared with the recently proposed scoring function (3D-partner [13]) on two data sets. The first data set, comprising 275 mutated residues selected from the ASEdb database [36], was adopted to reveal the Pearson correlations between ddG values and predicted energies of the 3D-interologs method applying five scoring functions (Figure 4), including E_{tot} (3D-interologs method), E_{cons} (only consensus), E_{vdw}+E_{SF} (only matrices), E_{sim} (only template similarity) and one matrix proposed by Lu, *et al.* [18], where E_{tot}, E_{cons}, E_{vdw}, E_{sim} and E_{SF} are defined in Equation (1). Among these five scoring functions, the E_{tot} is the best (0.92) and one matrix is the worst (0.55, i.e. Lu, *et al.*). The correlations are 0.91 (only

matrices), 0.88 (only template similarity) and 0.84 (only consensus).

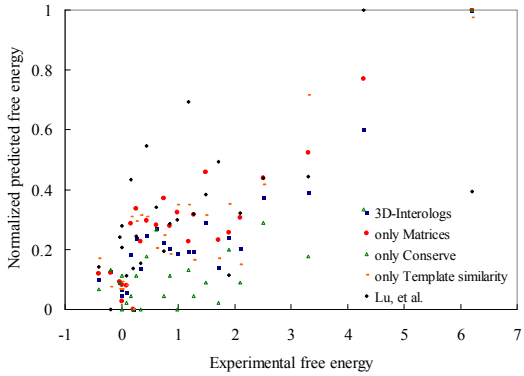

Figure 4. Evaluation of the 3D-interologs method in binding affinities prediction. The Pearson correlations between experimental free energies (ddG) and the predicted values of the 3D-interologs using five scoring functions, including E_{tot} (3D-interologs, blue), E_{vdw}+ E_{SF} (only Matrices, red circle), E_{cons} (only Consensus, green), E_{sim} (only template similarity, orange) and one matrix (black) proposed by Lu, *et al.*, on 275 mutated residues selected from Alanine Scanning Energetics database.

3.4 Interactions Prediction in *S. cerevisiae*

Additionally, a non-redundant set (NR-563), comprising 563 dimer complexes from the 3D-dimer library, was adopted to evaluate the performance of this enhanced scoring function for interacting partner predictions in *S. cerevisiae*. This set comprised 5,882 protein-protein interactions, which were recorded as the core subset in the DIP database as the positive cases, and 2,708,746 non-interacting protein pairs, defined by Jansen *et al*. [7] as the negative cases. The average precisions, which calculated as $(\sum_{i=1}^{A} i/T_h^i)/A$, where T_h^i denotes the number of compounds in a hit list including i correct hits, were 0.84 (this function), 0.82 (3D-partner), and 0.67 for one matrix (proposed by Lu *et al.* [18]). Above results demonstrated that the proposed new scoring function can achieve good agreement for the binding affinity in protein-protein interactions, and can provide statistical significance (Z-value) for predicting protein-protein interactions.

3.5 Interactions Prediction on Multiple Species

To evaluate the performance of the 3D-domain interologs on multiple species, 563 non-redundant dimer complexes (NR-563) were used as queries searching on the Integr8 database (Release 65) which comprises 2,102,196 proteins in 549 species [37]. The Integr8 is an integrated database for organisms with completely deciphered genomes, which are mainly obtained from the non-redundant sets of UniProt entries. Experimentally determined protein-protein interactions dataset were collected from IntAct [3] as the gold standard positive set (110,686 interactions). The gold standard negative set was generated according to the assumption that two proteins acting in the same biological process are more likely to interact than two proteins involved in different processes [38]. This study applied the relative specificity similarity (RSS), proposed by Wu *et al*. [39], to measure the biological process similarity and the location similarity of two proteins based on the GO terms of the biological process (BP) and the cellular component (CC), which describes locations at levels of subcellular structures and macromolecular complexes, respectively. Among 110,686 interactions recorded in the IntAct database, 51,049 interactions can be used to calculate the BP and the CC RSS scores. The BP RSS scores of 4,753 interactions (8.9%) are less than 0.3. Conversely, the CC RSS score of each interaction is more than 0.3. This study considered an interacting protein pair as a negative case if the BP RSS < 0.3 and the CC RSS < 0.3.

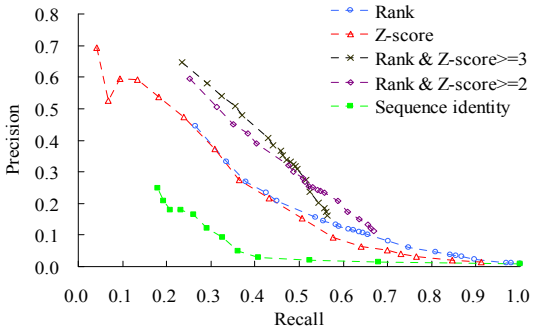

Figure 5. Relationships between precisions and recalls of the 3D-interologs server on the data set NR-563 and the Integr8 database. The 3D-interologs server uses five scoring schemes, including rank in a species (blue), Z-score (red), rank and Z-score >=3 (black), rank and Z-score >=2 (purple), and sequence identity (green).

The structures in the NR-563 as queries to search the Integr8 database yielded 1,063 protein-protein interactions recorded in the IntAct database and 131,831 protein pairs, whose RSS scores were less than 0.3, as the negative set. The precision and recall were adopted to access the predicted quality of the 3D-domain interologs using the five different scoring schemes (Figure 5). The precision was defined as $A_h/(A_h+F_h)$, where A_h and F_h denote the numbers of hit positive cases and hit negative cases, respectively. The recall was defined as A_h/A, where A is the total number of positives (here A=1,063).

Figure 5 shows that the accuracy of the new scoring function (Z-score, red line) is significantly better than that of the sequence identity (green line), and similar to that of candidate ranks (blue line) in one species. The precision is increasing and the recall is decreasing if the Z-value is increasing. Adopting ranks in one species as the scoring function is useful

for distinguishing between positives and negatives when the 3D-domain interologs yield many protein-protein interactions (e.g. > 200) for one species from a structure template. For instance, the 3D-domain interologs obtained about 148 and 125 candidates on average for *Homo sapiens* and *Mus musculus*, respectively, when using NR-563 as queries searching on Integr8 database. The main reason is that a eukaryote genome frequently contains multiple paralogous genes. However, the rank cannot be utilized to calculate the binding affinity of an interacting candidate, and incomplete genome data reduces the performance of the rank. In contrast, the Z-score cannot be adopted to identify the orthologs and in-paralogs arising from a duplication event following the speciation [27]. Figure 5 indicates that the performance of the 3D-interologs using both Z-scores and ranks (black and purple) is the best among these scoring schemes. These results reveal that Z-scores and ranks are complementary, and the accuracy is improved by combining these two scoring methods. For instance, the precision was 0.52 and the recall was 0.34 when $Z > 3.0$ and the rank in one species less than 25.

4. Conclusions

This work demonstrates that the 3D-interologs database is robust and feasible for the interacting evolution of PPIs and DDIs across multiple species. This database can provide couple-conserved residues, interacting models and interface evolution through 3D-domain interologs and a scoring function. The scoring function achieves good agreement for the binding affinity in protein-protein interactions. The 3D-domain interologs method is effective for inferring protein-protein interactions for multiple species. Additionally, the 3D-domain interologs can provide a key source of across-species information for refining sequence-based homology searches.

5. Acknowledgments

J.-M. Yang was supported by National Science Council and partial support of the ATU plan by MOE. Authors are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University.

6. References

[1] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H. W. Mewes, A. Ruepp, and D. Frishman, "The MIPS mammalian protein-protein interaction database," *Bioinformatics*, vol. 21, pp. 832-834, 2005.

[2] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Research*, vol. 28, pp. 289-291, 2000.

[3] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob, "IntAct-open source resource for molecular interaction data," *Nucleic Acids Research*, vol. 35, pp. D561-D565, 2007.

[4] 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 Science of the USA*, vol. 98, pp. 4569-4574, 2001.

[5] A. Pandey and M. Mann, "Proteomics to study genes and genomes," *Nature*, vol. 405, pp. 837-846, 2000.

[6] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399-403, 2002.

[7] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, pp. 449-453, 2003.

[8] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Science of the USA*, vol. 96, pp. 4285-4288, 1999.

[9] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal, "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"," *Genome Research*, vol. 11, pp. 2120-2126, 2001.

[10] J. Wojcik and V. Schachter, "Protein-protein interaction map inference using interacting domain profile pairs," *Bioinformatics*, vol. 17, pp. S296-S305, 2001.

[11] P. Aloy and R. B. Russell, "Interrogating protein interaction networks through structural biology," *Proceedings of the National Academy of Science of the USA*, vol. 99, pp. 5896-5901, 2002.

[12] M. P. Cary, G. D. Bader, and C. Sander, "Pathway information for systems biology," *FEBS Letters*, vol. 579, pp. 1815-1820, 2005.

[13] Y.-C. Chen, Y.-S. Lo, W.-C. Hsu, and J.-M. Yang, "3D-partner: a web server to infer interacting partners and binding models," *Nucleic Acids Research*, pp. W561-W567, 2007.

[14] B. P. Kelley, R. Sharan, R. Karp, E. T. Sittler, D. E. S. Root, B. R., and T. Ideker, "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 11394-11399, 2003.

[15] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker, "Conserved patterns of protein interaction in multiple species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 1974-1979, 2005.

[16] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein, "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs," *Genome Research*, vol. 14, pp. 1107-18, 2004.

[17] N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R. K. Green, J. L. Flippen-Anderson, J. Westbrook, H. M. Berman, and P. E. Bourne, "The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema," *Nucleic Acids Research*, vol. 33, pp. D233-D237, 2005.

[18] L. Lu, H. Lu, and J. Skolnick, "MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading," *Proteins: Structure, Function and Bioinformatics*, vol. 49, pp. 350-364, 2002.

[19] A. Stein, R. B. Russell, and P. Aloy, "3did: interacting protein domains of known three-dimensional structure," *Nucleic Acids Research*, vol. 33, pp. D413-D417, 2005.
[20] R. D. Finn, M. Marshall, and A. Bateman, "iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions," *Bioinformatics*, vol. 21, pp. 410-412, 2005.
[21] Y.-C. Chen, H.-C. Chen, and J.-M. Yang, "DAPID: A 3D-domain annotated protein-protein interaction database," *Genome Informatics*, vol. 17, pp. 206-215, 2006.
[22] L. Lu, A. K. Arakaki, H. Lu, and J. Skolnick, "Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome," *Genome Research*, vol. 13, pp. 1146-1154, 2003.
[23] P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. B. Russell, "Structure-based assembly of protein complexes in yeast," *Science*, vol. 303, pp. 2026-2029, 2004.
[24] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Research*, vol. 34, pp. D187-D191, 2006.
[25] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces," *Journal of molecular biology*, vol. 325, pp. 377-387, 2003.
[26] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data," *Nucleic Acids Research*, vol. 32, pp. D226-D229, 2004.
[27] L. Li, C. J. J. Stoeckert, and D. S. Roos, "OrthoMCL: identification of ortholog groups for eukaryotic genomes," *Genome Research*, vol. 13, pp. 2178-2189, 2003.
[28] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, pp. D258-D261, 2004.
[29] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of The National Academy of Sciences of The United States of America*, vol. 89, pp. 10915-10919, 1992.
[30] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, pp. 89-102, 2001.
[31] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 28, pp. 10-14, 2000.
[32] B. E. Black, L. Levesque, J. M. Holaska, T. C. Wood, and B. M. Paschal, "Identification of an NTF2-related factor that binds Ran-GTP and regulates nuclear protein export," *Molecular and Cellular Biology*, vol. 19, pp. 8616-8624, 1999.
[33] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, pp. 1173-1178, 2005.
[34] A. Herold, M. Suyama, J. P. Rodrigues, I. C. Braun, U. Kutay, M. Carmo-Fonseca, P. Bork, and E. Izaurralde, "TAP (NXF1) belongs to a multigene family of putative RNA export factors with a conserved modular architecture," *Molecular and Cellular Biology*, vol. 20, pp. 8996-9008, 2000.
[35] I. C. Braun, A. Herold, M. Rode, E. Conti, and E. Izaurralde, "Overexpression of TAP/p15 heterodimers bypasses nuclear retention and stimulates nuclear mRNA export," *The Journal of Biological Chemistry*, vol. 276, pp. 20536-20543, 2001.
[36] K. S. Thorn and A. A. Bogan, "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions," *Bioinformatics*, vol. 17, pp. 284-285, 2001.
[37] P. Kersey, L. Bower, L. Morris, A. Horne, R. Petryszak, C. Kanz, A. Kanapin, U. Das, K. Michoud, I. Phan, A. Gattiker, T. Kulikova, N. Faruque, K. Duggan, P. McLaren, B. Reimholz, L. Duret, S. Penel, I. Reuter, and R. Apweiler, "Integr8 and Genome Reviews: integrated views of complete genomes and proteomes," *Nucleic Acids Research*, vol. 33, pp. D297-302, 2005.
[38] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, "Probabilistic model of the human protein-protein interaction network," *Nature Biotechnology*, vol. 23, pp. 951-959, 2005.
[39] X. Wu, L. Zhu, J. Guo, D. Y. Zhang, and K. Lin, "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations," *Nucleic Acids Research*, vol. 34, pp. 2137-2150, 2006.