This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

IAS: Interaction specific GO term associations for predicting Protein-Protein Interaction Networks

Satwica Yerneni, Ishita K. Khan, Qing Wei, and Daisuke Kihara

Abstract— Proteins carry out their function in a cell through interactions with other proteins. A large scale Protein-Protein Interaction (PPI) network of an organism provides static yet an essential structure of interactions, which is valuable clue for understanding the functions of proteins and pathways. PPIs are determined primarily by experimental methods; however, computational PPI prediction methods can supplement or verify PPIs identified by experiment. Here we developed a novel scoring method for predicting PPIs from Gene Ontology (GO) annotations of proteins. Unlike existing methods that consider functional similarity as an indication of interaction between proteins, the new score, named the protein-protein Interaction Association Score (IAS), was computed from GO term associations of known interacting protein pairs in 49 organisms. IAS was evaluated on PPI data of six organisms and found to outperform existing GO term-based scoring methods. Moreover, consensus scoring methods that combine different scores further improved performance of PPI prediction.

Index Terms— Bioinformatics, Proteins, Computational Systems Biology, Biological Interactions

1. INTRODUCTION

Proteins conduct various biological functions through interactions with other proteins. There are complexes of proteins where two or more proteins physically interact and permanently maintain the resulting assembly, which include transporters, molecular machineries in transcription and translation, and molecular chaperones. On the other hand, proteins in signaling pathways interact with each other in a transient fashion and pass signals to downstream proteins. Because protein interactions provide crucial information about how functions of proteins are orchestrated in a cell, tremendous efforts have been paid to develop experimental methods for elucidating protein-protein interactions (PPIs) on a large scale. Experimental methods developed include yeast two hybrid system [1], affinity column-coupled with mass spectrometry [2], and liquid chromatography-coupled mass spectrometry [3]. Revealed PPIs of organisms are stored in databases such as IntAct [4], DIP [5], and GenoBase (for *Escherichia coli* data) [6]. Large PPI data are not only valuable as reference of known interactions but also useful for identifying functional pathway of proteins as well as predicting function of proteins by applying bioinformatics approaches [7-13]. Essentially conventional function prediction methods from PPI are based on the observation that interacting proteins tend to share common function [14, 15].

Experimental methods for determining PPI are costly both in time and in resources. Moreover, an experiment usually identifies a small subset of PPI network of an organism and the rest of the interactions remain unknown. Thus, there is a strong need for computational methods that predict PPIs from various sources. It would be also noted that these PPI prediction methods can be useful as verification tools for PPIs detected by experiments to handle potential errors in experimental results, i.e. false positives and false negatives of PPIs. Indeed potential errors of PPI-detecting experiments have been long discussed, having observed discrepancies of PPIs detected by independently performed experiments [16-18]. Along the same line of de-noising of PPI networks, some computational methods perform missing PPI link prediction from an exisiting PPI topology [19, 20].

Computational PPI prediction methods can be classified according to the source of data used, which include sequence-based, structure-based, expression-based, network-toplology-based, and function-based features. Sequence-based methods can be further classified into two sub-categories, those which use protein sequence features and the other that use comparative genomics approaches. Examples of the former are methods by Shen et al. [21] and Martin et al. [22], which characterized a protein sequence by frequency of n-mer fragments in the sequence and used support vector machine (SVM) to predict interaction between protein pairs. PIPE considers common sequence fragments between a query protein pair to known interacting proteins [23] while PPIevo uses position specific scoring matrices (PSSM) [24] as sequence features in framework of machine learning. Ben-Hur and Noble used sequence similarity defined by several different forms between a query protein pair to known interact-

- *S. Yerneni is with the Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN-55905, USA. E-mail: Yerneni.Satwica@mayo.edu .*
- *I.K. Khan is with the Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA. Email: khan27@purdue.edu.*
- *Q. Wei is with the Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA. Email: wei72@purdue.edu.*
- *D. Kihara is with the Department of Biological Sciences and Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA. Email: dkihara@purdue.edu.*

ing protein pairs for a kernel method [25]. Physico-chemical features of amino acids, such as charge and hydrophobicity, were also used to represent a query protein sequence [26]. The latter sub-category of sequence-based methods compares many genome sequences and identify protein pairs that are coded in close neighbors in genome sequences [27], co-exist/co-absent in genomes [28], or fused into single genes in some genomes [29]. Although these methods primarily predict functionally-related, but not physically-interacting proteins, predicted functional relevance can be strong indicators for predicting protein-protein interactions because proteins often carry out biological function through physical interaction. The STRING database [30] contains a pre-computed list of identified protein pairs of various organisms by the comparative genomics approaches. Pazos et al. proposed to consider similarity in phylogenetic tree of proteins to predict their interactions [31].

Structure-based methods compare the tertiary structure of query protein pairs to structures of known interacting proteins (i.e. protein complex structures in the Protein Data Bank [32]). If the tertiary structure of query proteins has not been solved yet, computational models can be used [33, 34]. PRISM compares query protein structures to a database of known structural interface regions [35]. Wass et al. claimed that interacting proteins may be identified by performing protein-protein docking prediction and examining the distribution of docking scores of decoys [36].

Network-topology-based computational PPI prediction methods rely on exising PPI links to predict missing interactions. OS et al. developed a topological feature-based machine learning model to predict PPI links in fission yeast [20]. Hulovatty et al. used extended neighborhood of proteins in order to extract their topological features to perform missing link prediction in PPI networks [19].

Gene expression data are commonly used for predicting interacting proteins because interacting proteins are expected to have similar expression patterns over different conditions. Typically, gene expression data is combined with other features of protein pairs, such as sequence features, in a machine learning framework [33, 37].

The function of a protein, usually described as a set of Gene Ontology (GO) terms [38], also provides good clue for predicting protein interactions since there are many cases that proteins with the same or similar function form permanent complexes or take part in the same pathway and interact to carry out their biological function. This is reverse from aforementioned protein function prediction methods that use PPI data. Typical PPI prediction methods from GO terms consider similarity of GO terms as an indication of interaction [39-42]. In our previous work [43], we developed two scores for quantifying the "functional coherence" of proteins by considering association of GO terms observed in two biological contexts, co-occurrences in protein annotations and co-mentions in literature in the PubMed database. These two scores are called the Co-occurrence Association Score (CAS) and the PubMed Association Score (PAS), respectively. The scores were shown to be capable of detecting protein pairs that interact with each other [43]. CAS and PAS are not quantifying functional similarity, rather, associations of GO terms, as will be further explained in the Method section.

In this work, we developed a new GO term-based score that was designed to identify interacting proteins. Unlike existing methods that consider GO term similarity as a signature of interacting proteins, we specifically mined GO term pairs that are frequently observed in known interacting protein pairs. The GO term associations were mined and quantified in a similar way as the procedure used for computing CAS and PAS. The new score, named the protein Interaction Association Score (IAS), was first characterized in comparison with the semantic similarity score [44]. Then, IAS was tested on PPI data of six organisms to examine whether it can identify interacting proteins. IAS performed in general better than CAS and PAS. PPI prediction was further improved when IAS, PAS, and CAS were combined to form consensus scores. IAS performs well for PPI prediction by itself and will be a strong component scoring term when combined with various different types of scores in machine learning methods.

2. MATERIALS AND METHODS

2.1. Materials

2.1.1. BIOGRID Database

Protein-Protein Interaction (PPI) data including both physical and genetic interactions were obtained from BIOGRID database [43] (build 3.2.107). This version contains 496,635 non-redundant interactions and 54,236 unique proteins of 49 organisms, out of which 47,239 proteins are associated with a known functional annotation in terms of GO. Along with the interaction data, we also obtained a mapping file that associates a BIOGRID protein identifier to its UniProt ID.

2.1.2 UniProt Database

We downloaded ID to GO Mapping file from UniProt database [44] (version 2014-09), which maps the UniProt ID's to its set of GO terms.

2.1.3 GOA Database

We obtained GO term ontology category (Biological Process, Molecular function, and Cellular Component) information from the Gene Ontology (GO) database (version 2015-01).

2.1.4 True Positive (TP) and True Negative (TN) Protein-Protein Interaction data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

We have considered six organisms (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophilla Melanogaster*, *Caenorhabditis elegans*, *Homo sapiens*, and *Arabidopsis thaliana)* to test the performance of IAS. We eliminated interactions containing proteins that are shared between different organisms. For example, protein-protein interaction with identifiers *13519* and *111825* was found in both *Arabidopsis thaliana* and *Homo sapiens*. Table 1 summarizes the number of interactions and proteins obtained in each organism after removing such shared interactions. We also considered only those interactions which have both the protein identifiers mapped to at-least one GO term annotation. To test each interaction type, we categorized the interactions into three groups as 'All', 'physical', and 'genetic' using the 'Experimental System type' information provided in the interaction data. To compute Receiver Operating Characteristic (ROC) curve of interacting protein pairs for each score discussed in Methods section, we first randomly generated same number of true negative interactions as the number of true positive interactions for each organism. Next, we ranked the interactions in the decreasing order of their corresponding scores to calculate the sensitivity and specificity from the number of True Positives (TP) and True Negatives (TN) captured at different points in the ranked list. The Area under the curve for these ROCs are later discussed in the first half of Section 3.2 and in Table 3.

2.2 Methods

2.2.1 Identifier mapping and retrieval of GO term associations

To retrieve GO term associations, we first mapped each protein interactor from the interaction data to its UniProt IDs and then to its corresponding GO term annotations. During the mapping stage we used only unique GO terms for each protein. Then for each of the 49 organisms' unique interactions, we retrieved all true positive GO term pairs by associating all GO terms from one interactor with the entire set of GO terms annotating the other interactor in the participating pair. Number of associations for each GO term pair was computed from the total GO term pair associations obtained from all the interactions.

2.2.2 GO term Interaction Association Score (GO_IAS)

The GO term Interaction Association Score (GO_IAS) captures the relationship between two GO terms by normalizing the GO term pair co-occurrence count with the total number of protein-protein interactions and also with the number of times each of the participating GO term is used for protein annotation in the interaction data. Once we obtained the number of GO term pair counts and the number of individual GO term counts as mentioned in 2.2.1, the GO_IAS for each GO term pair was computed as follows:

$$GO\_IAS(GOx, GOy) = \frac{\frac{N(GOx, GOy)}{\#T.Edges}}{\left(\frac{N(GOx)}{\#T.Nodes}\right)\left(\frac{N(GOy)}{\#T.Nodes}\right)} \quad (1)$$

where *N(GOx-GOy)* is the number of times GO term pair *GOx* and *GOy* interact in the PPI network, #*T.Edges* is the total number of true positive protein-protein interactions, *N(GOx) and N(GOy)* are the number of times GO term *GOx* and GO term *GOy* independently occur in the network, and #*T.Nodes* is the total number of unique proteins in the interaction network.

2.2.3 Protein pair Interaction Association Score (PPI_IAS)

To test the prediction of two interacting proteins *Pi* and *Pj* using IAS, we re-computed the GO_IAS by removing the two proteins and GO terms in the two proteins in the interaction data. Consequently, interactions between the two proteins (*Pi* and *Pj*) and surrounding proteins were also removed. This is to remove the prior contribution of the target protein pair from IAS. The re-computed GO_IAS score (GO_IASr) for each GO term pair of a given protein pair is calculated as follows:

$$GO\_IASr(GOx, GOy) = \frac{\frac{N(GOx,GOy)-n(GOx,GOy)}{(\#T.Edges-\#nE)}}{\left(\frac{N(GOx)-n(GOx)}{\#T.Nodes-\#nN}\right)\left(\frac{N(GOy)-n(GOy)}{\#T.Nodes-\#nN}\right)} \quad (2)$$

N(GOx, GOy), #*T.Edges*, *N(GOx)*, *N(GOy)* and #*T.Nodes* would remain the same as mentioned in (1). *n(GOx, GOy)* is the number of times the GO term pair *GOx* and *GOy* occur in the interactions containing protein *Pi* and *Pj* either interacting with each other or with other proteins. #nE is the total number of protein-protein interactions containing the proteins *Pi* and *Pj*. *n(GOx)* and *n(GOy)* are the number of times one or both of the proteins in the respective participating pairs carry the same GO term. #*nN* is the number of proteins in the participating pair which is always 2.

With all the GO term pair scores re-computed for a given protein pair, we calculated PPI_IAS that quantifies how likely the two proteins interact. This computation is based on a matrix of GO term pair scores, where row values consist of *GO_IASr* scores of each GOterm of protein *Pi* with every GO term of the protein *Pj* and the maximum score per row is captured to compute the final protein pair score. Similarly, the column values consist of the *GO_IASr* scores of each GO term of protein *Pj* with every GO term of protein *Pi*. The maximum score of each column are then captured for the final score computation. Following that, the PPI_IAS score is defined as:

$$PPI\_IAS(Pi, Pj) = \max\{Row\_Score, Column\_Score\}, \quad (3)$$

where

$$Row\_Score = \frac{1}{N}\sum_{i=1}^{N} \max_{1 \le j \le M} GO\_IASr_{ij} \quad (4)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

and $Column\_Score = \frac{1}{M}\sum_{j=1}^{M} \max_{1\le i\le N} GO\_IASr_{ij}$ (5)

Here, *N* is number of GO terms annotated to the protein *Pi* and *M* is the number of GO terms annotated to the protein *Pj* for a given protein pair. *GO_IASrij* is the re-computed GO term pair IAS for each associating GO term pair of *Pi* and P*j*.

Using this method, we calculated the Protein IAS scores for both true positive and true negative interactions in each organism.

2.2.4 Co-occurrence Association Score (CAS) and PubMed Association Score (PAS)

PPI prediction performance by IAS was compared with the Co-occurrence Association Score (CAS) and PubMed Association Score (PAS) for GO term pairs previously designed by our group [41]. The protein pair Co-occurrence Association Score (PPI_CAS) and protein pair PubMed Association Score (PP_PAS) were computed in a similar way as how PPI_IAS was computed in 2.2.3, except that the GO term scores are not re-computed as the two scores are not directly dependent on the interaction data for their GO score computations.

CAS was designed to quantify the frequency of co-occurrences of two GO terms in a single gene annotation relative to random chance [41] and is computed as follows:

$$CAS(i,j) = \frac{\frac{C(i,j)}{\sum_{i,j} C(i,j)}}{\left(\frac{C(i)}{\sum_k C(k)}\right)\left(\frac{C(j)}{\sum_k C(k)}\right)} \quad (6)$$

where *C(i,j)* is the number of sequences in the database that contain both the GO terms *i* and *j*. Similarly, *C(i)* is the total number of sequences annotated with the GO term *i*, and so is the *C(j)*. CAS also includes GO hierarchy information in scoring the term pairs.

On the other hand, PAS is based on the number of times a given GO term pair occurs in the PubMed Abstracts of National Center for Biotechnology information (NCBI) [41]. PAS was computed in the same way as CAS.

$$PAS(i,j) = \frac{\frac{Pub(i,j)}{\sum_{ij} Pub(i,j)}}{\left(\frac{Pub(i)}{\sum_k Pub(k)}\right)\left(\frac{Pub(j)}{\sum_k Pub(k)}\right)} \quad (7)$$

Here, *Pub(i,j)* is the PubMed abstracts count which contain both the GO terms *i* and *j*. Similarly, *Pub(i)* is the number of abstracts that contain GO term *i* and the same is applicable for *Pub(j)*.

Using these pre-computed GO term scores, we calculated protein pair CAS and PAS scores to compute ROC curves for comparison with IAS and other derived scores mentioned in the following sections.

2.2.5 Average Z-Score (*Avg_Zscore*) of a protein pair

Along with protein CAS and PAS scores, we also computed two consensus scores: Average Z-score (Avg_Zscore) and Average Rank Score (Avg_Rank) by combining all the three scores (PPI_IAS, PPI_CAS, and PPI_PAS) for a give protein pair. We used these scores to test if these combination scores performed better than individual scores, i.e. IAS, CAS, and PAS. To compute Avg_Zscore for a particular protein pair, we first computed the row score (IAS, PAS, or CAS) of the all protein pairs of the organism and computed the Z-score of the protein pair using the distribution. The Z-score of IAS of protein *Pi* and *Pj* is defined as follows:

$$IAS\_Zscore(Pi,Pj) = \frac{(PPI\_IAS(Pi,Pj) - \mu_{IAS})}{(\sigma_{IAS})} \quad (8)$$

Here, PPI_*IAS(Pi, Pj)* was computed following Equation 3. μ_{IAS} is the mean and σ_{IAS} is the standard deviation of IAS of protein pairs of the organism.

$$PAS\_Zscore(Pi,Pj) = \frac{(PPI\_PAS(Pi,Pj) - \mu_{PAS})}{(\sigma_{PAS})} \quad (9)$$

PPI_PAS(Pi, Pj) is based on Equation 3 except for PAS is used instead of IASr. μ_{PAS} is the mean and σ_{PAS} is the standard deviation of PAS of protein pairs of the organism.

$$CAS\_Zscore(Pi,Pj) = \frac{(PPI\_CAS(Pi,Pj) - \mu_{CAS})}{(\sigma_{CAS})} \quad (10)$$

Similarly, *PPI_CAS(Pi, Pj)* is based on Equation 3 except for CAS is used instead of IASr. μ_{CAS} is the mean and σ_{CAS} is the standard deviation of CAS of protein pairs of the organism.

After the IAS_Zscore, CAS_Zscore, and PAS_Zscore are calculated for a given protein pair, we average the three scores to obtain the Avg_Zscore as shown:

$$Avg\_Zscore(Pi,Pj) = \frac{1}{3}\begin{pmatrix} (IAS\_Zscore(Pi,Pj)) \\ +(CAS\_Zscore(Pi,Pj)) \\ +(PAS\_Zscore(Pi,Pj)) \end{pmatrix} \quad (11)$$

2.2.6 Average Rank Score (*Avg_Rank*) of a protein pair

The Average Rank score was also computed for each protein pair by using all the three scores (PPI_IAS, PPI_PAS, and PPI_CAS). This time all the protein pairs in the organism were ranked in the descending order of their scores in the PPI data of the organism. Protein pairs with the same score carry the same rank. The average rank score was computed as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

$$Avg\_Rank(Pi,Pj) = \frac{1}{3}\begin{pmatrix} IAS\_Rank(Pi,Pj) \\ +PAS\_Rank(Pi,Pj) \\ +CAS\_Rank(Pi,Pj) \end{pmatrix} \quad (12)$$

Here, *IAS_Rank(Pi, Pj)* is the rank of *PPI_IAS* for the protein pair *Pi* and *Pj* in the organism. Similarly, *PAS_Rank(Pi, Pj)* and *CAS_Rank(Pi, Pj)* are the ranks of *PPI_PAS* and *PPI_CAS* for the protein pair (*Pi, Pj)* in the organism, respectively.

2.2.7 Semantic Similarity (SS) Score

Later in the results, we characterized IAS at the GO term level and compared IAS of GO term pairs (GO_IAS) with a similar existing score designed by Schlicker *et al*. [42]. This score called the semantic similarity (SS) score measures the similarity of two GO terms, *c1* and *c2*, by using the commonality information in terms of common ancestors of the two GO terms in a given GO term pair. It is computed as follows:

$$sim(c1,c2) = \max_{c \in S(c1,c2)} \left(\frac{2\log(p(c))}{\log p(c1) + \log p(c2)} * (1 - p(c))\right) \quad (13)$$

where *S(c1,c2)* is the set of common ancestors of GO terms *c1* and *c2,* GO term *c* is one of the ancestral term of GO terms *c1* and *c2,* and p*(c)* is the frequency of occurrences of GO term *c* in the GOA annotation database.

2.2.8 Evaluation of PPI Prediction

For testing each of IAS, CAS, and PAS, we generated test data by combining TP and TN interactions for each organism and ordering them in the descending order of their scores, to compute the True Positive Rate (TPR) and False Negative Rate (FNR) for ROC curve.

TABLE 1
STATISTICS OF INTERACTIONS

Organism	Interaction Type	Number of Interactions	Number of Proteins
Saccharomyces Cerevisiae	All	143779	5337
	Physical	49271	5030
	Genetic	94508	4887
Schizosaccharomyces Pombe	All	50876	3916
	Physical	43344	2481
	Genetic	7532	3001
Drosophila Melanogaster	All	37130	7992
	Physical	34822	7824
	Genetic	2308	955
Caenorhabditis Elegans	All	7780	3813
	Physical	5542	3123
	Genetic	2238	1110
Homo sapiens	All	19578	7440
	Physical	19472	7414
	Genetic	106	117
Arabidopsis Thaliana	All	15143	6602
	Physical	15064	6596
	Genetic	79	66

3. RESULTS

3.1 Characteristics of IAS

In this section, we characterize IAS (Eqn. 1). The GO database used in this study contains 12,835 Biological Process (BP), 4,443 Molecular Function (MF) and 1,783 Cellular Component (CC) terms, resulting in a total of 19,061 terms. Among 181,670,391 possible GO term pairs, 13,945,909 (7.7%) contain non-zero values of IAS.

Fig. 1 shows the distribution of IAS. The maximum, median, and minimum scores are 11845.9, 19.48, and 0.0052, respectively. One example among the 307 GO term pairs that have the highest IAS in the distribution is GO:0035718 *macrophage migration inhibitory factor binding* and GO:0019883 *antigen processing and presentation of endogenous antigen*. The first term is in MF and the second in BP. A protein annotated with both of these terms is HG2A_Human (UniProt ID P04233), which plays a critical role in antigen processing and serves as a cell surface receptor for cytokine, a macrophage migration inhibitory factor.

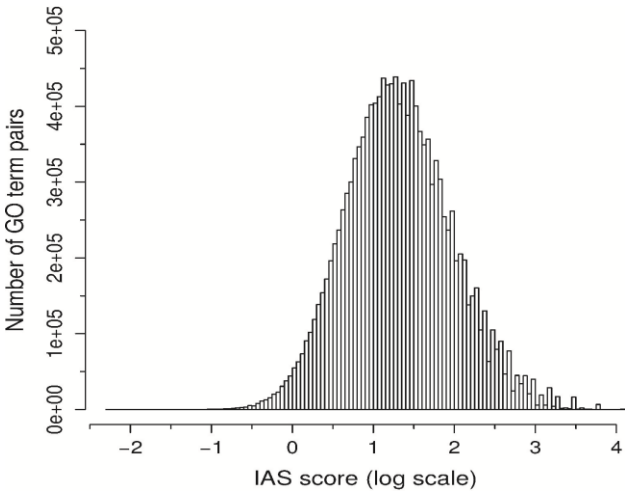


Fig. 1. Distribution of IAS for all GO term pairs.

IAS can be defined not only between terms of the same GO category but also across different categories. Among 1,39,45,909 GO term pairs with non-zero IAS score, 67,02,934 (48.1%) pairs fall in BP, 4,45,769 (3.2%) in MF and 2,46,637 (1.8%) in CC. The rest of them are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

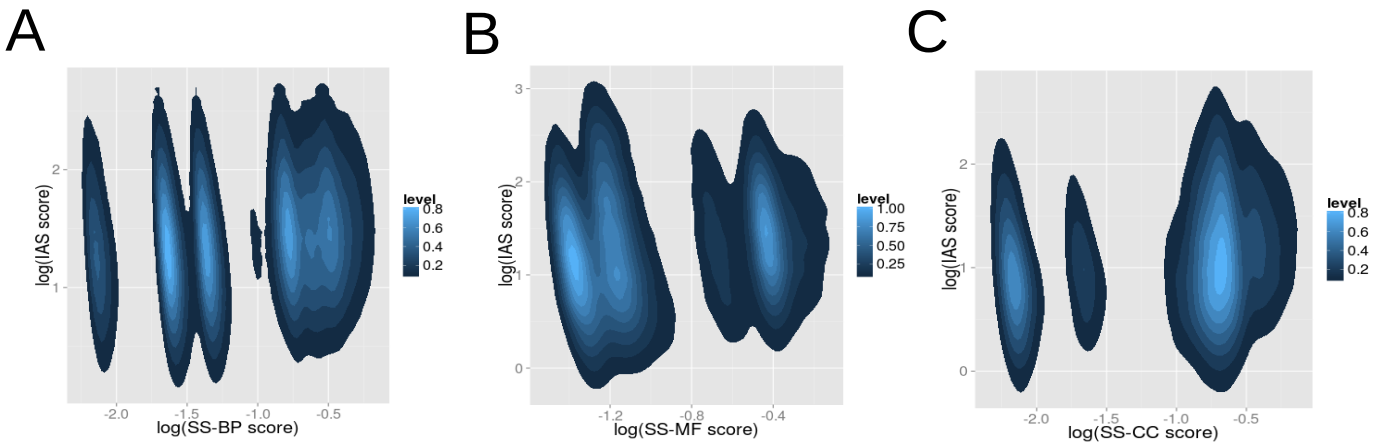


Fig. 2. Comparison of IAS score with Semantic Similarity (SS) score A. BP domain; B. MF domain; C. CC domain.

cross-domain pairs: 3,341,961 (23.9%) are across BP-MF, 2,435,331 (17.5%) across BP-CC, and 673,296 (4.83%) are across MF-CC. IAS scores in the same domain and cross-domain show similar distribution.

Next, in Fig. 2 we compare IAS with the semantic similarity (SS) score, which reflects the depth of the common ancestor of a GO term pair in the GO hierarchy and the number of gene products annotated with it in the database [44]. Since SS score can only capture GO association within the same domain, comparison is made separately for three different GO domains (BP, MF, and CC in Fig. 2A-C). Only the GO pairs that have non-zero IAS and SS are plotted. Overall, IAS does not show significant correlation with SS (with correlation coefficients of 0.1183, 0.1096, and 0.1691 for IAS against SS-BP, SS-MF, and SS-CC respectively), indicating that IAS captures very different information from SS. There are a large number of GO term pairs where both IAS and SS are consistently high and close to the maximum (e.g., GO:0015821 *methionine transport* and GO:0000101 *sulfur amino acid transport*) or low and close to zero (e.g., GO:*0000001 mitochondrion inheritance* and GO:0007155 *cell adhesion*). On the other hand, there are cases where IAS is high but SS is low (e.g., GO:0003213 *cardiac right atrium morphogenesis* and GO:1901201 *regulation of extracellular matrix assembly*) and cases with the opposite scenario (e.g., GO:0060322 *head development* and GO:0060325 *face morphogenesis*).

In Table 2 we show examples of GO pairs in the same domain, which have a large score either in IAS or in SS. The first five examples are cases of GO pairs that have a high IAS but a low SS score. The first example is GO:0019058 *viral life cycle* and GO:0000279 *M phase*, both in BP. These two GO terms have two common ancestral GO terms in the GO hierarchy, GO:0008150 *biological process* (depth 0) and GO:0009987 *cellular process* (depth 1). Since the lowest common ancestor of this pairs is too shallow (i.e. general) in the GO hierarchy, the SS score for this pair is as low as 0.0091. On the other hand, in PPI, there are 49 PPIs with these GO terms appearing in interacting proteins. An example of one of these interactions is between protein ILF3_Human (Q12906) and RL5_Human (P46777). Q21906 is involved in the M-phase of the cell cycle, a phase where nuclear division occurs [45]. P46777 is a rRNA maturation protein that is involved in viral mRNA translation [46]. Because of the high number of such PPI interactions, IAS score of this GO pair is very high (2398.55).

The second example is a BP GO pair, GO:0034085 *establishment of sister chromatid cohesion* and GO:0000727 *double-strand break repair via break induced replication*. The lowest common ancestor of these two terms in the GO hierarchy is GO:0044699 *single-organism process* at the depth 2, hence the low SS score 0.0382. In the PPI network, there are 30 interactions where the interacting partners hold GO:0000727 or GO:0000727. One such interaction is between protein CTF4_Yeast (Q01454) and PSF2_Yeast (P40359). The first protein, Q01454, functions as an accessory factor in DNA replication and has a role in duplicating the genome in vivo. It has both the GO terms [47, 48]. The second protein P40359 plays an essential role in the initiation of DNA replication by binding to DNA replication origins [48] and has the second GO term GO:0000727 *double-strand break repair via break induced replication* in the pair in question. Due to such PPI pairs, IAS gets a high value of 3290.53 for this GO pair.

The third example is GO pair GO:0071033 *nuclear retention of pre-mRNA at the site of transcription* and GO:0000973 *posttranscriptional tethering of RNA polymerase II gene DNA at nuclear periphery*. Similar to the first example, the low SS is because they share only a very general term as their common ancestor (GO:0009987 *cellular process*, at the depth of 1). In PPI, this pair has 51 interactions where interacting proteins hold one or both of the GO terms. One such interaction is between proteins RRP6_Yeast (Q12149) and NUP42_Yeast (P49686). The first protein Q12149 has both GO terms and functions as nuclear-specific catalytic component of the RNA exosome complex and participates in various cellular RNA processing and degradation events [49, 50]. The second protein (P49686) has the second GO term GO:0000973 *posttran-*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

scriptional tethering of RNA polymerase II gene DNA at nuclear periphery and functions as a component of the nuclear pore complex (NPC) that can play the role of docking of interaction partners for transiently associated nuclear transport factors [51]. A number of such PPI interactions result in a high IAS score of 3775.88 for this GO pair.

The fourth example is a MF pair GO:0009884 *cytokine receptor activity* and GO:0043424 *protein histidine kinase binding*. SS is 0.0 for this pair because these GO terms do not share common ancestor up to the root term of MF, GO:0003674 *molecular function.* But has a high IAS because the number of interacting proteins with the GO terms is relatively high, 25. An example is AHK3_ARATH (Q9C5U1) and Y5436_ARATH (Q8RY18). The first protein is a cytokines receptor that functions as a histidine kinase [52, 53]. The second protein is a meprin and TRAF-homology domain containing protein that has GO:0043424 *protein histidine kinase binding*.

The fifth example is a CC pair GO:0034515 *proteasome storage granule* and GO:0042175 *nuclear outer membrane-endoplasmic reticulum membrane network*. This GO pair has 170 PPIs. One such interaction is between RPN1_Yeast (P38764) and PSA7_Yeast (P21242). The first protein P38764 is a proteasome regulatory subunit and has the GO term GO:0034515 *proteasome storage granule* [54]. The second protein is a probable proteasome subunit alpha type-7 and has the GO term GO:0042175 *nuclear outer membrane-endoplasmic reticulum membrane network* [55].

The next five examples in Table 2 illustrate cases where IAS is low while the SS score is high. The first of these examples is a pair of BP GO terms GO:*0000819 sister chromatid segregation* and GO:0000070 *mitotic sister chromatid segregation*. Since the second term is a direct descendant of the first term in the GO hierarchy, the similarity of these two terms (the SS score) is very high (0.9946). On the other hand, in the PPI network, there is only one interaction between proteins with these two GO terms despite of not very small number of proteins with the two GO terms (10 proteins with the first GO term and 70 proteins with the second GO term), which made IAS very low (3.48).

The rest of the four examples essentially have the same situation as the sixth example. Two GO terms listed for each example share a high functional similarity because the common ancestral term is deep in the GO hierarchy. On the other hand, IAS is low because the number of interacting protein pairs with the two GO terms is small relative to individual proteins that have one of the GO terms (Eqn. 1). The seventh example is BP pair GO:0001843 *neural tube closure* and GO:0016331 *morphogenesis of embryonic epithelium.* The SS score is high because the lowest common ancestor between them is the second term itself GO:0016331 *morphogenesis of embryonic epithelium* at the depth of 8. On the other hand, IAS is low because although there are 127 proteins with the first GO term and 37 proteins with the second GO term, only two protein pairs interact among them.

The eighth example is BP pair GO:0006417 *regulation of translation* and GO:0010608 *posttranscriptional regulation of gene expression*. The high SS core is due to their lowest common ancestor, GO:0010467 *gene expression* at the depth of six in the GO hierarchy, while in the PPI database, there are only two PPIs among combinations of 215 proteins with the first GO term and 14 proteins with the second GO term.

The ninth example is MF pair GO:0019829 *cation-transporting ATPase activity* and GO:0042626 *ATPase activity, coupled to transmembrane movement of substances*. Obviously, these two GO terms are closely related. However, only two PPIs were observed in the database, although 61 and 92 proteins are annotated with the first or the second GO terms, respectively.

The final example, CC pair GO:0001669 *acrosomal vesicle* and GO:0002080 *acrosomal membrane*, have a high SS of 0.9281 due to their lowest common ancestor at the depth 10 (GO:0030141 *secretory granule*). But the IAS is low because of only one interacting protein pair among 88 proteins that have the first GO term and 22 proteins with the second GO term.

TABLE 2
EXAMPLES OF IAS SCORES THAT ARE DIFFERENT FROM SS SCORES

GO ID 1	Description	Domain	GO ID 2	Description	Domain	IAS	SS
GO:0019058	Viral life cycle	BP	GO:0000279	M phase	BP	2398.55	0.0091
GO:0034085	Establishment of sister chromatid cohesion	BP	GO:0000727	Double-strand break repair via break induced replication	BP	3290.53	0.0382
GO:0071033	Nuclear retention of pre-mRNA at the site of transcription	BP	GO:0000973	Posttranscriptional tethering of RNA polymerase II gene DNA at nuclear periphery	BP	3775.88	0.0091
GO:0009884	Cytokine receptor activity	MF	GO:0043424	Protein histidine kinase binding	MF	2350.38	0.00

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

GO:0034515	Proteasome storage granule	CC	GO:0042175	Nuclear outer membrane-endoplasmic reticulum membrane network	CC	1202.98	0.0078
GO:0000819	Sister chromatid segregation	BP	GO:0000070	Mitotic sister chromatid segregation	BP	3.48	0.9946
GO:0001843	Neutral tube closure	BP	GO:0016331	Morphogenesis of embryonic epithelium	BP	2.52	0.9237
GO:0006417	Regulation of translation	BP	GO:0010608	Posttranscriptional regulation of gene expression	BP	3.94	0.9716
GO:0019829	Cation-transporting ATPase activity	MF	GO:0042626	ATPase activity, coupled to transmembrane movement of substances	MF	2.11	0.9561
GO:0001669	Acrosomal vesicle	CC	GO:0002080	Acrosomal membrane	CC	3.06	0.9281

First five are the cases where IAS is high and SS is low. Next five are cases where IAS is low and SS is high.

To summarize these examples, differences of IAS and SS reflect observations that interacting proteins do not necessarily have similar functions (i.e. high IAS and low SS) and proteins with similar functions do not always physically interact (i.e. low IAS and high SS).

We also compared IAS with CAS and PAS, which are GO term association scores that can also capture cross-domain associations. Fig. 3A and 3B show result of comparison for IAS with CAS and PAS, respectively. Only GO pairs that have non-zero score for both IAS and CAS/PAS were used in this analysis. Among 13,945,609 GO term pairs with non-zero IAS, 1,549,864 (11.1%) have non-zero CAS and 1,480,407 (10.6%) have non-zero PAS. Overall, both CAS and PAS show moderate correlation with IAS with correlation coefficient 0.5621 for IAS-CAS and 0.4220 for IAS-PAS. An example of a low scoring GO pair, both in IAS and CAS is GO:0000310 *xanthine phosphoribosyltransferase activity* in MF and GO:0032265 *xanthosine monophosphate (XMP) salvage* in BP. An example of GO pairs with a very low score both in IAS and CAS is GO:0000001 *mitochondrion inheritance* in BP and GO:0005773 *vacuole* in CC.

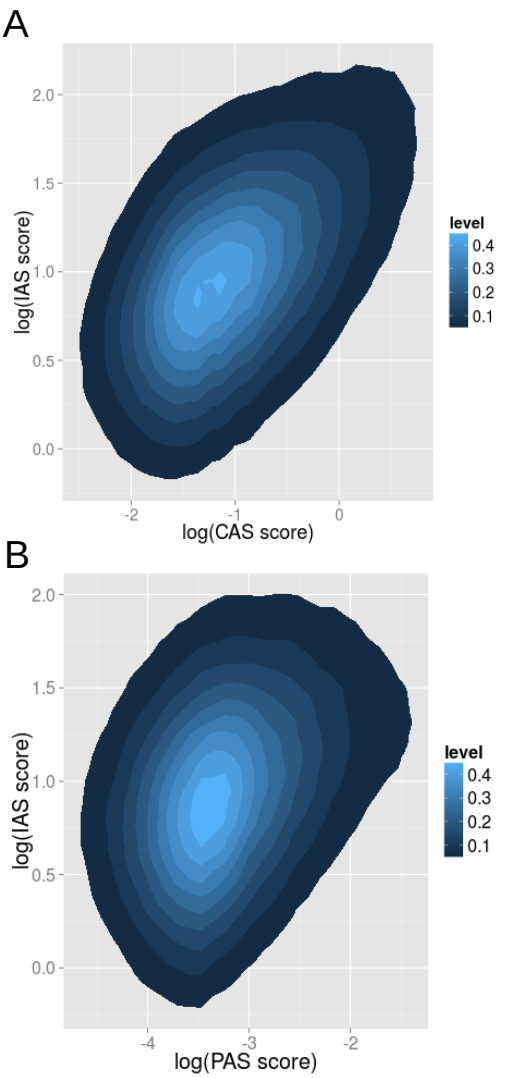


Fig. 3. Comparison of IAS score. A. IAS against CAS; B. IAS against PAS.

3.2 Prediction of interacting proteins using IAS

We tested the practical performance of IAS in predicting PPIs in organisms. The benchmark dataset consists of 274,286 interactions between 35,100 proteins from six organisms as shown in Table 1. For each organism, the PPI_IAS score (Eqn. 3) was computed for all pairs of proteins and the protein pairs were sorted by the descendant order of their score. Each time the PPI_IAS score was computed for a protein pair, the two proteins and their GO annotations were removed from the statistics and the IAS scores were re-computed. The prediction performance was evaluated by the Area Under the Curve (AUC) of the Receiver Operator Characteristics (ROC).

Table 3 summarizes the PPI prediction results in comparison with CAS and PAS. Obviously, all IAS, PAS, and CAS (more precisely, PPI_IAS, PPI_PAS, and PPI_CAS) performed substantially better than random (which has an AUC of 0.5). AUC of IAS for "All" inter-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

actions (i.e. physical and genetic) of the six organisms ranged from 0.6692 (*D. melanogaster*) to 0.7230 (*S. cerevisiae*). Since both physical and genetic interactions in BIOGRID were used to compute IAS, both physical interactions and genetic interactions were predicted with similar accuracy by IAS. Comparing with CAS and PAS, overall IAS performed better than these two scores. Considering "all" the interactions of the six organisms, IAS performed better than CAS and PAS for four organisms while PAS was the best in the rest of the two organisms (*H. sapiens* and *A. thaliana*) with small margins to IAS. Considering three types of interactions from the six organisms (thus 18 interactions), IAS was best for 11 interactions while PAS was best for the rest of the seven interactions. Figure 4 shows ROC curves of IAS, CAS, and PAS for "all" interactions of four organisms, *S. cerevisiae*, *D. melanogaster*, *H. Sapiens,* and *A. thaliana*. For the *S. cerevisiae* (Fig. 4A), IAS clearly outperformed (AUC: 0.7230) while CAS and PAS performed in this order with similar AUC values (0.6931 and 0.6871, respectively). For *D. melanogaster* (Figure 4B), IAS performed marginally better than the other two scores (IAS: 0.6692, CAS: 0.6614, PAS: 0.6610). In the last two panels (Figure 4C & D) PAS performed the best, although the difference of the three scores is not very clear on the ROC curves since their AUC values are similar (human: IAS: 0.7081, CAS: 0.7025, PAS: 0.7184; *A. thaliana*: IAS: 0.7190, CAS: 0.7095, PAS: 0.7227). Note that in our previous work [43], we compared performance of CAS and PAS in PPI prediction with five other existing GO term-based scores, namely, Funsim, BP-Funsim [56], Chagoyen [57], Pandey [58, 59], and Funsim scores and showed that CAS and PAS outperformed those scores. The main difference of CAS and PAS against those five scores was that the four scores quantify similarity of GO terms while CAS and PAS quantify "coherence" of GO terms by counting associations, i.e. co-occurrence of GO terms. Here Table 3 further shows IAS, where GO term associations were taken from interacting proteins, performs better than CAS and PAS in PPI prediction.

TABLE 3 AUC VALUES FOR DIFFERENT SCORE CATEGORIES

Organism	Interaction Type	IAS	PAS	CAS	Average Z-score	Average Rank
Saccharomyces cerevisiae	All	0.7230	0.6872	0.6931	0.7307	0.7304
	Physical	0.7633	0.7520	0.7580	0.7754	0.7857
	Genetic	0.6925	0.6510	0.6560	0.6994	0.6946
Schizosaccharomyces pombe	All	0.6882	0.5861	0.6037	0.6870	0.6447
	Physical	0.8420	0.6789	0.7176	0.8459	0.7862
	Genetic	0.6474	0.5658	0.5794	0.6442	0.6121
Drosophila melanogaster	All	0.6692	0.6610	0.6614	0.6768	0.6800
	Physical	0.6451	0.6396	0.6388	0.6525	0.6541
	Genetic	0.7492	0.8049	0.7830	0.7874	0.8112
Caenorhabditis elegans	All	0.7155	0.6923	0.6715	0.7176	0.7158
	Physical	0.6489	0.6623	0.6313	0.6521	0.6541
	Genetic	0.8409	0.7439	0.7594	0.8310	0.8112
Homo sapiens	All	0.7081	0.7184	0.7025	0.7163	0.7397
	Physical	0.7047	0.7177	0.7031	0.7153	0.7382
	Genetic	0.6317	0.5779	0.5842	0.6106	0.5633
Arabidopsis thaliana	All	0.7190	0.7227	0.7095	0.7308	0.7406
	Physical	0.7160	0.7219	0.7071	0.7273	0.7384
	Genetic	0.7294	0.7951	0.7611	0.8087	0.8120

In Table 4, we show examples of protein pairs where PPI_IAS exhibits contrasting results from the Funsim score, which is a score given to a protein pair by applying SS (Eqn. 13) to Eqn. 5 instead of IAS. Two examples each from yeast and human are selected. The first example is a protein pair, P50945 and P40341. P50945 is a component of a large protein complex of mitochondrial inner membrane that plays a crucial role

in the maintenance of inner membrane architecture. The second protein P40341 is a mitochondrial respiratory chain complex's assembly protein that is involved in the degradation of non-assembled mitochondrial inner membrane proteins. This is an example that has a high IAS and a low Funsim score. The high PPI_IAS between the pair originates from high IAS between GO pairs across the two proteins. One such GO pair is a CC GO terms pair, GO:0061617 (MICOS complex) from the first protein and GO:0005743 (mitochondrial inner membrane) from the second protein. This GO pair has a high IAS of 133.44 due to 83 PPIs in BIOGRID among a combination of 5 proteins that have the first and 737 proteins that have the second GO term. However, SS between the same pair of GO terms is low because the common ancestor of the two terms is GO:004446 (intracellular organelle part) which is a relatively general GO term (depth 5). A similar scenario between the other GO pairs for this protein pair results in a high IAS and low Funsim score.

The second one is an opposite case in yeast, which has a low PPI-IAS and a high Funsim (P25383 and Q99303). Both proteins are capsid proteins and are the structural components of the virus-like particle. The almost identical functionalities of the two proteins lead to the high Funsim score of 0.8115. However, even the GO pair across the two proteins with the highest IAS, GO:0000943 (retrotransposon nucleocapsid) and GO:0032197 (transposition, RNA-mediated), has as low IAS 12.86, because there are only 17 PPIs among a combination of 87 proteins that has the first GO term and 90 proteins that has the second GO term.

The third example in Table 4 is P52333 and P42229 from human. The first protein is a kinase that phosphorylates STAT protein. The second protein is STAT protein, which carries out a dual function: signal transduction and activation of transcription. Among over 2000 GO pairs across these two proteins, there are many that have very high IAS, which led to the high IAS between this protein pair. One such GO pair is GO:0004715 (non-membrane spanning protein tyrosine kinase activity) in MF and GO:0060397 (JAK-STAT cascade involved in growth hormone signaling pathway) in BP, which has IAS of 213, which reflects the fact that there are 102 PPIs in BIOGRID among a combination of 109 proteins that have the first GO term and 26 proteins with the second GO term. However, the semantic similarity score between these two GO terms is 0, since they are from different categories. Due to many such GO-pairs, this protein pair has a high IAS sand a low Funsim score.

The last example is the Q96SU4 and Q9BXB4 from human. Both are oxysterol-binding protein (OSBP)-related proteins, hence the pair has a large Funsim score of 0.9912. However, even the GO pair across these two proteins that have the highest IAS is BP terms GO:0006869 (lipid transport) and GO:0010890 (positive regulation of sequestering of triglyceride), has a small IAS of 10.84, because there are only one PPI among a combination of 78 proteins with the first GO term and 7 proteins with the second GO term. The low IAS among such GO pairs results in the low PPI_IAS for this protein pair. Similar examples from all six species can be found at Supplemental Material.

TABLE 4 Comparison of Protein-pair IAS and Funsim Scores

Organism	Protein1	Function	Protein2	Function	IAS score	Funsim score
Saccharomyces cerevisiae	P50945	MICOS complex subunit	P40341	Mitochondrial respiratory chain complex assembly protein	3653.25	0.2048
Saccharomyces cerevisiae	P25383	Transposon Ty2-C Gag polyprotein	Q99303	Transposon Ty2-DR3 Gag polyprotein	9.9688	0.8115
Homo sapiens	P52333	Tyrosine-protein kinase JAK3	P42229	STAT 5A	583.2310	0.2741
Homo sapiens	Q96SU4	Oxysterol-binding protein-related protein 9	Q9BXB4	Oxysterol-binding protein-related protein 11	12.9753	0.9912

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2476809, IEEE/ACM Transactions on Computational Biology and Bioinformatics

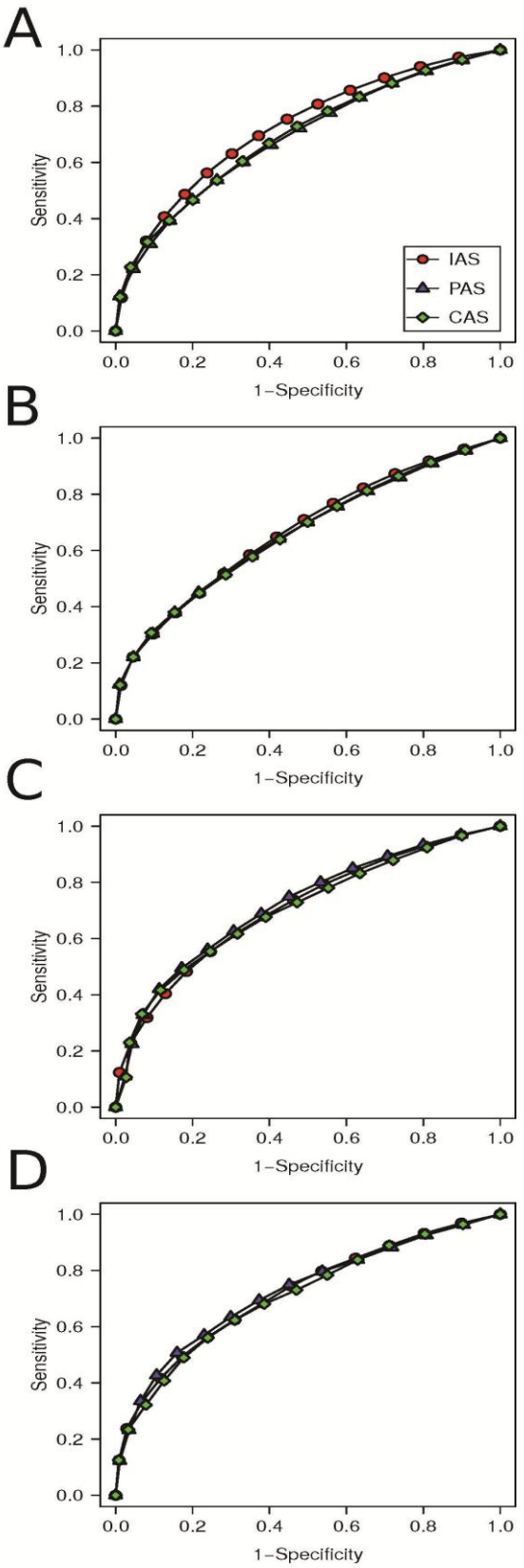


Fig. 4. ROC curves of IAS, PAS, and CAS. A, *S. cerevisiae;* B, *D. melanogaster;* C, *H. Sapiens;* D, *A. thaliana.*

3.3. Consensus methods

Next, we designed two consensus methods that combine IAS, CAS, and PAS. The first method, named the Average Z-score method (Eqn. 11), computes the average of Z-scores of the three scores, IAS, CAS, and PAS, for each protein pair. The second method, the average rank score (Eqn. 12), computes the average rank by the three scores for each protein pair. Table 3 shows that both methods outperformed the individual scores, IAS, CAS, and PAS. The average Z-score method showed a higher AUC value than all of the three individual scores for 14 interactions out of 18 interactions. On the other hand, the average rank score was better than the individual scores for 16 interactions. Finally, a head-to-head comparison between the average Z-score method and the average rank score shows that the average rank score had a higher AUC value for 10 interactions out of 18. It is noteworthy that the two consensus methods overall outperform all individual methods as often consensus methods just show performance that is near the average of component methods.

4. DISCUSSION

PPIs and protein function are intertwined; interacting proteins tend to share common functions and conversely, functionally related proteins are more likely to interact with each other. Thus, from a bioinformatics point of view, PPI network can be a source of protein function prediction while functional relationship of proteins can be used for verifying experimentally detected PPIs. The same relationship to protein function is also observed between other omics data, including gene expression patterns and phylogenetic profiles.

In this work, we developed IAS, a novel GO term-based score for predicting PPIs. To make IAS specific for PPI prediction, we mined GO term pair associations directly from interacting protein pairs. Consequently, IAS captured new relationships of GO terms, which is very different from conventional functional similarity. Moreover, unlike functional similarity of GO terms, such as the SS score, IAS can be defined also for GO term pairs that are from different categories. The method of computing IAS, i.e. counting observed GO pairs followed by normalization with an expected number of such counts, is essentially the same as statistical atom-atom or residue-residue contact potentials [60, 61], which are very successful in the protein structure prediction field. This simple yet powerful approach could also be applied to prediction of other behaviors of proteins and genes, such as co-expression of genes from GO term annotations.

5. CONCLUSIONS

We have developed a novel score of GO terms named IAS for predicting interacting proteins. Unlike existing works which consider functional similarity to predict PPI, IAS quantifies associations of GO terms that are frequently observed in known interacting proteins. IAS performed better in predicting PPIs than existing GO term-based scores. Moreover, consensus methods further improved the accuracy of PPI prediction.

6. ADDITIONAL FILES

IAS scores of GO term pairs are made available at http://kiharalab.org/IAS/. Moreover, 60 examples in accordance with Table 4 are provided as Supplemental Material. IAS score of proteins can be computed at http://kiharalab.org/compare.php.

7. ACKNOWLEDGEMENTS

Meghana Chitale has contributed to an early stage of this work. The authors are grateful to Lyman Monroe for proofreading the manuscript.
Funding: This work is supported partly by the National Institutes of Health (R01GM097528), the National Science Foundation (IIS1319551, DBI1262189, IOS1127027).

8. REFERENCES

1. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *ProcNatlAcadSciUSA* 2001, **98**(8):4569-4574.
2. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A *et al*: **Large-scale identification of protein-protein interaction of Escherichia coli K-12**. *Genome research* 2006, **16**(5):686-691.
3. Aryal UK, Xiong Y, McBride Z, Kihara D, Xie J, Hall MC, Szymanski DB: **A proteomic strategy for global analysis of plant protein complexes**. *The Plant cell* 2014, **26**(10):3867-3882.
4. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U *et al*: **The IntAct molecular interaction database in 2012**. *Nucleic acids research* 2012, **40**(Database issue):D841-846.
5. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32 Database issue**:D449-D451.
6. Otsuka Y, Muto A, Takeuchi R, Okada C, Ishikawa M, Nakamura K, Yamamoto N, Dose H, Nakahigashi K, Tanishima S *et al*: **GenoBase: comprehensive resource database of Escherichia coli K-12**. *Nucleic acids research* 2014:D606-D617.
7. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T: **Assessment of prediction accuracy of protein function from protein--protein interaction data**. *Yeast* 2001, **18**(6):523-531.
8. Brun C, Herrmann C, Guenoche A: **Clustering proteins from interaction networks for the prediction of cellular functions**. *BMCBioinformatics* 2004, **5**:95.
9. Chua HN, Sung WK, Wong L: **Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions**. *Bioinformatics* 2006, **22**(13):1623-1630.
10. Letovsky S, Kasif S: **Predicting protein function from protein/protein interaction data: a probabilistic approach**. *Bioinformatics* 2003, **19 Suppl 1**:i197-204.
11. Deng M, Zhang K, Mehta S, Chen T, Sun F: **Prediction of protein function using protein-protein interaction data**. *JComputBiol* 2003, **10**(6):947.
12. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks**. *NatBiotechnol* 2003, **21**(6):697-700.
13. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M: **Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps**. *Bioinformatics* 2005, **21 Suppl 1**:i302-i310.
14. Hawkins T, Chitale M, Kihara D: **New paradigm in protein function prediction for large scale omics analysis**. *Mol Biosyst* 2008, **4**(3):223-231.
15. Hawkins T, Kihara D: **Function prediction of uncharacterized proteins**. *JBioinformComputBiol* 2007, **5**(1):1-30.
16. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.
17. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes**. *Trends in genetics : TIG* 2002, **18**(10):529-536.
18. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A *et al*: **Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae**. *Journal of biology* 2006, **5**(4):11.
19. Hulovatyy Y, Solava RW, Milenkovic T: **Revealing missing parts of the interactome via link prediction**. *PloS one* 2014, **9**(3):e90073.
20. Sarac OS, Pancaldi V, Bahler J, Beyer A: **Topology of functional networks predicts physical binding of proteins**. *Bioinformatics* 2012, **28**(16):2137-2145.
21. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: **Predicting protein-protein interactions based only on sequences information**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(11):4337-4341.
22. Martin S, Roe D, Faulon JL: **Predicting protein-protein interactions using signature products**. *Bioinformatics* 2005, **21**(2):218-226.

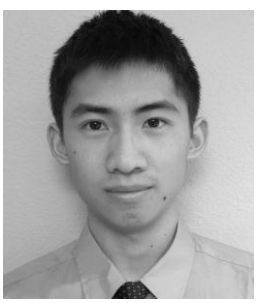
23. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N *et al*: **PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs**. *BMC bioinformatics* 2006, **7**:365.
24. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A: **PPIevo: protein-protein interaction prediction from PSSM based evolutionary information**. *Genomics* 2013, **102**(4):237-242.
25. Ben-Hur A, Noble WS: **Kernel methods for predicting protein-protein interactions**. *Bioinformatics* 2005, **21 Suppl 1**:i38-46.
26. Bock JR, Gough DA: **Predicting protein--protein interactions from primary structure**. *Bioinformatics* 2001, **17**(5):455-460.
27. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends BiochemSci* 1998, **23**(9):324-328.
28. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *ProcNatlAcadSciUSA* 1999, **96**(8):4285-4288.
29. Snel B, Bork P, Huynen MA: **The identification of functional modules from the genomic association of genes**. *ProcNatlAcadSciUSA* 2002, **99**(9):5890-5895.
30. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al*: **STRING v10: protein-protein interaction networks, integrated over the tree of life**. *Nucleic acids research* 2015, **43**(Database issue):D447-452.
31. Pazos F, Juan D, Izarzugaza JM, Leon E, Valencia A: **Prediction of protein interaction based on similarity of phylogenetic trees**. *Methods Mol Biol* 2008, **484**:523-535.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.
33. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T *et al*: **Structure-based prediction of protein-protein interactions on a genome-wide scale**. *Nature* 2012, **490**(7421):556-560.
34. Lu L, Lu H, Skolnick J: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading**. *Proteins* 2002, **49**(3):350-364.
35. Tuncbag N, Gursoy A, Nussinov R, Keskin O: **Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM**. *Nature protocols* 2011, **6**(9):1341-1354.
36. Wass MN, Fuentes G, Pons C, Pazos F, Valencia A: **Towards the prediction of protein interaction partners using physical docking**. *Molecular systems biology* 2011, **7**:469.
37. van Haagen HH, t Hoen PA, de Morree A, van Roon-Mom WM, Peters DJ, Roos M, Mons B, van Ommen GJ, Schuemie MJ: **In silico discovery and experimental validation of new protein-protein interactions**. *Proteomics* 2011, **11**(5):843-853.
38. Consortium GO: **Gene Ontology Consortium: going forward**. *Nucleic acids research* 2015, **43**(Database issue):D1049-1056.
39. Wu X, Zhu L, Guo J, Fu C, Zhou H, Dong D, Li Z, Zhang DY, Lin K: **SPIDer: Saccharomyces protein-protein interaction database**. *BMC bioinformatics* 2006, **7 Suppl 5**:S16.
40. Wu X, Zhu L, Guo J, Zhang DY, Lin K: **Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations**. *Nucleic acids research* 2006, **34**(7):2137-2150.
41. Maetschke SR, Simonsen M, Davis MJ, Ragan MA: **Gene Ontology-driven inference of protein-protein interactions using inducers**. *Bioinformatics* 2012, **28**(1):69-75.
42. Lin N, Wu B, Jansen R, Gerstein M, Zhao H: **Information assessment on predicting protein-protein interactions**. *BMC bioinformatics* 2004, **5**:154.
43. Chitale M, Palakodety S, Kihara D: **Quantification of protein group coherence and pathway assignment using functional association**. *BMC bioinformatics* 2011, **12**:373.
44. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology**. *BMCBioinformatics* 2006, **7**:302.
45. Matsumoto-Taniura N, Pirollet F, Monroe R, Gerace L, Westendorf JM: **Identification of novel M phase phosphoproteins by expression cloning**. *Molecular biology of the cell* 1996, **7**(9):1455-1469.
46. Goodman AG, Smith JA, Balachandran S, Perwitasari O, Proll SC, Thomas MJ, Korth MJ, Barber GN, Schiff LA, Katze MG: **The cellular protein P58IPK regulates influenza virus mRNA translation and replication through a PKR-mediated mechanism**. *Journal of virology* 2007, **81**(5):2221-2230.
47. Borges V, Smith DJ, Whitehouse I, Uhlmann F: **An Eco1-independent sister chromatid cohesion establishment pathway in S. cerevisiae**. *Chromosoma* 2013, **122**(1-2):121-134.
48. Lydeard JR, Lipkin-Moore Z, Sheu YJ, Stillman B, Burgers PM, Haber JE: **Break-induced replication requires all essential DNA replication factors except those specific for pre-RC assembly**. *Genes & development* 2010, **24**(11):1133-1144.
49. Rougemaille M, Gudipati RK, Olesen JR, Thomsen R, Seraphin B, Libri D, Jensen TH: **Dissecting mechanisms of nuclear mRNA surveillance in THO/sub2 complex mutants**. *The EMBO journal*

2007, **26**(9):2317-2326.

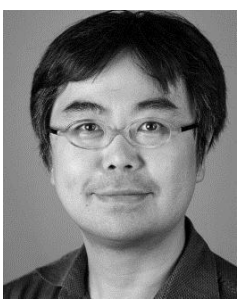
50. Vodala S, Abruzzi KC, Rosbash M: **The nuclear exosome and adenylation regulate posttranscriptional tethering of yeast GAL genes to the nuclear periphery**. *Molecular cell* 2008, **31**(1):104-113.
51. Light WH, Brickner DG, Brand VR, Brickner JH: **Interaction of a DNA zip code with the nuclear pore complex promotes H2A.Z incorporation and INO1 transcriptional memory**. *Molecular cell* 2010, **40**(1):112-125.
52. Franco-Zorrilla JM, Martin AC, Leyva A, Paz-Ares J: **Interaction between phosphate-starvation, sugar, and cytokinin signaling in Arabidopsis and the roles of cytokinin receptors CRE1/AHK4 and AHK3**. *Plant physiology* 2005, **138**(2):847-857.
53. Dortay H, Mehnert N, Burkle L, Schmulling T, Heyl A: **Analysis of protein interactions within the cytokinin-signaling pathway of Arabidopsis thaliana**. *The FEBS journal* 2006, **273**(20):4631-4644.
54. Laporte D, Salin B, Daignan-Fornier B, Sagot I: **Reversible cytoplasmic localization of the proteasome in quiescent yeast cells**. *The Journal of cell biology* 2008, **181**(5):737-745.
55. Groll M, Ditzel L, Lowe J, Stock D, Bochtler M, Bartunik HD, Huber R: **Structure of 20S proteasome from yeast at 2.4 A resolution**. *Nature* 1997, **386**(6624):463-471.
56. Hawkins T, Chitale M, Luban S, Kihara D: **PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data**. *Proteins* 2009, **74**(3):566-582.
57. Chagoyen M, Carazo JM, Pascual-Montano A: **Assessment of protein set coherence using functional annotations**. *BMC bioinformatics* 2008, **9**:444.
58. Pandey J, Koyuturk M, Subramaniam S, Grama A: **Functional coherence in domain interaction networks**. *Bioinformatics* 2008, **24**(16):i28-34.
59. Pandey J, Koyuturk M, Grama A: **Functional characterization and topological modularity of molecular interaction networks**. *BMC bioinformatics* 2010, **11 Suppl 1**:S35.
60. Yang YD, Park C, Kihara D: **Threading without optimizing weighting factors for scoring function**. *Proteins* 2008, **73**(3):581-596.
61. Skolnick J: **In quest of an empirical potential for protein structure prediction**. *CurrOpinStructBiol* 2006, **16**(2):166.

Satwica Yerneni received her BTech degree in Bioinformatics from Vellore Institute of Technology, India in 2011 and MS degree in Bioinformatics from Indiana University Bloomington, IN, USA in 2013. She is currently an Informatics Specialist in the Clinical Genome Sequencing Lab at Mayo Clinic, Rochester, MN. She also worked as a Genomics Systems Specialist in the DNA Sequencing and Genotyping Core at Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio. She worked as an intern in 2011 at Kihara Laboratory, Purdue University, IN. Her area of research is Bioinformatics and Computational Biology.

Ishita K. Khan received her BS degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET) in 2009. She is currently working toward the PhD degree in Computer Science at Purdue University, West Lafayette, IN, USA. Her research interest is in Bioinformatics/Computational Biology. She works in development of computational protein function prediction methods. She spent the summer of 2014 at Qatar Computing Research Institute (QCRI) as a doctoral research intern.



Qing Wei received his BS degree in computer science from Purdue University, West Lafayette, IN, in 2014. He is currently working toward the MS degree in computer science at Purdue University. His current research interests include machine learning/data mining, computational engineering, and bioinformatics. He works in protein function prediction area and biological visualizations.



Daisuke Kihara is professor of Department of Biological Sciences and Department of Computer Science at Purdue University, West Lafayette, Indiana, USA. He has received B.S. degree from University of Tokyo, Japan in 1994, M.S. and Ph.D. degree from Kyoto University, Japan in 1996 and 1999, respectively. His research area is protein bioinformatics, which include protein structure, docking, and function prediction. He is named Showalter University Faculty Scholar from Purdue University in 2013.