BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–7

Functional Coherence in Domain Interaction Networks

Jayesh Pandey[a], Mehmet Koyutürk[b], Shankar Subramaniam[c], and Ananth Grama[a]

[a]Dept. of Computer Science, Purdue Univ., [b]Dept. of Electrical Engineering & Computer Science, Case Western Reserve Univ., [c]Dept. of Biomedical Engineering, Univ. of California, San Diego.

ABSTRACT

Motivation: Extracting functional information from protein-protein interactions (PPI) poses significant challenges arising from the noisy, incomplete, generic, and static nature of data obtained from high-throughput screening. Typical proteins are composed of multiple domains, often regarded as their primary functional and structural units. Motivated by these considerations, domain-domain interactions (DDI) for network-based analyses have received significant recent attention. This paper performs a formal comparative investigation of the relationship between functional coherence and topological proximity in PPI and DDI networks. Our investigation provides the necessary basis for continued and focused investigation of DDIs as abstractions for functional characterization and modularization of networks.

Results: We investigate the problem of assessing the functional coherence of two biomolecules (or segments thereof) in a formal framework. We establish essential attributes of admissible measures of functional coherence, and demonstrate that existing, well-accepted measures are ill-suited to comparative analyses involving different entities (*i.e.*, domains vs. proteins). We propose a statistically motivated functional similarity measure that takes into account functional specificity as well as the distribution of functional attributes across entity groups to assess functional similarity in a statistically meaningful and biologically interpretable manner. Results on diverse data, including high-throughput and computationally predicted PPIs, as well as structural and computationally inferred DDIs for different organisms show that: (i) the relationship between functional similarity and network proximity is captured in a much more (biologically) intuitive manner by our measure, compared to existing measures, and (ii) network proximity and functional similarity are significantly more correlated in DDI networks than in PPI networks, and that structurally determined DDIs provide better functional relevance as compared to computationally inferred DDIs.

Contact: Jayesh Pandey, jpandey@cs.purdue.edu

1 INTRODUCTION

Availability of high-throughput protein-protein interaction (PPI) data makes it possible to study the function of biological systems from a network perspective. Recent advances in this area have focused on the development of computational infrastructure for network-based functional annotation (Sharan *et al.*, 2007), identification of functionally coherent modules (Spirin and Mirny, 2003), and evolutionary network analysis (Koyutürk *et al.*, 2006). However, the use of PPI data for computational assessment of network function poses several challenges: (i) PPI data generated by high-throughput screening is generally noisy and incomplete (Titz *et al.*, 2004), (ii) PPI data provides only a generic and static picture of the cellular network, *i.e.*, it does not capture the spatio-temporal dynamics of biological systems (Han *et al.*, 2004), and (iii) proteins themselves are typically composed of multiple functional domains. For this reason, significant efforts are devoted to increasing the quality and reliability of PPI data, as well as using other data sources and abstractions to study interaction data (Lee *et al.*, 2004).

An important limitation of PPI data that relates to the dynamics of cellular systems is that it does not explicitly capture the domain specificity of interactions. Domains in proteins are often regarded as primary functional and structural units (Bateman *et al.*, 2004). Therefore, the functional relevance of an interaction may be considered at the domain level as well. However, the specificity of interactions at this level cannot be captured by high-throughput screening. Consequently, domain-domain interactions (DDI) are often identified using either dedicated structural analysis (Gong *et al.*, 2005) or computational inference from PPI data (Deng *et al.*, 2002; Riley *et al.*, 2005). As DDI data and databases become commonplace (Ng *et al.*, 2003; Raghavachari *et al.*, 2007), DDI networks provide an attractive abstraction for functional network analysis (Schlicker *et al.*, 2007; Wuchty, 2006).

In this paper, we investigate how functional modularity manifests itself in a network of molecular interactions, considering different molecular entities – proteins and domains. This question is studied extensively on PPI and gene co-expression networks (Sevilla *et al.*, 2005), however, knowledge on interactions involving different molecular entities is relatively scarce. In order to provide a basis and motivation for computational analysis of DDI networks, we investigate how network proximity in a DDI network relates to the functional coherence of domains. For this purpose, we consider PPI networks as a reference, and compare PPI and DDI networks comprehensively in terms of the relationship between network proximity and functional similarity.

While comparing networks composed of different molecular entities, it is particularly difficult to quantify the functional similarity between two entities in an unbiased manner. This is because functional similarity may have different meanings for different molecular entities. Furthermore, from a practical standpoint, the functional information available for different types of molecular entities may have different characteristics. This is indeed the case for proteins and domains. Most of the available functional annotations for domains are derived from annotations for proteins (Schug *et al.*, 2002). Consequently, they are more general, scarce, and incomplete compared to protein annotations. Motivated by this observation, we develop a formal framework for evaluating metrics for assessing functional similarity between two molecular entities. We establish essential attributes of admissible measures of functional coherence, and demonstrate that existing, well-accepted measures are ill-suited

to comparative analyses involving different entities. We propose an information theoretic functional similarity measure that takes into account functional specificity as well as distribution of functional attributes across entities. This results in a more statistically meaningful and biologically interpretable functional similarity measure that relies on only positive evidence to quantify the functional coherence of molecular entities – thus eliminating any artifacts caused by incompleteness of annotations. On a comprehensive collection of PPI and DDI data, we show that our measure indeed captures the relation between network proximity and functional coherence in a more biologically interpretable manner.

Using our proposed functional similarity measure, we compare PPI and DDI networks for diverse species comprehensively. We consider PPIs from large public databases that integrate different sources of data, as well DDIs that are derived from different sources, ranging from structural analysis to computational inference. Our results show that functional coherence is more closely related to network proximity in DDI networks as compared to PPI networks, clearly motivating the use of DDI data in the analysis of networks for functional inference. We also show that, for different sub-ontologies of Gene Ontology (Ashburner *et al.*, 2000), functional coherence manifests itself differently in the networks.

2 METHODS

Understanding the relationship between network topology and functional modularity requires measures for assessing the functional similarity (or coherence) of a group of entities with respect to each other. For example, in testing the hypothesis that functional modularity is related to high connectivity in PPI networks, it is common to investigate the functional purity of groups of proteins that induce dense subgraphs in the network (Grossmann *et al.*, 2006). In this work, we focus on the relationship between the topological proximity of two entities in a network and their functional similarity. The eventual goal is to determine whether functional relationship manifests itself better in PPI or DDI networks.

There exist several approaches to assessing functional similarity of bio-molecules (*e.g.*, genes, proteins, domains) (Lord *et al.*, 2003; Schlicker *et al.*, 2007). Since functional categories are not isolated, but rather related to each other through a taxonomy (*e.g.*, Gene Ontology), it is necessary to consider the underlying taxonomy while comparing molecules in terms of their functional annotation (Resnik, 1995). Various approaches take into account different factors, including taxonomical distance, specificity/generality (rank in hierarchy) of common ancestor, and associated number of molecules for the functional terms being compared. Since most molecules are associated with multiple functional terms, assessment of functional similarity between two molecules poses an additional challenge, namely one of evaluating the similarity between two *sets* of terms, as opposed to a *pair* of terms. Common, and relatively straightforward approaches to this problem include taking the maximum (Schlicker *et al.*, 2007) or average (Lord *et al.*, 2003) of similarities among all pairs of terms in the two sets. We show that neither of these alternatives provide robust metrics for extending term similarity to set similarity. We develop an information theoretic measure for set-similarity that directly computes similarity of sets as a whole, as opposed to computing an aggregate of pairwise term-similarities. Our measure takes into account the information content of the most specific of the common ancestors of all terms, and quantifies positive reinforcement of similar terms, avoiding negative contributions arising from incomplete data. In order to motivate this approach, we provide a formal framework for the problem, and identify the desirable properties of a metric for evaluating the functional similarity between two molecules in this framework.

2.1 Concepts and Ontologies

Let $C = \{c_i | 1 \leq i \leq N\}$ be a finite partially ordered set of concepts. In terms of Gene Ontology (GO), these concepts represent the GO terms (*i.e.*, molecular function, biological process, and cellular component). Without loss of generality, we refer to concepts as terms throughout this paper. Terms are related to each other through *is a* and *part of* relationships, such that $c_i \rightarrow c_j$ denotes c_i *is a/part of* c_j. Note that, if $c_i \rightarrow c_j$, then the molecules associated with c_i are also associated with c_j, known as the *true path rule*. Based on these relationships, we define a binary relation over C, denoted by $\preceq$. We say c_j is an ancestor of c_i, denoted by $c_i \preceq c_j$ if and only if either $c_i \rightarrow c_j$, or for some $\ell \geq 1$, there exist $c_{k_l} \in C$ for $1 \leq l \leq \ell$ such that $c_i \rightarrow c_{k_1}$, $c_{k_l} \rightarrow c_{k_{l+1}}$ for $1 \leq l < \ell$, and $c_{k_\ell} \rightarrow c_j$ (c_j is an ancestor of c_i in GO hierarchy). Two terms c_i, c_j are comparable, denoted by $c_i \sim c_j$, if either $c_j \preceq c_i$ or $c_i \preceq c_j$. If c_i and c_j are comparable, then the shortest path between c_i and c_j is given by $L(c_i, c_j) = L(c_j, c_i) = \ell + 1$ for minimum such ℓ.

We denote the set of ancestors of a term c_i by $A_i = \{c_k \in C | c_i \preceq c_k\}$. Note that, not all ancestors of a term are comparable, since the GO hierarchy is a directed acyclic graph, as opposed to a tree. We represent the root term of GO with a terminal concept r, such that $c_i \preceq r \; \forall c_i \in C$.

2.2 Semantic Similarity of Terms

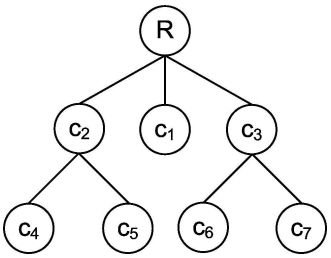
Semantic similarity measures are intended to quantify the similarity between two terms based on the underlying taxonomical relationships. For a semantic similarity measure $\delta : C^2 \rightarrow \Re$, we identify the following as properties that must be satisfied for the measure to be meaningful:

(1) $\delta(c_i, c_j) = \delta(c_j, c_i)$, for all c_i, c_j,
(2) $\delta(c_i, c_i) \leq \delta(c_j, c_j)$ for $c_j \preceq c_i$,
(3) $\delta(c_i, c_j) \leq \delta(c_j, c_j)$ for all c_i, c_j,
(4) if $\exists c_m \in A_i \cap A_j$ such that $c_m \preceq c_n$, $\forall c_n \in A_k \cap A_l$, then $\delta(c_i, c_j) \geq \delta(c_k, c_l)$.

The first property states that a semantic similarity measure must be symmetric. The second property states that more specific terms should have at least as much self-similarity as more general terms. The third property states that a term should not be less similar to itself than to any other term. The fourth property states that terms with more specific common ancestors should be more similar to each other, compared to those with less specific common ancestors. Note that if δ satisfies properties (3) and (4), then $\delta(r, r) \leq \delta(c_i, c_j)$ and $\delta(c_i, r) = \delta(r, r) \; \forall i, j$.

We now discuss candidate measures for assessing the semantic similarity of two terms.

Distance. Let the depth of a term c_i be $d(c_i) = L(c_i, r)$ and the depth of the entire hierarchy be $D = \max_{1 \leq i \leq N} L(c_i, r)$. For terms c_i and c_j that are not comparable, let $L(c_i, c_j) = \min_{c_k \in A_i \cap A_j} L(c_i, c_k) + L(c_k, c_j)$. Then, the distance between two terms in the hierarchy is defined as $\delta_E(c_i, c_j) = 2D - L(c_i, c_j)$.



$S_1 = \{c_4, c_6, c_7\}$, $S_2 = \{c_4\}$, $S_3 = \{c_4, c_6\}$, $S_4 = \{c_6, c_7\}$, $S_5 = \{c_4, c_3\}$

Fig. 1. Sample ontology and associated annotations. Each node of the hierarchy represents a GO term, each set of terms represents a protein (or domain).

This measure satisfies all properties except (4), *i.e.*, it does not take into account the specificity of the common ancestor.

Information content. This measure takes into account the distribution of terms among molecules. Let G_c be the set of molecules that are associated with term c. Then, the information content of a term is defined as $I(c) = -\log_2(|G_c|/|G_r|)$ (where G_r is the set of all molecules)(Resnik, 1995). Clearly, $I(r) = 0$, and as a consequence of the true path rule, $I(c_j) \geq I(c_i)$ for $c_j \preceq c_i$. Then, the semantic similarity between two terms is defined as

$$\delta_I(c_i, c_j) = \max_{c \in A_i \cap A_j} I(c) = I(\lambda(c_i, c_j)). \quad (1)$$

Note that, $\lambda(c_i, c_j) = \text{argmax}_{c \in A_i \cap A_j} I(c)$is said to be the *minimum common ancestor* of c_i and c_j.

This measure satisfies all four properties described above, but does not take into account the specificity of terms with identical common ancestors, as illustrated in Figure 1. In the figure, we have $\delta_I(c_2, c_3) = \delta_I(c_5, c_6)$. Although c_5 and c_6 are more specific concepts farther apart from each other, their similarity is equal to that of their parents, c_2 and c_3. This problem can be alleviated by normalizing the similarity between two terms by the self-similarities of the terms being compared, *e.g.*, $\delta_L(c_i, c_j) = \frac{2\delta_I(c_i, c_j)}{I(c_i)+I(c_j)}$ (Lin, 1998), and $\delta_{JC}(c_i, c_j) = \frac{1}{1-2\delta_I(c_i, c_j)+I(c_i)+I(c_j)}$ (Jiang and Conrath, 1997). It is evident in Figure 1 that these measures satisfy $\delta_L(c_2, c_3) \geq \delta_L(c_5, c_6)$ and $\delta_{JC}(c_2, c_3) \geq \delta_{JC}(c_5, c_6)$. We now generalize these term-similarity measures to set-similarity.

2.3 Functional Similarity of Molecules

Since most molecules are associated with multiple molecular functions and sometimes involved in multiple processes, the annotation of a molecule consists of a set of terms. While assessing the similarity of term sets, we assume that each set consists of terms that are not comparable, *i.e.*, each branch of the hierarchy is represented by at most one term in each set. In GO, this involves considering only the most specific annotations associated with a gene, which enables non-redundant representation of functional annotation. In this representation, the association between the gene and the ancestors of the most specific term is implied by the true path rule.

A set of terms $S \subseteq C$ is said to be non-redundant if $\forall c_i, c_j \in S, c_i \not\preceq c_j$. Note that, to satisfy non-redundancy requirement for any set, we define the trim of a term set S as $\Upsilon(S) = \{c_i \in S : \exists \text{ no } c_j \in S \text{ s.t. } c_j \preceq c_i\}$. By definition, $\Upsilon(S)$ is non-redundant for any S. Now we can define the semantic similarity measure for sets assuming that the sets are non-redundant, since any set of terms has a unique trim[1]. For two non-redundant sets $S_i, S_j \subseteq C$, we need a measure $\rho(S_i, S_j)$ to assess their semantic similarity. We identify the following as properties of any such measure:

(i) $\rho(S_i, S_j) = \rho(S_j, S_i)$ for all S_i, S_j,

(ii) $\rho(S_i, S_j) \leq \rho(S_i \cup c_k, S_j \cup c_k)$ where $\forall c_l \in S_i, \cup S_j, c_l \not\sim c_k$ for all S_i, S_j,

(iii) $\rho(S_i, S_j) \leq \rho(S_i, S_j \cup S_k)$ for all S_i, S_j, S_k,

(iv) $\rho(S_i, S_j) \leq \rho(S_j, S_j)$ for all S_i, S_j.

Property (ii) states that adding a common annotation for two molecules should not decrease the similarity between these two molecules. Property (iii) states that if new annotations are added for a new molecule, the similarity of this molecule to any other molecule should not decrease. This seemingly unintuitive property is motivated by the fact that existing annotations are quite incomplete. For this reason, we require semantic similarity measures to rely only on positive evidence, avoiding negative conclusions based on lack of annotations. Property (iv) states that a set of annotations should be at least as similar to itself as it is to any other set.

Common approaches to computing similarity between sets include, taking the *average* or *maximum* of the similarities between all pairs in the two sets. We first discuss the limitations of such straightforward approaches, and propose a generalization of Resnik's information-content based term similarity measure to sets of terms. This also relates to the statistical significance of the similarity between two term sets.

Average: $\rho_A(S_i, S_j) = \frac{1}{|S_i||S_j|} \sum_{c_k \in S_i} \sum_{c_l \in S_j} \delta(c_k, c_l)$ (Lord *et al.*, 2003). The idea behind this measure is that the semantic similarity between any two pairs of annotations contributes to the functional similarity between two molecules, so that molecules with more similar functions are assigned a higher similarity score. By letting $S_i = \{a\}$, $S_j = \{b\}$, and choosing c_k such that $3\delta(a, b)/4 > \delta(c_k, c_k) + \delta(a, c_k) + \delta(b, c_k)$, it can be shown that this measure does not satisfy property (ii). Furthermore, it satisfies property (iii) only if $\rho_A(S_i, S_j) = \rho_A(S_i, S_k)$. Similarly, letting $S_i = \{a, b\}$, $S_j = \{c\}$ and $2(\delta(a, c) + \delta(b, c) - \delta(a, b)) > \delta(a, a) + \delta(b, b)$, it can be seen that this measure violates property (iv) as well.

Maximum: $\rho_M(S_i, S_j) = \max_{c_k \in S_i, c_l \in S_j} \delta(c_k, c_l)$ (Sevilla *et al.*, 2005). This measure is based on the notion that if two molecules perform similar functions in at least one context, then they can be considered functionally similar. While this measure satisfies all properties, it satisfies (ii) weakly, *i.e.*, $\rho_M(S_i, S_j) = \rho_M(S_i \cup c_k, S_j \cup c_k)$ unless there exists no $c_m \in S_i$ and $c_n \in S_j$ such that $\delta(c_m, c_n) \geq \delta(c_k, c_k)$.

Average of maximums: Average functional similarity between two proteins can be defined in terms of a compromise between these two measures (Schlicker *et al.*, 2007), namely $\rho_H(S_i, S_j) = \max\left\{\frac{1}{|S_i|}\sum_{c_k \in S_i} \max_{c_l \in S_j} \delta(c_k, c_l), \frac{1}{|S_j|}\sum_{c_l \in S_j} \max_{c_k \in S_i} \delta(c_k, c_l)\right\}$.This

[1] To see that $\Upsilon(S)$ is unique for S, recall that the underlying hierarchy of terms is represented by a directed acyclic graph. Consequently, its transitive closure is also an acyclic graph, in which an edge represents ancestral relationship between two terms. Observe that the trim of a term set is equivalent to the set of nodes with no incoming arcs in the subgraph induced by the term set on this transitive closure, therefore it is uniquely defined.

modification provides a more biologically sound formulation of average functional similarity between two molecules, since a function of one molecule may be considered to be shared by another molecule as long as the other molecule is associated with a sufficiently similar function. However, this measure also fails to satisfy properties (ii), (iii), and (iv).

Information content: Observing that the notion of minimum common ancestor can be extended to sets of terms, we propose a set-similarity measure that is defined on entire sets, as opposed to a composite of pairwise similarities. Let $\Lambda(S_i, S_j) = \bigsqcup_{c_k \in S_i, c_l \in S_j} \lambda(c_k, c_l)$ be the minimum common ancestor set of term sets S_i and S_j. Here, $\sqcup$ denotes a generalized union operator that preserves non-redundancy, *i.e.*, $A \sqcup B = \Upsilon(A \cup B)$. We define the similarity between two term sets as the information content of the set of minimum common ancestors, *i.e.*,

$$\rho_I(S_i, S_j) = I(\Lambda(S_i, S_j)) = -\log_2\left(\frac{|G_{\Lambda(S_i,S_j)}|}{|G_r|}\right), \quad (2)$$

where $G_{\Lambda(S_i,S_j)} = \bigcap_{c_k \in \Lambda(S_i,S_j)} G_{c_k}$ is the set of molecules that are associated with all terms in the minimum common ancestor set of S_i and S_j. Note that the above definition also generalizes the concept of information content from a single term to a set of terms.

Example. Consider the ontology in Figure 1. The root term in this ontology is R. The annotation sets for five molecules are also shown in the figure. Consider the similarity between the two molecules with annotation sets S_1 and S_2. Since $\lambda(c_4, c_4) = c_4$, $\lambda(c_6, c_4) = \lambda(c_7, c_4) = R$, and $c_4 \preceq R$, we have $\Lambda(S_1, S_2) = \{c_4\}$. Consequently, $\rho_I(S_1, S_2) = -\log_2(|G_{c_4}|/|G_R|) = -\log_2(|\{S_1, S_2, S_3, S_5\}|/|\{S_1, S_2, S_3, S_4, S_5\}) = \log_2(5/4)$. On the other hand, since $\Lambda(S_1, S_3) = \{c_4, c_6\}$, we have $\rho_I(S_1, S_3) = \log_2(5/2) > \rho_I(S_1, S_2)$. Observe that $\rho_M(S_1, S_2) = \rho_M(S_1, S_3)$, illustrating that ρ_I is stronger than ρ_M in terms of property (ii).

THEOREM 1. *ρ_I satisfies all properties required for a measure of semantic similarity between two sets of terms.*

PROOF. (i) Trivially, $\rho_I(S_i, S_j) = \rho_I(S_j, S_i)$ for all S_i, S_j.

(ii) Since $c_k \nsim c_n$ for all $c_n \in S_i \cup S_j$, we have $\Lambda(S_i \cup c_k, S_j \cup c_k) = \Lambda(S_i, S_j) \sqcup \Lambda(S_i \sqcup S_j, \{c_k\}) \sqcup \{c_k\} \supseteq \Lambda(S_i, S_j) \cup \{c_k\}$, leading to $G_{\Lambda(S_i \cup c_k, S_j \cup c_k)} \subseteq G_{\Lambda(S_i,S_j)}$ and $|G_{\Lambda(S_i \cup c_k, S_j \cup c_k)}| \leq |G_{\Lambda(S_i,S_j)}|$. Consequently, $\rho_I(S_i \cup c_k, S_j \cup c_k) \geq \rho_I(S_i, S_j)$.

(iii) $\Lambda(S_i, S_j \cup S_k) = \Lambda(S_i, S_j) \sqcup \Lambda(S_i, S_k) \supseteq \Lambda(S_i, S_j)$. Therefore, $G_{\Lambda(S_i, S_j \cup S_k)} \subseteq G_{\Lambda(S_i,S_j)}$, leading to $\rho_I(S_i, S_j \cup S_k) \geq \rho_I(S_i, S_j)$.

(iv) Clearly, $c_k \preceq \lambda(c_k, c_l)$ for any c_k, c_l. Now consider any $c_m \in \Lambda(S_i, S_j)$. Since $c_m = \lambda(c_k, c_l)$ for some $c_k \in S_i$ and $c_l \in S_j$, there always exists $c_n \in \Lambda(S_i, S_i)$ such that $c_n \preceq c_k \preceq c_m$. Consequently, we must have $G_{\Lambda(S_i,S_i)} \subseteq G_{\Lambda(S_i,S_j)}$, leading to $\rho_I(S_i, S_j) \leq \rho_I(S_i, S_i)$.

Note that, ρ_I also has the problem associated with Resnik's measure (Section 2.2) and that this problem can be alleviated through normalization by self-similarities, *e.g.*, $\rho_L = \frac{2\rho_I(S_i,S_j)}{\rho_I(S_i,S_i)+\rho_I(S_j,S_j)}$ or $\rho_{JC} = 1/(\rho_I(S_i, S_i) + \rho_I(S_j, S_j) - 2\rho_I(S_i, S_j) + 1)$.

Table 1. Protein-protein interaction dataset.

	C.eleg	D.mela	H.sapi	S.cere	S.pomb
Proteins	2308	5151	6718	4673	745
Interactions	3577	14529	19316	35833	1277

3 MATERIALS

In order to evaluate the suitability of PPIs and DDIs to different functional analyses, we obtain protein and domain interaction data for five well-studied eukaryotic species from public databases. These datasets contain physical protein-protein interactions, as well as structural and computationally inferred domain-domain interactions.

3.1 Protein-Protein Interactions

We obtain protein interaction data for five species, *C. elegans*, *D. melanogaster*, *H. sapiens*, *S. cerevisiae*, and *S. pombe*, from the BioGrid database (Breitkreutz *et al.*, 2007). The networks are chosen to be largest among available networks in the database, with the expectation that larger networks are relatively more comprehensive. We filter the dataset to obtain a set of physical interactions between proteins, *i.e.*, genetic interactions are removed based on experiment type (*e.g.*, knockout experiments). The interaction data is binary, *i.e.*, no confidence score is associated with the interactions. The numbers of proteins and interactions in each PPI network are shown in Table 1. Integr8 (Kersey *et al.*, 2005) is used to map the proteins in the interaction dataset to their Uniprot names. The data is filtered to keep only those proteins for which pfam domain decomposition is known using Integr8.

3.2 Domain interactions

We obtain domain interaction data from the DOMINE database (Raghavachari *et al.*, 2007). This dataset is composed of known, as well as predicted domain interactions. Interactions inferred from PDB entries of protein complexes are collected from iPfam and 3did. Predicted interactions are obtained through computational methods, which infer domain interactions from protein interaction networks or co-evolution of conserved sites (for details, see Raghavachari *et al.* (2007)). Based on the source and quality of the data, we partition this dataset into five classes:

- **Struct**: Only known domain interactions (structure based)
- **HC+NA** : High Confidence (HC) and Structure based (NA) interactions
- **HC+MC** : High Confidence (HC) and Medium Confidence (MC) interactions
- **Comp-2**: Interactions predicted by at least two computational approaches
- **Comp-1**: Interactions predicted by at least one computational approach

The numbers of domains and interactions in each class are shown in Table 2. Note that domain-domain interactions here are binary, *i.e.*, there is no confidence score associated with these interactions.

3.3 Gene Ontology & Annotations

Gene Ontology Annotation (GOA) (Camon *et al.*, 2006) is used to obtain annotation information for Uniprot proteins. The mapping of Pfam-A domains to their Gene Ontology functions is

Table 2. Domain-domain interaction dataset.

	Struct	HC + NA	HC + MC	Comp-2	Comp-1
Domains	2948	2978	1699	930	2933
Interactions	4349	5875	3957	1745	17781

obtained from pfam2go (`http://www.geneontology.org/external2go/pfam2go`). We use only the Biological Process and Molecular Function sub-ontologies of GO for evaluation, since the coverage for the Cellular Component sub-ontology is relatively low.

4 RESULTS

We first compare different semantic similarity measures on comprehensive PPI and DDI data. Then, using our proposed semantic similarity measure, we investigate the differences between PPI and DDI networks in terms of the relationship between network proximity and functional similarity.

4.1 Comparison of Semantic Similarity Measures

For each network, we compute the distance between all pairs of molecules (proteins or domains) in the network. Then, we group molecule pairs according to their distance and compute the average semantic similarity for each group. Since the distribution and range of semantic similarity scores varies across different measures, we normalize semantic similarity scores to obtain a mean similarity score of zero and standard deviation of one in each network. In other words, for each similarity measure ρ_x, the similarity score between two molecules $P_i, P_j \in \mathcal{P}$ is computed as $\hat{\rho}_x(P_i, P_j) = \frac{\rho_x(P_i, P_j) - \mu_x(\mathcal{P})}{\sigma_x(\mathcal{P})}$, where $\mathcal{P}$ denotes the set of molecules in the network. Note also that this normalization is useful in comparing PPI and DDI networks as well, since the distribution of available annotations across proteins and domains can be significantly different. In general, since domain annotations are generally derived from protein annotations, domain annotations are relatively scarce and more general (higher in the GO hierarchy) compared to protein annotations.

In Figure 2(a), the behavior of different semantic similarity measures with respect to network distance in the *C. elegans* PPI network is shown. We consider five measures, namely ρ_A/δ_I (average of Resnik's term similarity measure), ρ_H/δ_I (average of maximums for Resnik's term similarity measure), ρ_A/δ_{JC} (average of self-normalized Resnik's term similarity measure), ρ_I (proposed information content based molecule similarity measure), and ρ_{JC} (proposed information content based molecule similarity measure with self-normalization). As evident in the figure, all semantic similarity measures demonstrate a negative relation between network distance and functional similarity. However, if average term similarity score is used to compare molecules, an anomaly is observed in that average semantic similarity tends to increase for pairs of proteins at larger distances (≥ 4). This behavior demonstrates the inadequacy of average-based measures in handling randomness. Observe that in a network, the number of protein pairs with given distance grows with increasing distance and goes down after a point, which is the behavior of the curve for ρ_A/δ_I in Figure 2 in reverse direction. Consequently, the growth in average similarity

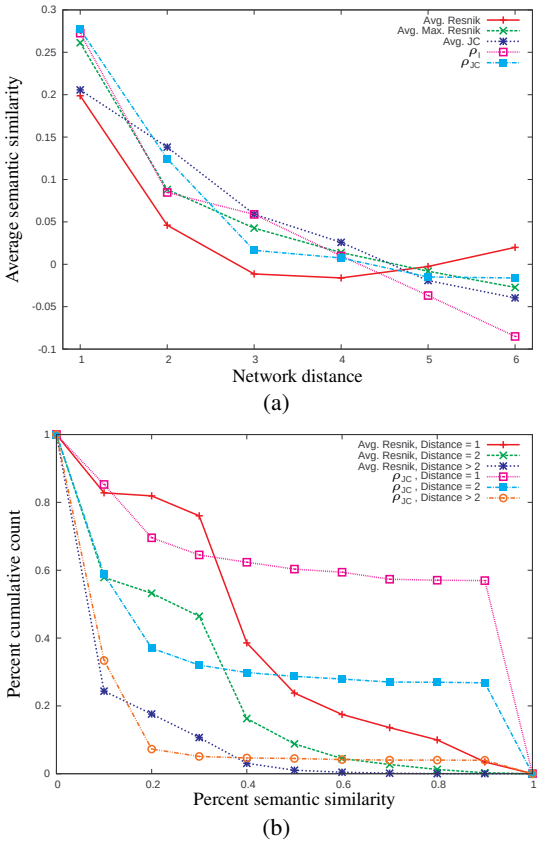


Fig. 2. Comparison of different semantic similarity measures in terms of their behavior with respect to network distance: (a) network distance vs. average semantic similarity for pairs of proteins in *C. elegans* PPI network, (b) distribution of semantic similarity scores for direct neighbors, indirect neighbors, and other domain pairs in the Struct DDI network.

with respect to network distance beyond a point can be explained by the decrease in the number of pairs with larger distance, which is likely to be an artifact of randomness. On the other hand, all other measures show a consistent decline in semantic similarity with respect to network distance, with saturation at distance ≥ 5. However, it is worth noting that the proposed information content based measure provides the sharpest decline in semantic similarity with increasing distance throughout, while it provides the sharpest decline for distance ≤ 3 when it is used with self-normalization.

In Figure 2(b), a comparison of the distribution of semantic similarity scores for the average information content (ρ_A/δ_I) and self-normalized information content (ρ_{JC}) measures is shown. In this figure, domain pairs are grouped according to their distances in the Struct DDI network, to obtain groups immediate neighbors (distance $= 1$), indirect neighbors (distance $= 2$), and other domain pairs (distance > 2). In the figure, the cumulative distribution of similarity score is shown for each group, *i.e.*, the vertical axis shows the fraction of domain pairs with similarity larger than the value on horizontal axis, where similarity scores are normalized to range from 0 to 1. Observe that, ρ_{JC} provides very large ($> 90\%$) similarity score for a much larger fraction ($> 60\%$) of neighboring domain pairs, as compared to ρ_A ($< 10\%$), while keeping fraction of highly similar domain pairs with distance > 2 considerably low ($< 10\%$). In general, the curves for ρ_{JC} demonstrate a sharper decline for similarity $\leq 20\%$ as compared to their counterparts

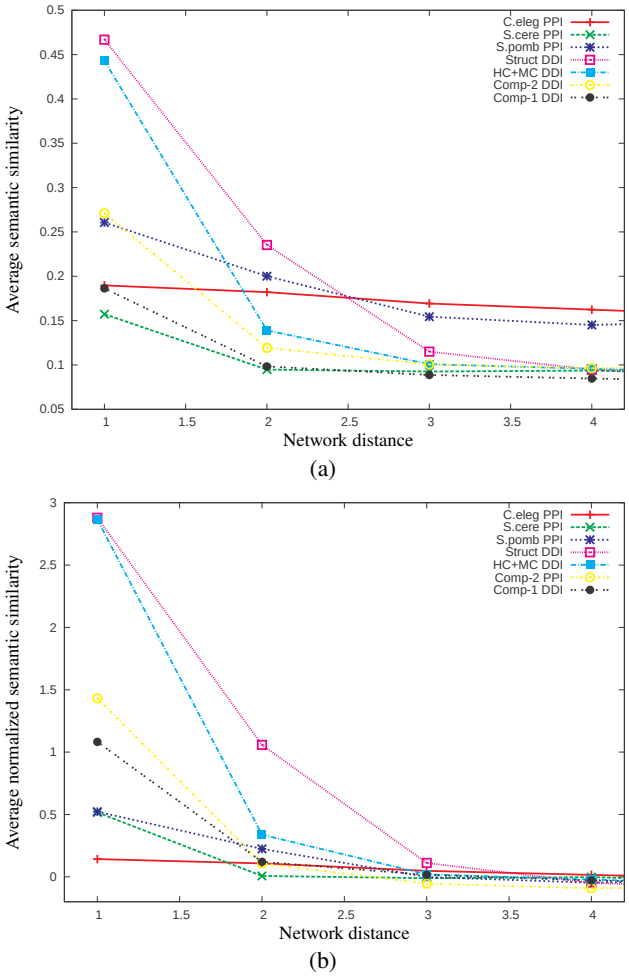


Fig. 3. Comparison of the relation between network proximity and semantic similarity with respect to molecular functions in PPI and DDI networks: (a) raw semantic similarity, (b) normalized semantic similarity with zero mean and unit standard deviation in each network. For distance 5,6 the similarity values are very close to that for distance 4. For annotation, see section 3.2.

for ρ_A and remain well above them, particularly for neighboring domains up to similarity $> 90\%$, illustrating that ρ_{JC} is more successful than ρ_A in reflecting the differences between (directly or indirectly) interacting and arbitrary pairs of domains, in terms of functional similarity.

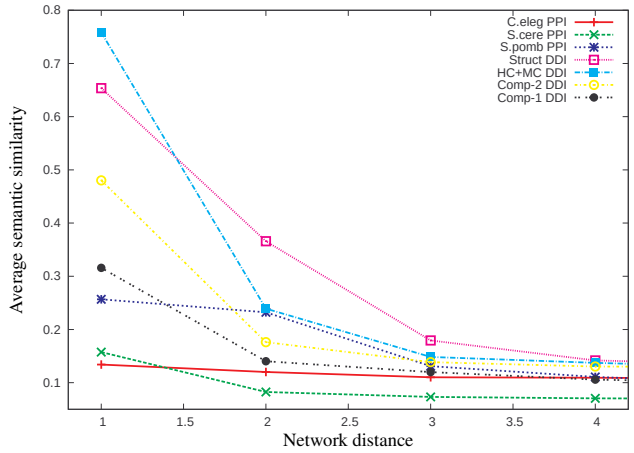
4.2 Comparison of PPI and DDI Networks

Using the proposed semantic similarity measure with self-normalization (ρ_{JC}), we compare the relationship between network proximity and functional similarity, using PPI and DDI networks described in section 3. We find that the following pairs of networks yield similar results: HC+MC and HC+NA DDI, *C. elegans*and *D. melanogaster*PPI, *S. cerevisiae*and *H. sapiens*PPI. For clarity, we do not display results for HC+NA DDI, *D. melanogaster*and *H. sapiens*PPI in Figure 3. The behavior of semantic similarity with respect to network distance is shown in Figure 3, for the molecular function sub-ontology of GO, *i.e.*, semantic similarity here refers to the similarity between the molecular functions of a pair of proteins or domains. Since the same semantic similarity measure is used for each network, the semantic similarity scores are compatible across different networks. The behavior of these raw semantic similarity scores for different networks is shown in Figure 3(a). Since the annotations of proteins and domains are largely incomplete, and the coverage of annotations may differ significantly across different networks, the distribution of semantic similarity scores can also vary significantly. For this reason, we normalize similarity scores using the procedure described in the previous section, to ensure that the similarity scores have zero mean and unit standard deviation in each network. The behavior of normalized similarity scores for different networks is shown in Figure 3(b).

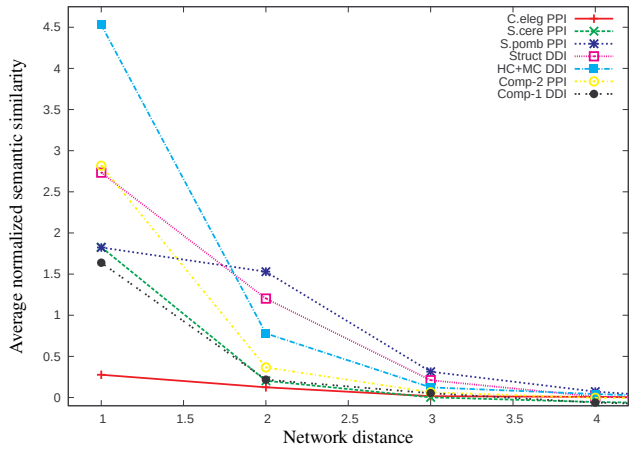
As evident in the figure, immediate and indirect neighbors perform (more) similar molecular functions. Furthermore, the negative correlation between network distance and functional similarity is stronger in the Struct DDI network, as compared to all other networks. This network is followed by other relatively more reliable HC+MC DDI network (and HC+NA DDI, not shown here). These observations suggest that network proximity is likely to be more relevant to, hence indicative of, functional coherence and modularity, However, this conclusion is tempered by the observation that the DDI networks that are based on structural information are relatively more reliable than PPI networks, which may come from noisy high-throughput screening.

The figure also shows that in PPI networks of relatively well-studied organisms such as *S. cerevisiae*, and *C. elegans*, functional similarity between two proteins that are further apart in the network is larger, on average, than that in the DDI and other PPI networks. This observation suggests that functional similarity between two arbitrary proteins in model organisms is expected to be larger than the functional similarity between two arbitrary domains or proteins in other organisms. This may be because more functional information is available for model organisms. As seen in Figure 3(b), network-based normalization alleviates this problem. Furthermore, after normalization, it becomes apparent that the relationship between functional similarity and network distance is stronger in computationally inferred DDI networks than that in PPI networks. Since computational inference of domain-domain interactions is generally based on protein-protein interactions, this observation provides further evidence that supports the notion that network proximity in DDI networks is likely to be a better indicator of functional modularity than PPI networks.

The behavior of semantic similarity with respect to network distance for the biological process sub-ontology of GO is shown in Figure 4. Here, semantic similarity refers to the similarity between the biological processes that a pair of proteins individually take part in. The behavior of process similarity with respect to network distance is generally similar to that of functional similarity, however, there are differences worth noting. First, when the similarity scores are not normalized with respect to network, the process similarity for arbitrary pairs of proteins in model organisms appears to be lower, on average, than that for arbitrary pairs of domains. This is in contrast to the argument based on annotation coverage. However, even after normalization, PPI networks demonstrate weaker relationship between network proximity and process similarity, as compared to DDI networks. Yet, the gap that is observed for functional similarity closes when processes are considered, particularly for the *S. pombe* PPI network, which shows similar process similarity between neighbors compared to computationally inferred DDI networks. Furthermore, indirect neighbors in the *S. pombe* PPI network



(a)



(b)

Fig. 4. Comparison of the relation between network proximity and semantic similarity with respect to biological processes in PPI and DDI networks. (a) Raw semantic similarity, (b) normalized semantic similarity with zero mean and unit standard deviation in each network. For annotation, see section 3.2.

have highest average process similarity among all networks considered. This might be indicative of the difference between molecular functions and biological processes in terms of their relationship to functional similarity. In general, it is possible to speculate that molecular function is a lower level property of a molecule that is directly related to its structure, while biological processes are higher level constructs, related to the wider neighborhood in the network. For this reason, while our results suggest that domain-domain interactions may be more informative in terms of identification of function and functional modularity, it may be necessary to consider DDI networks along with PPI networks to extract information about process modularity.

5 CONCLUSION

We investigate metrics for quantifying functional similarity in PPIs and DDIs. We present essential attributes of admissible metrics for term- and set-similarity, show that existing commonly used measures are not admissible, and present an admissible metric. We establish that the proposed metric provides highly intuitive biological interpretations from comprehensive comparative analysis of PPIs and DDIs. In doing so, we conclusively establish the metric, as well as validate the role of DDIs in quantifying functional coherence.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene Ontology: Tool for the unification of biology. the Gene Ontology consortium. *Nat Genet*, **25**(1), 25–29.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S. *et al.* (2004). The Pfam protein families database. *Nucleic Acids Research*, **32**, D138–D141.

Breitkreutz, B., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M. *et al.* (2007). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*

Camon, E., Barrell, D., Dimmer, E., and Lee, V. (2006). *In Silico Genomics and Proteomics: Functional Annotation of Genomes and Proteins.*, chapter The Gene Ontology Annotation (GOA) Database: Sharing Biological Knowledge with GO, pages 37–54.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**, 1540–1548.

Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J. *et al.* (2005). A protein domain interaction interface database: Interpare. *BMC Bioinformatics*, **6**.

Grossmann, S., Bauer, S., Robinson, P. N., and Vingron, M. (2006). An improved statistic for detecting over-represented gene ontology annotations in gene sets. In *RECOMB'06*, pages 85–98.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V. *et al.* (2004). Evidence for dynamically organized modularity in the yeast protein interaction network. *Nature*, **430**, 88–93.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *ICRCL*.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C. *et al.* (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, 297–302.

Koyutürk, M., Kim, Y., Subramaniam, S., Szpankowski, W., and Grama, A. (2006). Detecting conserved interaction patterns in biological networks. *J Comput Biol*, **13**(7), 1299–1322.

Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science*, **306**(5701), 1555–1558.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Lord, P., Stevens, R., Brass, A., and Goble, C. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Ng, S., Zhang, Z., Tan, S., and Lin, K. (2003). InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, **31**, 251–254.

Raghavachari, B., Tasneem, A., Przytycka, T., and Jothi, R. (2007). DOMINE: a database of protein domain interactions. *Nucleic Acids Res.*

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453.

Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**, R89.

Schlicker, A., Huthmacher, C., Ramrez, F., Lengauer, T., and Albrecht, M. (2007). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, **23**, 859–865.

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B., and Stoeckert, C. (2002). Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res.*, **12**, 648–655.

Sevilla, J., Segura, V., Podhorski, A., Guruceaga, E., Mato, J., Martnez-Cruz, L., Corrales, F., and Rubio, A. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform*, **2**, 330–338.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, **3**.

Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *PNAS*, **100**(21), 12123–12128.

Titz, B., Schlesner, M., and Uetz, P. (2004). What do we learn from high-throughput protein interaction data? *Expert Review of Proteomics*, **1**(1), 111–121.

Wuchty, S. (2006). Topology and weights in a protein domain interaction network–a novel way to predict protein interactions. *BMC Genomics*, **7**.