SIGN: reliable protein interaction identification by integrating the Similarity In GO and the similarity in protein interaction Networks

Woochang Hwang† Taehyong Kim† Young-Rae Cho† Aidong Zhang† Murali Ramanathan†
†Department of Computer Science and Engineering, State University of New York at Buffalo, USA
††Department of Pharmaceutical Sciences, State University of New York at Buffalo, USA

Email: {whwang2, thkim7, ycho8, azhang}@cse.buffalo.edu, murali@acsu.buffalo.edu

Abstract—High-throughput techniques for protein-protein interaction detection in a genomic scale have provided us a genomic wide view of molecular interactions of many living organisms. A few approaches were proposed to scrutinize protein-protein interactions of living organisms. By the way, the binary nature of the current protein interaction data sets imposes challenges for effective analysis. Furthermore, their performance was suffered by the intrinsic defect, i.e., high noise level, of high-throughput data. This unpleasantly high false positive rate could lead many devoted researches to erroneous biological conclusions. We propose a novel reliability measurement for protein interactions integrating the similarity in Gene Ontology and the topological similarity in protein interaction networks. Our metric has been proven to be an effective reliability metric for identifying biologically more reliable interactions through the analyses performed from various view points, e.g., functional homogeneity, subcellular localizational homogeneity, and gene expression correlation, etc.

I. INTRODUCTION

Many important and useful information about underlying principles of living organisms can be extracted from protein-protein interactions (PPI) data. PPI data provide us opportunities to understand essential principles like essentiality, genetic interactions, functions, functional modules, protein complexes and cellular pathways. However, there are several challenges to overcome in order to extract valuable information from PPI. Among them, the binary nature of the current PPI data sets imposes challenges for effective analysis. The performance of many researches analyzing PPI was suffered by the intrinsic weakness of high-throughput data, e.g., a significant proportion, about 50%, of these high-throughput data has been found to be false positives [15], [5]. These false positive interactions might not occur in real living tissues. Therefore, these false positives could lead many devoted researches to erroneous biological conclusions. So, screening false positives from reliable interactions should be an indispensable work for more effective analysis on PPI data.

II. RELATED WORK

The GO is an important annotatory information resource of genes and gene products. Resnik [9] and Lin [6] measures the semantic similarity of a gene pair using two terms in GO which the gene pair belongs to. Jiang [4] developed a distance functions between GO terms reflecting dissimilarity between two GO terms. These proposed semantic similarity metrics shows significant relationships between GO based semantic similarity of gene pairs in GO and their structure, sequence similarity, and other functional properties; however, it only considers the cardinality of annotated genes for each GO term. The topological specificity in GO Directed Acyclic Graph (DAG) structure and the topological information in protein interaction network (PIN) are also essential information to investigate protein interactions, but not addressed in the previous mentioned approaches.

Many useful topological similarity metrics were proposed to measure reliability of a protein pair in PIN. Interaction generality (IG1) [10] and improved interaction generality (IG2) [11] was developed to quantize the reliability of each interaction. However, limited deliberation on the local topology connectivity limits the effectiveness of these metrics. The Interaction reliability by Alternative Path (IRAP) [1] and PathRatio [8] was proposed to measure the reliability of an interaction as the strongest path of the alternative paths and the sum of the most alternative paths between a node pair respectively. But, they only focused on the the PIN topology even though interaction reliability is more involved with other biological relationships, e.g., functional or subcellular localizational relationships.

In this paper, we propose a weighted PIN model in which the weight for each interaction will be the integrated information of the informational similarity in the Gene Ontology (GO) structure and the topological similarity of the end nodes. We term this weighted approach as SIGN, which integrates the similarity in GO and the similarity in the PIN in order to quantitatively measure the reliability of an interaction.
A. Gene Ontology DAG

Gene Ontology can be built as a directed acyclic graph (DAG) like the example illustrated in Figure 1. A directed graph $DG = (V, A)$ consists of a set $V$ of nodes and a set $A$ of arcs, $A \subseteq V \times V$. An arc $a = (i, j)$ connects nodes $i$ and $j$ and the direction of the arc is from $i$ to $j$, $a \in A$. The length of a node $i$ from $j$, Depth$(j, i)$, in a directed graph is the number of arcs on the shortest path from node $j$ to $i$. The Height$(i)$ of a node $i$ is the number of arcs on the shortest path from node $i$ to its furthest child in a directed graph. Two examples of Depth and Height are presented in Figure 1.

B. Similarity in GO

In this section, we will describe how to measure the similarity between two GO terms using GO DAG structure and GO annotation. The GO project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO collaborators developed three structured, controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [13].

SIGN measured similarity based on the integration of three GO categories. In the following two definitions, the specificity of the cardinality on annotated genes and the hierarchical position of GO term have been measured in the integrated GO DAG.

**Definition 1.** The cardinal specificity of a GO term $T$ is defined as the proportion of genes, a GO term annotates, out of all genes involved in one study:

$$\mu(T) = \frac{|T|}{N}$$

where $|T|$ is the number of genes annotated by a GO term $T$ and $N$ is the total number of genes in GO DAG.

$\mu(T)$ is the probability of finding a member of a term $T$ in the annotation database being analyzed.

**Definition 2.** The structural specificity of a GO term $T$ is defined as relative height of a term $T$ in GO DAG:

$$\psi(T) = \frac{\text{Height}(T) + 1}{\text{Height}(T) + \text{Depth}(\text{Root}, T) + 1}$$

where Depth$(\text{Root}, T)$ is the shortest distance of a GO term $T$ from the Root term and Height$(T)$ is the height of GO term $T$ in GO DAG.

$\psi(T)$ is the normalized height of term $T$ in GO DAG. Figure 1 describes cardinal and structural specificity of Term 2 and 3. It clearly shows that the cardinal specificity of these two terms is the same but the structural specificity of them is different because they are on the different hierarchical location in GO DAG. Term 3 should have higher specificity since it has more specific structural information than Term 2 even though they have the same cardinal specificity. The depth and height of a term in GO tree are measured in shortest path length since it can be solved in polynomial time among other measurements.

In information theory, self information is a measure of the information content associated with the outcome of a random variable. So, the amount of self information contained in a probabilistic event $c$ depends only on the probability $p(c)$ of that event. More specifically, the smaller this probability is, the larger is the self information associated with receiving information that the event indeed occurred. Based on this theory, we measure the information contained in a GO term $T$ based on the cardinal specificity and the structural specificity of term $T$ in GO DAG structure. The cardinal and structural specificity information contained in a GO term $T$ are measured as follows respectively:

$$\phi_{\mu}(T) = -\log(\mu(T))$$

$$\phi_{\psi}(T) = -\log(\psi(T))$$

Total information contained in a GO term $T$ will be the summation of cardinal specificity and structural specificity information since the total information of two independent event will be the summation of the information of each event. The self information combining the cardinal and structural information of a GO term $T$ in GO DAG is measured as follows:

$$\phi(T) = \phi_{\mu}(T) + \phi_{\psi}(T)$$

**Definition 3.** The most specific GO term, $\text{MST}(p)$, of a protein $p$ is a GO term $T$ that has the highest self information, defined in Definition 5, among the GO terms annotating protein $p$.

**Definition 4.** The most common specific GO term, $\text{MCSST}(p_1, p_2)$, of a protein pair, $p_1$ and $p_2$, is a GO term $T$ that has the highest self information, defined in Definition 5, among the GO terms annotating both proteins $p_1$ and $p_2$. 
For example, in Figure 1, the most specific GO terms for gene g1 and g2 are Term 5 and 6, respectively. The most common specific GO term of genes g1 and g2 is Term 2.

Similarity of a protein pair p1 and p2, which are the two incident nodes incident to an interaction e, in GO DAG is measured based on the normalized similarity metric proposed by Lin [6].

\[
\rho(p_1, p_2) = \frac{2 \times \phi(MCST(p_1, p_2))}{\phi(MST(p_1)) + \phi(MST(p_2))}
\]  

(C. Similarity in PIN)

In this section, we will describe how to measure the topological similarity between two proteins, which are incident to an interaction in PIN.

In Figure 2, we classify the local connectivity for each interaction into four different categories to summarize the topological property within the neighbors directly connected to the protein pairs, p1 and p2 which are incident to an interaction e:

- O(p1, p2): set of nodes that have an interaction to p1 and p2, (Figure 2a)
- I(p1, p2): set of interactions between two nodes i and j, node i interacts with p1 but not with p2 and node j interacts with p2 but not with p1, (Figure 2b)
- IO(p1, p2): set of interactions between two nodes i and j, node i interacts with p1 but not with p2 or vice versa and node j in the O(p1, p2) set. (Figure 2c)
- OO(p1, p2): set of interactions between two nodes i and j, nodes i and j are in the O(p1, p2) set. (Figure 2d)

Our topological similarity of a protein pair, p1 and p2, incident to an interaction e will use these four types of connectivity properties in PIN.

The node overlapping among neighbors of a protein pair p1 and p2 in a PIN will be measured as follows:

\[
\alpha(p_1, p_2) = \frac{|O(p_1, p_2)|}{\min(|N(p_1)|, |N(p_2)|)}
\]  

The inter-connectivity among mutually exclusive neighbors of a protein pair p1 and p2 in a PIN will be measured as follows:

\[
\beta(p_1, p_2) = \frac{|I(p_1, p_2)|}{|O(p_1)| \times |O(p_2)|}
\]  

where \(O(p_1)\) is the complement set, \(O(p_1) = N(p_1) - O(p_1, p_2)\), and \(O(p_2)\) is the complement set, \(O(p_2) = N(p_2) - O(p_1, p_2)\).

The inter-connectivity between mutually exclusive neighbors and overlapping nodes of a protein pair p1 and p2 in a PIN will be measured as follows:

\[
\gamma(p_1, p_2) = \frac{|IO(p_1, p_2)|}{|O(p_1)| \times |O(p_2)|}
\]  

The inter-connectivity among overlapping neighbors of a protein pair p1 and p2 in a PIN will be measured as follows:

\[
\delta(p_1, p_2) = \frac{2 \times |OO(p_1, p_2)|}{|O(p_1)| \times |O(p_2)| - 1}
\]  

Once again, information theoretic approach is utilized to measure the topological information shared in the neighbor region by two incident nodes of an interaction based on the four topological properties defined above.

Topological common information for each category shared by the neighbors of a protein pair p1 and p2 incident to an interaction e is measured as follows:

\[
\phi_{\alpha}(p_1, p_2) = -\log(1 - \alpha(p_1, p_2)) \quad (11)
\]

\[
\phi_{\beta}(p_1, p_2) = -\log(1 - \beta(p_1, p_2)) \quad (12)
\]

\[
\phi_{\gamma}(p_1, p_2) = -\log(1 - \gamma(p_1, p_2)) \quad (13)
\]

\[
\phi_{\delta}(p_1, p_2) = -\log(1 - \delta(p_1, p_2)) \quad (14)
\]

\(\phi_{\alpha}(p_1, p_2)\) is the information about the node overlapping among neighbors of a protein pair p1 and p2, \(\phi_{\beta}(p_1, p_2)\) is the information about the inter-connectivity among mutually exclusive neighbors of a protein pair p1 and p2, \(\phi_{\gamma}(p_1, p_2)\) is the information about the inter-connectivity between a mutually exclusive neighbor and an overlapping node of a protein pair p1 and p2, \(\phi_{\delta}(p_1, p_2)\) is the information about the inter-connectivity among overlapping neighbors of a protein pair p1 and p2.

Topological similarity of an interaction e, which p1 and p2 are the incident protein pair, is measured as the summation of these four topological common information.

\[
\tau(p_1, p_2) = \phi_{\alpha}(p_1, p_2) + \phi_{\beta}(p_1, p_2) + \phi_{\gamma}(p_1, p_2) + \phi_{\delta}(p_1, p_2)
\]  

(D. Interaction Reliability)

So far, we have defined the similarity in GO DAG structure, \(\rho(p_1, p_2)\), and the topological similarity in a PIN, \(\tau(p_1, p_2)\), of a protein pair p1 and p2 which are the two incident nodes of an interaction e. The reliability of an interaction e is defined as the integration of these two similarities.

\[
\text{Reliability}(e) = \rho(p_1, p_2) + \tau(p_1, p_2)
\]
IV. RESULTS

A. Protein-Protein Interaction Data

One combined data set from three different data sets, which has 4847 proteins and 16125 interactions, is used for this research. Ito data set includes 3278 proteins and 4393 interactions by Ito et al. [3]. Uetz data set includes 1003 proteins and 904 interactions in [14]. DIP database includes 2521 proteins and 5716 interactions obtaining the yeast (S. cerevisiae) PPI data [2].

B. Functional and Localizational Homogeneity Analysis

To evaluate that our SIGN reliability metric is effective on identifying reliable interactions, we first analyzed whether a protein pair, which are the two incident nodes to an interaction, is in the same functional or localizational category. MIPS functional and localizational categories were used as the analyzing ground truth [7]. Performance of the other four existing reliability metrics, IG2, MCC, IRAP, and PathRatio, were compared.

We ordered interactions according to the reliability value for each measurement from the highest and to the lowest, and measured the functional homogeneity by counting the number of interactions such that the incident protein pair shares the same functional role. Figure 3 shows the functional homogeneity performance for each reliability measurement. It is clear that SIGN identified more interactions that have the same functional role than any other metrics did in the top 50 percentile. PathRatio shows a comparable performance in the top 6 percentile, but its performance dramatically worsen after the 6 percentile. The functional homogeneity of SIGN is the worst among the other comparing metrics after the 50 percentile. In other words, SIGN identified the most number of reliable interactions sharing the same functional role in the higher rank than the other metrics did in the same interval.

Subcellular localizational homogeneity was also analyzed. Interactions were ordered according to their reliability values in the same manner as we did in the previous analysis, and measured the localizational homogeneity by counting the number of interactions such that the incident protein pair shares the same subcellular location. As you can see in Figure 4, SIGN shows the best result above other 4 metrics in the top 50 percentile, and shows lower localizational homogeneity after 62 percentile. Therefore, SIGN identified more reliable interactions in the higher rank and less reliable interactions in the lower rank for the functional and localizational homogeneity viewpoint.

C. Gene Expression Correlation Analysis

To analyze effectiveness of SIGN reliability metric, we analyzed the relationship between the reliability of protein pairs and their gene expression correlation in a microarray data set. Spellman gene expression data set was used to measure gene expression correlation among genes [12]. Figure 5 plots the gene expression correlation between a protein pair which are the two proteins incident to each interaction. More reliable interaction should have higher gene expression correlation. SIGN shows the highest correlation above other 4 metrics in the top 25 percentile. PathRatio shows comparable performance in the top 6 percentile than SIGN does, but its performance is getting worse after the top 6 percentile than SIGN. So, SIGN metric have higher gene expression correlation for the interactions in the higher rank and lower
correlation for the interactions in the lower rank.

D. Interaction Reproduction Analysis

Three different interaction data sets were combined for this research. Protein interactions that are detected by multiple independent experiments are often regarded as highly reliable. We estimated an interaction as a true positive, i.e., a reproducible interaction, if it appears in more than one experiments, and used it as the gold standard to analyze the effectiveness of reliability metrics. Figure 6 shows that the accumulated number of experimentally reproducible interactions for each metric. SIGN shows the best performance in the top 37 percentile identifying more number of reproducible interactions, i.e., interactions detected by more than one experiments, than any other measurements. Therefore, SIGN identified more reliable interactions in the higher rank, top 37 percentile, than the other 4 metrics.

V. DISCUSSION

Analyzing protein interactions should be one of the most essential process in understanding the underlying biological principles of living organisms. High throughput methods for protein interactions detection provided as a genomic level view of molecular interactions of living organisms. But, high noise level of high throughput data has confined the effectiveness of many devoted researches.

A novel reliability metric, termed ‘SIGN’, is introduced in this paper to settle the problem mentioned in the above. The performance of SIGN was compared with the other various types of existing approaches. SIGN has been proven to be effective in identifying more reliable interactions than the other 4 competing metrics. SIGN outperformed the other 4 methods on each performance analysis.

Another essential aspect that we should not overlook is the behavior consistency on the analyses performed in the previous sections. The performance of SIGN showed consistent, i.e., monotonic, behavior throughout all intervals on each analysis carried out, but the other 3 metrics, IG2, MCC, TRAP, failed to show monotonic behavior. Their fluctuating behavior is resulted from the deficiency of their discriminating ability among interactions, i.e., the coarseness of their reliability values for interactions. For example, IRAP has the tie reliability score for the top 5130 interactions, and the next 2408 interactions also has the same reliability. The same phenomenon was observed in IG2 and MCC approaches though there is a difference in the extent of manifestation. This discriminating deficiency made these metrics give the same reliability values to the interactions which have to be differentiated in their reliability. In other words, they failed to discriminate between reliable and unreliable interactions. Therefore, this drawback on the other 3 metrics should prevent them to be an effective reliability metric. PathRatio, which is the most comparable method to SIGN, did not have a weakness in its discriminating ability. By the way, its comparable performance was limited in the restricted top small portion, e.g., top 6% in the functional and localizational homogeneity analysis. Its performance was suddenly deteriorated after the top restricted portion.

Consequently, SIGN discriminates reliable and unreliable interactions better than the other comparing methods do. Moreover, it shows desirable monotonic behavior throughout all intervals, while the other metrics do not.

REFERENCES