bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A New Family of Similarity Measures for Scoring Confidence of Protein Interactions using Gene Ontology

Madhusudan Paul and Ashish Anand

Abstract—The large-scale protein-protein interaction (PPI) data has the potential to play a significant role in the endeavor of understanding cellular processes. However, the presence of a considerable fraction of false positives is a bottleneck in realizing this potential. There have been continuous efforts to utilize complementary resources for scoring confidence of PPIs in a manner that false positive interactions get a low confidence score. Gene Ontology (GO), a taxonomy of biological terms to represent the properties of gene products and their relations, has been widely used for this purpose. We utilize GO to introduce a new set of specificity measures: Relative Depth Specificity (RDS), Relative Node-based Specificity (RNS), and Relative Edge-based Specificity (RES), leading to a new family of similarity measures. We use these similarity measures to obtain a confidence score for each PPI. We evaluate the new measures using four different benchmarks. We show that all the three measures are quite effective. Notably, RNS and RES more effectively distinguish true PPIs from false positives than the existing alternatives. RES also shows a robust set-discriminating power and can be useful for protein functional clustering as well.

Index Terms—Protein-protein interaction, semantic similarity measures, gene ontology, specificity, information content, set-discriminating power, KEGG pathways, ROC curve, Pfam.

1 Introduction

A significant amount of protein-protein interaction (PPI) data has become available due to high-throughput technologies. PPI data play a central role towards a systems-level understanding of cellular processes with important applications in disease diagnosis and therapy. A considerable fraction of interactions are false positives due to limitations of experiments used in detecting protein interactions [1]. Hence, a ranking or a scoring mechanism distinguishing between true and false interactions is important for any downstream analysis. There have been continuous efforts to utilize additional knowledge resources, such as Gene Ontology (GO) [2], in scoring confidence of PPIs in a manner that false positive interactions get a low confidence score [3]. The primary objective of this work is to introduce a new family of semantic similarity measures (SSMs) between gene products using GO for scoring confidence of PPIs.

GO has been effectively utilized in predicting and validating PPIs [4], [5], [6], and confidence scoring of PPIs [7], [8], [9], [10], [11], [12] among other genomic applications such as predicting protein functions [13], [14], [15], analyzing pathways [16] etc. It is a taxonomy of biological terms to represent the properties of genes and gene products (e.g., proteins) and is organized as a directed acyclic graph (DAG) to describe the relationship among the terms. GO is made up of three independent ontologies: biological process (BP), cellular component (CC), and molecular function (MF). A section of GO DAG (Release March 2015) is shown in the Supplementary Material. Terms closer to the root are more generic in nature and specificity of terms gradually increases as we move towards the leaves. The more specific a term is, the more informative it is. Ontology-based SSM is a quantitative function that measures the similarity between two terms based upon their relations over a set of terms organized as an ontology. Formally, it is a function of two ontology terms (or two sets of ontology terms) that returns a real number indicating the closeness between the terms in the context of semantic meaning [3]. Gene or gene products in different model organisms are annotated to GO terms based on various evidences and is available through annotation corpora. An annotation corpus of a species (e.g., yeast) is an association between gene products of the species and GO terms.

1.1 Motivation and Hypothesis

The notion of Information Content (IC) is widely used in defining SSMs. It quantifies specificity of a term in an ontology, i.e., how specific a term in an ontology is. The IC is explained formally in section 2. The IC based SSMs assume that the given ontology is complete and define specificity of a term by considering the whole ontology. However, GO is being updated regularly with the addition of new terms and removal of old terms. Furthermore, when new information of gene or gene product is discovered, annotation data corresponding to the appropriate terms are updated as well. Some proteins are annotated with a large number of terms, while many proteins are annotated to one term

- *M. Paul is with the Department of Computer and System Sciences, Visva-Bharati, Santiniketan 731235, West Bengal, India. He is also with the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India.*
  E-mail: madhusudan@iitg.ac.in
- *A. Anand is with the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India.*
  E-mail: anand.ashish@iitg.ac.in

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

only, i.e., annotations are not uniformly distributed among the terms (annotation bias). Thus the continuous evolution of the GO DAG, regular updates in annotation and non-uniform distribution of terms (as well as annotations) over the ontology are likely to impact confidence scores of several PPIs with each update.

A GO term is more closely related to its ancestors and descendants as the ontology is hierarchically organized. The major part of contribution towards specificity of a term is accumulated through the properties of its ancestors and descendants. Therefore for quantifying specificity of a term in an ontology like GO (which is very large, complex, continuously evolving and not uniformly distributed), it is safe to consider the properties of the subgraph consisting of the term itself along with its ancestors and descendants only instead of considering the whole ontology, to minimize the impact of continuous evolution.

Our main hypothesis is that the explicit encoding of the aforementioned unexplored subgraph-based specificity notions into a new family of SSMs could be useful for scoring confidence of PPIs.

1.2 Definition of the Problem and Contribution

The main problem of the current study is to define the specificity of a GO term, based on the properties of the subgraph consisting of the term itself along with its ancestors and descendants only, that could be useful for scoring confidence of PPIs.

With the aforementioned unexplored notion of specificity, we introduce three simple yet effective specificity measures: Relative Depth Specificity (RDS), Relative Node-based Specificity (RNS), and Relative Edge-based Specificity (RES). This new set of specificity measures led to a new family of SSMs.

We compare the performance of the new SSMs with six state-of-the-art SSMs proposed by Resnik [17], Lin [18], Schlicker *et al.* [19], Jiang & Conrath [20], Wang *et al.* [21], and Jain & Bader [22], referred to as *Resnik*, *Lin*, *Rel*, *Jiang*, *Wang*, and *TCSS*, respectively, in the rest of the paper. Resnik and TCSS have been considered to be the best SSMs for scoring confidence of PPIs by several studies such as Guo *et al.* [23], Xu *et al.* [24], Jain & Bader [22], and Pesquita [25]. We use four different benchmarks to evaluate the new SSMs. The four benchmarks are: 1) correlation with reference dataset from HIPPIE database [26], 2) ROC curve analysis with DIP database [27], 3) set-discriminating power of KEGG pathways [28], and 4) correlation with protein family (Pfam) using CESSM dataset [29]. The first benchmark is for human PPIs only as HIPPIE is an integrated database of human PPIs and the rest of the three benchmarks are applied to both yeast (S. cerevisiae) and human (H. sapiens) PPIs.

The rest of the paper is organized in the following manner. A brief survey of the literature is presented in section 2. The new family of SSMs is explained in section 3. Section 4 describes the experimental design, evaluation metrics, datasets used, implementation and results. In section 5, results are analyzed and discussed. Finally, section 6 introduces the conclusions and future work.

2 Related Work

This section introduces a brief review of the literature on PPI confidence scoring methods and GO-based SSMs. For an in-depth review of the family of GO-based SSMs, we refer the reader to the surveys by Pesquita *et al.* [3], Harispe *et al.* [30], Mazandu *et al.* [31], and Pesquita *et al.* [25].

2.1 PPI Confidence Scoring Methods

Computational approaches for scoring confidence of PPIs mainly differ in the selection of information used in the prediction model. The common sources of this information are three-dimensional protein structures [32], protein sequences [33], gene expression profiles [34], phylogenetic trees [35], [36], phylogenetic profiles [37], GO [7], [8], [9], [10], [11], [12] etc. Some approaches utilize topology of interaction network from already existing true PPIs [38], [39], [40]. Text mining on peer-reviewed literature is also used for scoring confidence of PPIs [41]. A few approaches utilize multiple sources of information [42], [43]. However, GO is a very comprehensive resource for the properties of gene products and their functional relationships across species. It provides a promising way to infer functional information of gene products. The idea of semantic similarity is a common way to utilize GO for scoring confidence of PPIs. Semantic similarity between two proteins (see section 2.3) involved in a PPI may be treated as a confidence score of the interaction. The current study is primarily focused on the SSMs by exploiting GO for scoring confidence of PPIs.

2.2 GO-based SSMs

Ontology-based SSMs were originally introduced in the fields of cognitive sciences by Tversky [44] and Natural Language Processing (NLP) and Information Retrieval (IR) by Rada *et al.* [45]. Since then a plethora of semantic similarity measures based on WordNet (a large lexical database of English) were developed such as the pioneering works introduced by Resnik [17], Jiang & Conrath [20] and Lin [18]. However, the first pioneering work was introduced by Lord et al. [46], [47] in the field of biology and this work has started the research on the development of GO-based SSMs and their applications in genomics such as [19], [21], [22], [48], [49], [50]. Here, we provide a brief overview of different SSMs.

Existing SSMs are classified broadly into two categories: *edge- and node-based* [3]. Edge-based measures are the natural and direct way of defining SSMs. Rada *et al.* [45] introduced a SSM of this kind in a lexical taxonomy, which was then applied in GO by Nagar and Al-Mubaid [48]. Subsequently, several edge-based SSMs have been developed and used in GO [51], [52], [53], [54]. In the edge-based approach, shared paths between two terms are primarily considered to compute the similarity between them. It assumes that terms at the same level have similar specificity and edges at the same level represent same semantic distances between two terms [3], which are seldom true in GO. Furthermore, an edge-based approach does not account annotation information of terms and entirely relies on the topological structure of the GO DAG. Hence edge-based methods are more sensitive to the intrinsic structure of the GO DAG.

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The most commonly used SSMs are node-based that compute the similarity between two terms by comparing their properties, common ancestors, or their descendants. As mentioned earlier, majority of the node-based approaches use the notion of information content (IC) to define the specificity of a term. The IC of a term t is defined as

$$IC(t) = -ln\ p(t) \tag{1}$$

where $p(t)$ is the probability or frequency of occurrence of t. Usually, the descendants of t are also considered for computing $IC(t)$. The probability of occurrence, $p(t)$ of term t in GO is defined as:

$$p(t) = \frac{|\{t\} \cup Des(t)|}{N} \tag{2}$$

where $Des(t)$ is the set of descendants of t and N is the number of terms in the ontology. Since gene products are annotated to terms in GO, $p(t)$ is estimated as the frequency of annotations of t, i.e.,

$$p(t) = \frac{|Ant(\{t\} \cup Des(t))|}{M} \tag{3}$$

where $Ant(T)$ is the set of annotations to the set of terms T and M is the total number of annotations in the GO. In words, it is the ratio of the number of annotations to t and its descendants to the total number of annotations. The aforementioned two definitions are commonly known as an *intrinsic* and *extrinsic* way of defining the probability function $p(t)$, respectively.

The most commonly used node-based SSMs are Resnik [17], Lin [18], and Jiang & Conrath [20], which were initially developed for WordNet and subsequently applied to GO by Lord *et al.* [46], [47]. Thereafter, a number of node-based SSMs have been proposed in order to improve the existing SSMs in different perspectives and applications [19], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66]. The major drawbacks of IC based SSMs are already pointed out in section 1.1. SSMs such as [21], [50], [67], [68] combine both node- and edge-based approaches and are commonly referred to as hybrid approaches. Recently, some complex structural-based SSMs are also developed [22], [69], [70].

2.3 SSM between Two Sets of Terms

A gene product may be annotated with more than one term in the same GO. Suppose, p_1 and p_2 are two gene products annotated to the set of terms S and T, respectively. The similarity between p_1 and p_2 are calculated as the similarity between two sets S and T, i.e., $SSM(p_1, p_2) = SSM(S, T)$. Therefore we need to combine GO terms of S and T. Generally, the following three types of strategies used in the literature:

Maximum (MAX) - In MAX strategy [71], similarity between S and T is calculated as the maximum of the set $S \times T$.

$$SSM_{MAX}(S,T) = \max_{s \in S, t \in T} SSM(s,t) \tag{4}$$

Average (Avg) - In 'average' strategy [46], [47], similarity between S and T is calculated as the average of the set $S \times T$.

$$SSM_{avg}(S,T) = \frac{\sum_{s \in S, t \in T} SSM(s,t)}{m \times n} \tag{5}$$

where $m = |S|$ and $n = |T|$.

Best-match average (BMA) - SSMs between two sets of terms form a matrix. BMA [19], [72], [73] is defined as the average of all maximum SSMs on each row and column of the matrix.

$$SSM_{BMA}(S,T) = \frac{\sum_{i=1}^{m} \max_{1 \le j \le n} SSM(s_i,t_j) + \sum_{j=1}^{n} \max_{1 \le i \le m} SSM(s_i,t_j)}{m+n} \tag{6}$$

where $s_i \in S$ and $t_j \in T$.

2.4 SSMs used in Evaluation

Resnik - Resnik considers IC of the *most informative common ancestor* (MICA) only [17]. The similarity between two terms s and t in Resnik is defined as

$$SSM_{Resnik}(s,t) = \max_{c \in C} IC(c) = IC(MICA(s,t)) \tag{7}$$

where C is the set of common ancestors of s and t, and IC is the information content defined earlier. It is the IC of the closest common ancestor or *lowest common ancestor* (LCA) of s and t.

Lin and Jiang - Although Resnik is very effective for computing information shared by two terms, it cannot distinguish between pairs of terms having the same MICA. To overcome the problem, Lin and Jiang are developed by considering ICs of both the terms along with their MICAs in different ways [18], [20]. The similarity between two terms is calculated by these two methods as

$$SSM_{Lin}(s,t) = \frac{2 \times IC(MICA(s,t))}{IC(s) + IC(t)}, \tag{8}$$

$$SSM_{Jiang}(s,t) = 1 - [IC(s) + IC(t) - 2 \times IC(MICA(s,t))]. \tag{9}$$

Rel - Lin and Jiang overestimate when one term is an ancestor of another. For example, when both the terms are same, the similarity score will be 1, irrespective of its specificity. Rel combines Resnik and Lin in order to capture relevance information by multiplying one minus the *extrinsic* probability of MICA to SSM_{Lin} [19]. As per Rel, the similarity between two terms is calculated as

$$SSM_{Rel}(s,t) = \frac{2 \times IC(MICA(s,t))(1 - p(MICA(s,t)))}{IC(s) + IC(t)}. \tag{10}$$

Wang - Wang is a hybrid measure that combines both edge- and node-based approaches [21]. Let $G_t = (V_t, E_t)$ be a DAG for a term t in GO such that V_t is the set of ancestors of t including t itself and E_t is the set of edges connecting terms in G_t. Terms closer to term t in G_t contribute more of its semantics to the semantics of term t. The semantic contribution of a term c to the semantics of term t in G_t is denoted as S-value of c or $S_{G_t}(c)$ and defined as:

$$\begin{cases} S_{G_t}(t) = 1 \\ S_{G_t}(c) = \max\{w_e \times S_{G_t}(c') : c' \in \text{ children of } c\} \text{ if } c \neq t \end{cases} \tag{11}$$

where w_e $(0 < w_e < 1)$ is semantic contribution factor for edge $e \in E_t$ from term c' to term c. For example, semantic contribution factors (w_e) of *is_a* and *part_of* relationships may be treated as 0.8 and 0.6, respectively. To compare

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

semantics of two terms, a semantic value $SV(t)$ is computed as the aggregate contribution of the semantics of all the terms in G_t to term t and defined as:

$$SV(t) = \sum_{c \in V_t} S_{G_t}(c). \quad (12)$$

Now, SSM between two terms s and t with respect to their DAGs $G_s = (V_s, E_s)$ and $G_t = (V_t, E_t)$ is defined as:

$$SSM_{Wang}(s,t) = \frac{\sum_{c \in V_s \cap V_t}(S_{G_s}(c) + S_{G_t}(c))}{SV(s) + SV(t)}. \quad (13)$$

The numerator is the summation of S-values of common terms between the two DAGs. S-values of common terms between the two DAGs may not be same as the locations of s and t may differ in GO.

TCSS - TCSS exploits the unequal depth of biological knowledge representation in different branches of GO DAG [22]. The objective of TCSS is to identify subsets of similar GO terms (e.g., terms related to nucleus and terms related to mitochondrion belong two different subsets) and score PPIs higher if participating proteins belong to the same subset compared to PPIs whose participating proteins belong to different subsets. The authors have introduced a structural-based IC, referred to as topological information content (ICT), to identify sub-graph root terms during pre-processing stage.

$$ICT(t) = -ln\Big(\frac{|Child(t)|}{N}\Big) \quad (14)$$

where $Child(t)$ is the set of children of t and N is the number of terms in the ontology.

3 The New GO-based SSMs

In this section, we introduce the new family of SSMs based on the proposed set of specificity measures. To define specificity of a GO term we consider the properties of the subgraph consisting of the term itself along with its ancestors and descendants only and ignore the rest of the ontology. The new specificity models quantify how specific a term in ontology is. The specificity of a parent (term) always will be less than any of its children. RDS considers a specific path of the aforementioned subgraph, while RNS and RES consider the whole subgraph. However, RNS relies on the properties of the nodes only, whereas RES considers the edges of the subgraph as well.

3.1 Relative Depth Specificity (RDS)

Let $d_{t,r}$ and $d_{l,t,r}$ are length of the longest path from term t to the root r and length of the longest path from any leaf l to the root r via the term t, respectively. Then, RDS of a term t in GO is defined as

$$RDS(t) = \frac{d_{t,r}}{d_{l,t,r}} = \frac{d_{t,r}}{d_{l,t} + d_{t,r}}. \quad (15)$$

In words, $RDS(t)$ is the ratio between the length of the longest path from the term t to the root and the length of the longest path from any leaf to the root via the term t. This is the simplest SSM that does not consider annotation information. The specificity of the leaves and the root would be highest (1) and lowest (0), respectively. When multiple paths are present between two terms, we consider the longest one as it is likely to be more informative than others.

3.2 Relative Node-based Specificity (RNS)

Let $G_1(V_1, E_1)$ be the subgraph consisting of the term t itself along with its ancestors; and $G_2(V_2, E_2)$ be the subgraph consisting of the term t itself along with its ancestors and descendants. The RNS of a term t in GO is defined as

$$RNS(t) = \frac{|Ant(V_1)| + |V_1|}{|Ant(V_2)| + |V_2|} \quad (16)$$

where $Ant(T)$ be the set of annotations to the set of terms T as mentioned earlier. In words, it is the ratio of the sum of nodes along with its annotations of the subgraph consisting of the term t and its ancestors to the sum of nodes along with its annotations of the subgraph consisting of t, its ancestors and descendants. Thus, RNS of the leaves and the root would be highest (1), and lowest (close to 0), respectively.

3.3 Relative Edge-based Specificity (RES)

We define the weight of an edge $e(t_1, t_2)$ between terms t_1 and t_2 in GO as:

$$w(e) = |Ant(\{t_1\})| + |Ant(\{t_2\})|. \quad (17)$$

It is the summation of the number of annotations of terms t_1 and t_2. The weight of a set of edges E is defined as:

$$W(E) = \sum w(e_i) : e_i \in E. \quad (18)$$

It is the summation of weights of all edges in the set of edges E. Let $G_1(V_1, E_1)$ be the subgraph consisting of the term t itself along with its ancestors and $G_2(V_2, E_2)$ be the subgraph consisting of the term t itself along with its ancestors and descendants as in RNS. The Relative Edge-based Specificity of a term t in GO is defined as

$$RES(t) = \frac{W(E_1) + |E_1|}{W(E_2) + |E_2|}. \quad (19)$$

In words, it is the ratio of the summation of weighted and unweighted edges of the subgraph consisting of the term t itself along with its ancestors to the summation of weighted and unweighted edges of the subgraph consisting of t itself along with its ancestors and descendants. Thus, specificity of the leaves and the root would be highest (1), and lowest (0), respectively.

The similarities between the two terms s and t are calculated as:

$$SSM_{RDS}(s,t) = \max_{c \in C} RDS(c) = RDS(MICA(s,t)), \quad (20)$$

$$SSM_{RNS}(s,t) = \max_{c \in C} RNS(c) = RNS(MICA(s,t)), \quad (21)$$

$$SSM_{RES}(s,t) = \max_{c \in C} RES(c) = RES(MICA(s,t)) \quad (22)$$

where C is the set of common ancestors of s and t as mentioned earlier.

We have chosen the MICA to define the shared specificity between the two terms similar to Resnik. It is noteworthy to mention that the proposed specificity models are different from IC models as they do not rely on probability functions. Therefore we cannot directly apply the new specificity models to other IC-based similarity measures such as Lin, Rel, and Jiang.

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4 EVALUATION

In this section, we detail the experimental design, evaluation metrics, datasets used, implementation and results. As already mentioned, six state-of-the-art SSMs are chosen as baseline methods and four benchmarks are considered for evaluation of the new SSMs.

4.1 Experimental Setup

Our experimental design for evaluation is based on the following two assumptions. First, two proteins involved in the same biological process(es) are more likely to interact than proteins involved in different processes [5, p.953] and [22]. Second, two proteins need to come in close proximity (at least transiently) for interaction, hence co-localization also provides evidence of interaction [74, p. 689] and [22]. However, if two proteins interact physically, there is no guarantee that they share the same molecular function [75, p. 27]. The 'average' strategy underestimates when two gene products share many similar terms as it considers all possible term pairs of the two gene products [76]. By contrast, the MAX strategy overestimates when two gene products share few similar terms as it is indifferent to the number of dissimilar terms between the gene products [76]. The BMA strategy, which considers both similar and dissimilar terms [76], does not suffer from the aforementioned limitations. Further, in PPIs, proteins need to be in close proximity (share similar CC terms) and participate in the same biological process (share similar BP terms) once, among all possible combinations, to become biologically relevant [22]. Hence MAX and BMA are considered better strategies than the 'average' for scoring confidence of PPIs. In light of the above discussion, we use BP and CC ontologies of GO along with MAX and BMA strategies for performance evaluation. We exclude electronically inferred annotations (IEA) of GO terms which lack manual curation. We consider only those protein pairs which are having both the proteins annotated with at least one GO term other than the root in their respective ontologies.

As mentioned earlier, the new SSMs are evaluated on the four benchmarks: 1) correlation with reference dataset from HIPPIE database, 2) ROC curve analysis in predicting true PPIs from DIP database, 3) set-discriminating power of KEGG pathways, and 4) correlation with Pfam on CESSM dataset. Evaluation is done using both yeast (S. cerevisiae) and human PPIs except for the first benchmark, as it contains only human PPIs. We use Entrez and ORF gene ids for human and yeast, respectively, except while comparing with TCSS where UniProtKB and SGD gene ids were used for human and yeast, respectively. We have not performed the comparison with TCSS on the second and third benchmarks (for human) as some UniProtKB ids (after mapping from Entrez ids) were not found in the annotation.

4.2 Evaluation Metrics and Baselines

This section introduces how and why each benchmark is used for evaluation. A brief outline and formulation of each metrics used are presented here.

4.2.1 Correlation with Reference Dataset from HIPPIE Database

The HIPPIE database [26] integrates most of the publicly available PPI databases like BioGRID [77], DIP [78], HPRDS [79], IntAct [80], MINT [81], BIND [82], MIPS [83]. It also includes interactions from several manually selected studies. The HIPPIE score of a PPI is defined by considering the following parameters: the number of studies where the PPI was detected, the number and quality of the experimental techniques used to detect the PPI, and the number of non-human organisms where the PPI was reproduced. The authors of HIPPIE showed that their scoring scheme of interactions correlates with the quality of the experimental characterization. We use a reference dataset from HIPPIE database to evaluate different SSMs. Pearson correlation is calculated between the HIPPIE score and PPI confidence score obtained using an SSM.

4.2.2 ROC Curve Analysis

Similarity measures can be treated as binary classifiers to classify a given PPI as positive or negative with a reasonable cutoff similarity score. PPIs having similarity score greater than the cutoff are treated as positive. Receiver operating characteristic (ROC) curve analysis is used to evaluate the performance of a binary classifier. ROC curve is a graph plotting of true positive rate (TPR or sensitivity) against false positive rate (FPR or 1-specificity) by varying discrimination threshold or cutoff. The area under the ROC curve (AUC) is the measure of discrimination, i.e., the ability of the classifier to classify correctly. An AUC of 1 represents perfect classifier. We utilize the core subsets of yeast and human PPIs from the DIP database [27] to evaluate different SSMs for AUCs.

4.2.3 Set-discriminating Power of KEGG Pathways

A biological pathway is a sequence of biochemical steps to accomplish a specific biological process within a cell. Therefore proteins involved in a pathway are more likely to interact among themselves than the proteins belonging to different pathways. Proteins within a pathway are likely to be annotated with the same or similar terms in GO too and should show high similarity scores. We consider three sets of selected KEGG pathways [28] to evaluate different SSMs for their discriminating power as discussed in the following paragraph.

For each KEGG pathway, an *intra-set average similarity* is calculated as the average of all pairwise similarities of proteins within the pathway. An *inter-set average similarity* for every two pathways is also calculated as the average of all pairwise SSMs of proteins between the two pathways. During calculation of *inter-set average similarity*, we do not consider those pairs whose both the proteins are same. A discriminating power (DP) of a pathway is defined in [84] as the ratio between *intra-set average similarity* and the average of all *inter-set average similarities* between the chosen pathway and rest other pathways. Let $\mathcal{P} = \{P_1, P_2, ..., P_n\}$

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be the set of KEGG pathways, each pathway P_k contains m_k number of proteins and p_{ki} denotes i^{th} protein in P_k.

$$Intra\_set\_avg\_sim(P_k) = \frac{\sum_{i=1}^{m_k}\sum_{j=1}^{m_k} SSM(p_{ki}, p_{kj})}{m_k^2}. \quad (23)$$

$$Inter\_set\_avg\_sim(P_k, P_l) = \frac{\sum_{i=1}^{m_k}\sum_{j=1}^{m_l} SSM(p_{ki}, p_{lj})}{m_k \times m_l}. \quad (24)$$

$$DP(P_k) = \textit{Intra-set average similarity of } P_k / \text{Avg. of all } \textit{inter-set average similarities} \text{ between } P_k \text{ and other pathways} = \frac{(n-1) \times Intra\_set\_avg\_sim(P_k)}{\sum_{i=1, i\neq k}^{n} Inter\_set\_avg\_sim(P_k, P_i)}. \quad (25)$$

4.2.4 Correlation with Protein Family (Pfam)

A protein family (Pfam) is a group of proteins that are evolutionarily related, i.e., they share a common evolutionary ancestor. Proteins belonging to a family often show functional similarity. The Jaccard index is used to calculate Pfam similarity. The Jaccard index of two proteins is calculated as the ratio of the number of protein families they share to the total number of protein families they belong. We utilize dataset of protein pairs used in CESSM [29]. For each pair, Pfam similarity (Jaccard index) and similarity scores of different SSMs are calculated and finally, the Pearson correlation between the two scores is obtained.

4.3 Datasets

In this section, we describe the sources of different datasets used in the evaluation and the corresponding preprocessing steps. A summary of the datasets used is presented in Table 1.

4.3.1 Reference Dataset from HIPPIE Database

We download Human Integrated Protein-Protein Interaction rEference (HIPPIE) dataset on 09.01.2015 [26]. We extract one reference dataset from HIPPIE consisting of PPIs detected by four top-scored experimental techniques: far-Western blotting, isothermal titration calorimetry, nuclear magnetic resonance, and surface plasmon resonance experiments as in [85]. The interaction detected by any of the chosen four experimental techniques have a high probability of being an actual interaction [85]. The number of PPIs present in the reference datasets is shown in Table 1.

4.3.2 Datasets for ROC Curve Analysis

We download the core subsets of PPIs from the Database of Interacting Proteins (DIP) [27] for S.cerevisiae and H.sapiens on 29.10.2015. DIP is a database of experimentally detected PPIs from various sources. We assume that these interactions are real and treat them as positive instances of interactions. DIP uses UniProt Ids for proteins. We map UniProt Ids into Entrez and ORF gene Ids for human and yeast, respectively. Table 1 shows the number of PPIs of DIP dataset used in this study. As done in [22], an equal number of negative PPI datasets are independently generated by randomly choosing protein pairs annotated in BP and CC, and are not present in the iRefWeb database [86] (version date: 27.11.2015), a combined database of all known PPIs.

TABLE 1
Summary of Datasets used in Evaluation

Benchmark datasets	Species	Ontology	Number of PPIs or protein pairs or length of pathways
HIPPIE	Human	BP	1748
		CC	1757
DIP	Yeast	BP	4962
		CC	4992
	Human	BP	4279
		CC	4283
Pfam	Yeast	BP	366
		CC	351
	Human	BP	1212
		CC	1211
KEGG	Yeast Set-1	-	11 - 14
	Yeast Set-2		Specified in Table 2
	Human	-	11 - 16

4.3.3 KEGG Pathways

We extract two sets of KEGG pathways [28] of each of the two organisms, S.cerevisiae and H.sapiens, using org.Sc.sgd.db and org.Hs.eg.db packages with R 3.1.2 version. The first set contains a number of genes between 11 to 14 and the second set 11 to 16. We choose the above ranges so that each set contains the same (11) number of pathways and takes a reasonable time to compute. The two sets have three common pathways: Terpenoid backbone biosynthesis (sec00900 and hsa00900), Riboflavin metabolism (sec00740 and hsa00740), and Pantothenate and CoA biosynthesis (sec00770 and hsa00770). However, each of them is from different organisms and may not show similar results. Another set of 11 yeast KEGG pathways (Table 2) with more diverse functionality is also considered to get a broader insight into the inter-set discriminating power.

4.3.4 CESSM Dataset for Correlation with Pfam

The Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) is an online tool for evaluation of GO-based SSMs against sequence, Pfam and EC similarities [29]. Since CESSM has been published around ten years ago, it uses ten years old dataset (August 2008 GO and GOA-UniProt). In the meanwhile, GO DAG, its annotation, as well as Pfam have substantially changed. Moreover, we use GO.db (version:3.1.2) and org.Hs.eg.db (version:3.1.2) R packages that utilize March 2015 GO and annotations, respectively, in the evaluation. Hence we could not use CESSM automated tool. However, we utilize the dataset of protein pairs used in CESSM to find correlation against Pfam similarity only, since GO captures the functional aspect of gene or gene products primarily. All pairs of proteins are mapped into Entrez and ORF gene Ids for human and yeast, respectively. The dataset involves 13,430 protein pairs of 1,039 proteins from various species. The authors of CESSM reported that both proteins of each pair are manually annotated to at least one term within all the three GOs with a uniform IC of at least 0.5 and have at least one EC class and one Pfam class. The number of protein pairs used for this evaluation is shown in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLE 2
List of 11 Yeast Pathways with More Diverse Functionality used in the Study.

Category	Subcategory	Pathway Id	Pathway Name	No. of Genes
Metabolism	Carbohydrate metabolism	sce00040	Pentose and glucuronate interconversions	10
	Energy metabolism	sec00920	Sulfur metabolism	15
	Lipid metabolism	sec00565	Ether lipid metabolism	5
	Amino acid metabolism	sec00360	Phenylalanine metabolism	9
	Glycan biosynthesis and metabolism	sec00514	Other types of O-glycan biosynthesis	13
	Metabolism of cofactors and vitamins	sec00750	Vitamin B6 metabolism	11
	Metabolism of terpenoids and polyketides	sec00900	Terpenoid backbone biosynthesis	13
	Metabolism of other amino acids	sec00410	beta-Alanine metabolism	8
Genetic Information Processing	Folding, sorting and degradation	sec04122	Sulfur relay system	8
	Replication and repair	sec03450	Non-homologous end-joining	10
Environmental Information Processing	Signal transduction	sec04070	Phosphatidylinositol signaling system	15

4.4 Implementation

The new SSMs are implemented in the R programming language [87]. We use GOSemSim R package (version: 1.26.0) [88] for implementations of Resnik, Lin, Rel, Jiang, and Wang SSMs. For GO and corresponding annotations, we use GO.db, org.Sc.sgd.db (for yeast), and org.Hs.eg.db (for human) R packages (version:3.1.2, March, 2015 release) [89], [90], [91]. We maintain versions of all R packages so that they use same GO and corresponding annotations. For TCSS, we use the implementation provided by the authors with the default set of parameters. The original implementation of TCSS uses MAX strategy only. Therefore we modify it to include BMA strategy as well. The implementation of TCSS needs the ontology and annotation as text files provided by Gene Ontology Consortium. Therefore we use the released version of GO (gene_ontology.obo) dated Mar 13, 2015. The same released version of GO is used in above R packages (version: 3.1.2) and annotation for yeast (gene_association.sgd) and human (gene_association.goa_human) released on Mar 17, 2015. We use ROC and ROCR R packages [92], [93] to plot the ROC curve and to calculate the area under ROC curves (AUC).

4.5 Results

Performance, in terms of Pearson correlation, of different SSMs with respect to the reference dataset from HIPPIE are shown in Table 3 (See Supplementary Material for barplots). The best correlations are shown in bold.

AUC obtained by different SSMs are tabulated in Table 4. The corresponding ROC curves are provided in the Supplementary Material. The best ROC scores are shown in bold.

As discussed earlier, the discriminating power quantifies the ability of an SSM to distinguish among various functionally different sets of proteins (e.g., KEGG pathways). Fig 1 and 2 demonstrate the discriminating power of different SSMs with BMA strategy against KEGG pathways in BP and CC ontology, respectively. Instead of pathway names, KEGG pathway identifiers are shown along the x-axis. The discriminating power for the selected yeast KEGG pathways (listed in Table 2) with more diverse functionality is shown in Fig 3. The results with MAX strategy are quite similar and kept in the Supplementary Material. The data tables are also provided in the Supplementary Material.

Finally, Table 5 demonstrates the performance of different SSMs on Pfam (See Supplementary Material for barplots). The best scores are shown in bold.

5 Discussion

This section analyzes and discusses the results presented in section 4.5. We have highlighted the key observations.

5.1 Correlation with Reference Dataset from HIPPIE Database

RDS achieves the highest correlation in BP, while TCSS shows the maximum correlation in CC. It may be noted that RDS is the simplest SSM among the proposed measures and does not even consider annotation information. Nevertheless, it shows good correlation. RNS and RES also perform quite well in BP, while Resnik shows good performance in both BP and CC.

All SSMs show greater correlations in BP. The average correlation over all SSMs in BP is 0.311/0.259 (MAX/BMA), whereas in CC it is 0.137/0.192 (MAX/BMA). However, all measures show less overall correlation since correlation is computed for positive PPIs only.

5.2 ROC Curve Analysis

RNS and RES, with both MAX and BMA strategies, effectively classify true PPIs from false in both BP and CC. Resnik-MAX and Rel-MAX too perform well compared to others, while RDS shows competitive performance. Although we could not compare TCSS for human, it performs well with MAX strategy in yeast. All SSMs with MAX strategy have quite similar AUCs in BP for both yeast and human. However, with BMA strategy, AUCs achieved by RES (yeast:0.893, human:0.898) and RNS (yeast:0.890, human:0.903) are significantly higher than others. Further, RES and RNS exhibit greater consistency, since they show

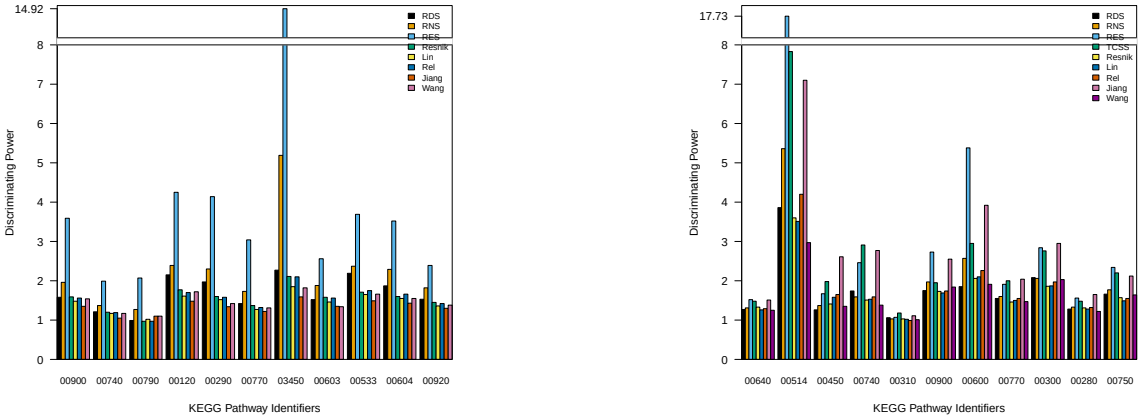
bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLE 3
Pearson Correlation with Reference Dataset from HIPPIE Database

Ontology	Strategy	RDS	RNS	RES	TCSS	Resnik	Lin	Rel	Jiang	Wang
BP	MAX	**0.358**	0.313	0.346	0.342	0.329	0.277	0.277	0.272	0.286
	BMA	**0.342**	0.332	0.310	0.270	0.238	0.220	0.218	0.211	0.193
CC	MAX	0.204	0.130	0.129	**0.232**	0.231	0.064	0.100	0.064	0.082
	BMA	**0.254**	0.227	0.198	0.232	0.230	0.148	0.164	0.118	0.158

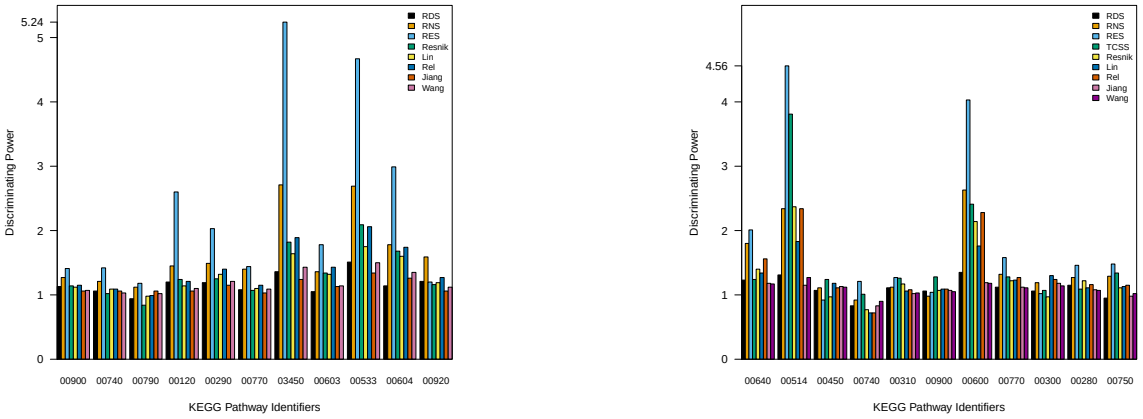
TABLE 4
The area under ROC curves of different SSMs

Species	Ontology	Strategy	RDS	RNS	RES	TCSS	Resnik	Lin	Rel	Jiang	Wang
Yeast	BP	MAX	0.896	0.908	0.903	0.907	0.908	0.912	**0.914**	0.910	0.895
		BMA	0.868	0.890	**0.893**	0.861	0.879	0.881	0.883	0.874	0.860
	CC	MAX	0.856	0.868	0.850	0.866	**0.870**	0.804	0.868	0.771	0.799
		BMA	0.826	0.848	0.843	0.831	**0.850**	0.805	0.838	0.709	0.783
Human	BP	MAX	0.907	**0.914**	0.904	-	0.908	0.900	0.913	0.887	0.895
		BMA	0.892	**0.903**	0.898	-	0.872	0.865	0.869	0.817	0.867
	CC	MAX	0.848	0.847	0.857	-	0.852	0.794	**0.858**	0.795	0.800
		BMA	0.824	**0.849**	**0.850**	-	0.814	0.773	0.791	0.708	0.791



(a) Human (b) Yeast

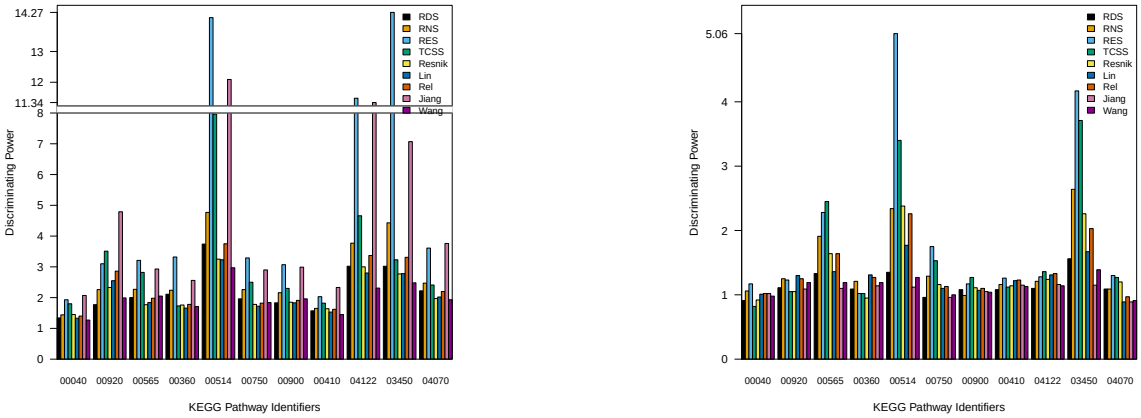
Fig. 1. **Inter-set discriminating power of different SSMs with BMA strategy in BP ontology.** The y-axis is splitted to accommodate high DP value.



(a) Human (b) Yeast

Fig. 2. **Inter-set discriminating power of different SSMs with BMA strategy in CC ontology.**

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) BP ontology

(b) CC ontology

Fig. 3. **Inter-set discriminating power of different SSMs with BMA strategy for the selected 11 yeast KEGG pathways with more diverse functionality.**

TABLE 5
Correlation with Protein Family (Pfam) on CESSM dataset

Species	Ontology	Strategy	RDS	RNS	RES	TCSS	Resnik	Lin	Rel	Jiang	Wang
Yeast	BP	MAX	0.280	**0.324**	0.283	0.290	0.304	0.308	0.314	0.268	0.302
		BMA	0.306	**0.347**	0.310	0.279	0.307	0.296	0.299	0.272	0.264
	CC	MAX	0.240	0.202	0.252	**0.259**	0.243	0.156	0.183	0.123	0.139
		BMA	0.218	0.204	**0.233**	0.204	0.225	0.226	0.225	0.205	0.201
Human	BP	MAX	0.158	0.157	0.160	0.258	**0.300**	0.152	0.156	0.143	0.156
		BMA	0.231	0.290	0.308	**0.347**	0.302	0.263	0.262	0.258	0.293
	CC	MAX	0.308	0.233	**0.390**	0.314	0.307	0.193	0.223	0.159	0.198
		BMA	0.356	0.383	**0.471**	0.437	0.347	0.349	0.365	0.269	0.349

less difference between MAX and BMA strategies in both BP and CC (for both yeast and human).

All SSMs show higher AUCs in BP. The average AUCs in BP are 0.906/0.877 (yeast:MAX/BMA) and 0.904/0.873 (human:MAX/BMA), whereas in CC these are 0.839/0.815 (yeast:MAX/BMA) and 0.831/0.800 (human:MAX/BMA).

5.3 Set-discriminating Power of KEGG Pathways

The discriminating power of RES is significantly higher than other SSMs for all the 11 human KEGG pathways. RES produces DP value greater than or equal to 1.81/1.99 (MAX/BMA) in BP, while the next minimum DP value is 1.17 (produced by RDS - MAX).

RES shows maximum functional discrimination among the pathways. RES produces very high DP value with 11.10/14.92 (MAX/BMA) for *Non-homologous end-joining* (hsa03450) pathway. This is the only pathway that belongs to the *Genetic Information Processing* category, while rest fall in the same *Metabolism* category. So, the functional characteristic of *Non-homologous end-joining* pathway is completely different from the rest. RES nicely captures this functional discrimination by producing very high DP value.

All SSMs produce greater DP values in BP. Although RES almost consistently produces higher DP values in both BP and CC (with both MAX and BMA), it shows comparatively lower DP values in CC.

Overall discriminating power of all the SSMs are quite similar and not so good for the first set of yeast KEGG pathways. If we examine the functional categories of all the 11 pathways, we find that all belong to the same *Metabolism* category with six pathways from two subcategories only. Further, the selected first set of yeast pathways contain merely 134 genes with 16 are shared. In contrast, the selected human pathways include 150 genes with 11 are common only. Hence the selected first set of yeast pathways are functionally closer to each other and this fact is reflected by low DP values.

To study further, we consider another set of 11 yeast pathways with more diverse functionality, where three pathways (sec00514, sec00750, and sec00900) were taken from the previous set. The pathways are listed in Table 2 and corresponding discriminating power for BMA strategy is shown in Fig 3 (See Supplementary Material for data table).

The discriminating power of all the SSMs is improved significantly for the pathways with more diverse functionality. In particular, DP values of RES and Jiang are higher than other measures for almost all the pathways. RES and Jiang produce DP value greater than or equal to 2/1.93 (MAX/BMA) and 1.84/2.07 (MAX/BMA), respectively, in BP, while the next minimum DP value is 1.73 (produced by TCSS - BMA). The maximum DP value (MAX/BMA:13.75/14.27 in BP) is again produced by RES for the pathway sec03450 (Non-homologous end-joining).

RES can be used for functional clustering. It may be noted that although Jiang produces competitive DP values with RES for yeast pathways, it is unable to show good DP

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

values for the human pathways. Therefore RES might be used for functional clustering (e.g., to characterize protein functional modules) as it shows consistently high discriminating power.

No SSM produces consistently good DP values in CC, particularly for the yeast pathways. Guo et al. [23] observed that all pairs of proteins involved in the same KEGG pathway have significantly higher similarity scores than randomly selected in BP, whereas similarity decreases exponentially as the distance between two proteins increases within the same pathway in CC and MF. These findings conform with current results as well.

5.4 Correlation with Pfam

Overall performances of TCSS, RES, and Resnik are well. Particularly, TCSS - MAX, RES - BMA, and Resnik - MAX perform well. Although RES does not show good correlation with MAX strategy in human, it produces a good correlation with BMA strategy. MAX strategy could overestimate while computing the general measure of functional similarity [22] and protein family captures a general aspect of protein function. Thus, BMA might be a better choice than MAX for Pfam similarity.

Further, it may be noted that correlation in CC is higher than BP in human for all measures which are quite unexpected. Therefore it might be challenging to draw comparative inference for the benchmark like Pfam that adopt a very general aspect of protein function with Jaccard index.

6 Conclusions and Future Work

The paper presents a new family of SSMs for scoring confidence of PPIs utilizing GO. This new family of SSMs is based on a new set of specificity measures namely, RDS, RNS and RES. Specificity of a term is redefined by considering the properties of its ancestors and descendants only along with its own properties so that maximum unwanted noises could be avoided. The evaluation shows that instead of simplicity, they are quite effective. Particularly, RNS and RES more effectively distinguish true interactions from false. RES can be useful for protein functional clustering as well since it shows a robust set-discriminating power over KEGG pathways. It also exhibits greater consistency and shows the best performance in BP with BMA strategy. Similar to the earlier studies, our evaluation also shows Resnik is one of the best SSMs for scoring confidence of PPIs. TCSS with MAX strategy and Rel also show competitive performance. Although RDS is the simplest SSM that does not even consider annotation information, it shows competitive performance as well. For almost all the four benchmarks, each SSM shows comparatively greater and consistent performances in BP. Therefore we believe that BP is more suitable than CC for scoring confidence of PPIs.

Although the newly developed SSMs are evaluated only on GO for scoring confidence of PPIs, it is not limited to any particular ontology. Therefore it would be worthy to evaluate how these SSMs perform on other ontologies and applications as future work.

Availabilty of Data and Script

An R script for the new SSMs along with the complete datasets used in the evaluation is freely available at https://github.com/msp-cse/NaiveSSMs.

Acknowledgments

The authors would like to thank Prof. V. Vijaya Saradhi for his valuable comments and suggestions.

References

[1] R. Gentleman, W. Huber *et al.*, "Making the most of high-throughput protein-interaction data," *Genome Biology*, vol. 8, no. 10, p. 112, 2007.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[3] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS computational biology*, vol. 5, no. 7, p. e1000443, 2009.

[4] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[5] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, "Probabilistic model of the human protein-protein interaction network," *Nature biotechnology*, vol. 23, no. 8, pp. 951–959, 2005.

[6] X. Wu, L. Zhu, J. Guo, D.-Y. Zhang, and K. Lin, "Prediction of yeast protein–protein interaction network: insights from the gene ontology and annotations," *Nucleic acids research*, vol. 34, no. 7, pp. 2137–2150, 2006.

[7] S. R. Maetschke, M. Simonsen, M. J. Davis, and M. A. Ragan, "Gene ontology-driven inference of protein–protein interactions using inducers," *Bioinformatics*, vol. 28, no. 1, pp. 69–75, 2012.

[8] G. D. Montanez and Y.-R. Cho, "Assessing reliability of protein-protein interactions by gene ontology integration," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*. IEEE, 2012, pp. 21–27.

[9] G. Montañez and Y.-R. Cho, "Predicting false positives of protein-protein interaction data by semantic similarity measures §," *Current Bioinformatics*, vol. 8, no. 3, pp. 339–346, 2013.

[10] G. Cui and K. Han, "Scoring protein-protein interactions using the width of gene ontology terms and the information content of common ancestors," *Emerging Intelligent Computing Technology and Applications*, vol. 2, pp. 31–36, 2013.

[11] G. Cui, B. Kim, S. Alguwaizani, and K. Han, "Assessing protein-protein interactions based on the semantic similarity of interacting proteins," *International journal of data mining and bioinformatics*, vol. 13, no. 1, pp. 75–83, 2015.

[12] S.-B. Zhang and Q.-R. Tang, "Protein–protein interaction inference based on semantic similarity of gene ontology terms," *Journal of theoretical biology*, vol. 401, pp. 30–37, 2016.

[13] L. J. Jensen, R. Gupta, H.-H. Staerfeldt, and S. Brunak, "Prediction of human protein function according to gene ontology categories," *Bioinformatics*, vol. 19, no. 5, pp. 635–642, 2003.

[14] Y. Chen and D. Xu, "Genome-scale protein function prediction in yeast saccharomyces cerevisiae through integrating multiple sources of high-throughput data." in *Pacific Symposium on Biocomputing*, vol. 10. World Scientific, 2005, pp. 471–482.

[15] N. Nariai, E. D. Kolaczyk, and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," *PLoS One*, vol. 2, no. 3, p. e337, 2007.

[16] R. Shen, A. M. Chinnaiyan, and D. Ghosh, "Pathway analysis reveals functional convergence of gene expression profiles in breast cancer," *BMC medical genomics*, vol. 1, no. 1, p. 28, 2008.

[17] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th international joint conference on Artificial intelligenc*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 448–453.

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[18] D. Lin, "An information-theoretic definition of similarity." in *Proceedings of the Fifteenth International Conference on Machine Learning*, vol. 98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

[19] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC bioinformatics*, vol. 7, no. 1, p. 302, 2006.

[20] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of 10th International Conference on Research In Computational Linguistics (ROCLING97)*, 1997.

[21] J. Z. Wang, Z. Du, R. Payattakool, S. Y. Philip, and C.-F. Chen, "A new method to measure the semantic similarity of go terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[22] S. Jain and G. D. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC bioinformatics*, vol. 11, no. 1, p. 562, 2010.

[23] X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman, "Assessing semantic similarity measures for the characterization of human regulatory pathways," *Bioinformatics*, vol. 22, no. 8, pp. 967–973, 2006.

[24] T. Xu, L. Du, and Y. Zhou, "Evaluation of go-based functional similarity measures using s. cerevisiae protein interaction and expression profile data," *BMC bioinformatics*, vol. 9, no. 1, p. 472, 2008.

[25] C. Pesquita, "Semantic similarity in the gene ontology," in *The Gene Ontology Handbook*. Springer, 2017, pp. 161–173.

[26] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, "Hippie: Integrating protein interaction networks with experiment based quality scores," *PloS one*, vol. 7, no. 2, p. e31826, 2012.

[27] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "Dip: the database of interacting proteins," *Nucleic acids research*, vol. 28, no. 1, pp. 289–291, 2000.

[28] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[29] C. Pesquita, D. Pessoa, D. Faria, and F. Couto, "Cessm: Collaborative evaluation of semantic similarity measures," *JB2009: Challenges in Bioinformatics*, vol. 157, p. 190, 2009.

[30] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," *Journal of biomedical informatics*, vol. 48, pp. 38–53, 2014.

[31] G. K. Mazandu, E. R. Chimusa, and N. J. Mulder, "Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 886–901, 2016.

[32] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter *et al.*, "Structure-based prediction of protein–protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, p. 556, 2012.

[33] C. Von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel, and P. Bork, "String 7recent developments in the integration and prediction of protein interactions," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D358–D362, 2006.

[34] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. suppl_1, pp. S233–S240, 2002.

[35] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein–protein interaction," *Protein engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[36] R. Jothi, M. G. Kann, and T. M. Przytycka, "Predicting protein–protein interaction by searching evolutionary tree automorphism space," *Bioinformatics*, vol. 21, no. suppl_1, pp. i241–i250, 2005.

[37] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences*, vol. 96, no. 8, pp. 4285–4288, 1999.

[38] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.

[39] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, "Increasing confidence of protein interactomes using network topological metrics," *Bioinformatics*, vol. 22, no. 16, pp. 1998–2004, 2006.

[40] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted ppi networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009.

[41] S. Jaeger, S. Gaudan, U. Leser, and D. Rebholz-Schuhmann, "Integrating protein-protein interactions and text mining for protein function prediction," in *BMC bioinformatics*, vol. 9, no. 8. BioMed Central, 2008, p. S2.

[42] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features," *BMC bioinformatics*, vol. 6, no. 1, p. 100, 2005.

[43] Y. Deng, L. Gao, and B. Wang, "ppipre: predicting protein-protein interactions by combining heterogeneous features," *BMC systems biology*, vol. 7, no. 2, p. S8, 2013.

[44] A. Tversky, "Features of similarity." *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[45] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE transactions on systems, man, and cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.

[46] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[47] P. Lord, R. Stevens, A. Brass, and C. Goble, "Semantic similarity measures as tools for exploring the gene ontology." in *Pacific Symposium on Biocomputing*, 2003, pp. 601–612.

[48] A. Nagar and H. Al-Mubaid, "A new path length measure based on go for gene similarity with evaluation using sgd pathways," in *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*. IEEE, 2008, pp. 590–595.

[49] K. Taha, "Determining the semantic similarities among gene ontology terms," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 512–525, 2013.

[50] S. Bandyopadhyay and K. Mallick, "A new path based hybrid measure for gene ontology similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 11, no. 1, pp. 116–127, 2014.

[51] H. Yu, L. Gao, K. Tu, and Z. Guo, "Broadly predicting specific gene functions with expression similarity and taxonomy similarity," *Gene*, vol. 352, pp. 75–81, 2005.

[52] J. Cheng, M. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, and M. A. Siani-Rose, "A knowledge-based clustering algorithm driven by gene ontology," *Journal of biopharmaceutical statistics*, vol. 14, no. 3, pp. 687–700, 2004.

[53] H. Wu, Z. Su, F. Mao, V. Olman, and Y. Xu, "Prediction of functional modules based on comparative genome analysis and gene ontology application," *Nucleic acids research*, vol. 33, no. 9, pp. 2822–2837, 2005.

[54] A. del Pozo, F. Pazos, and A. Valencia, "Defining functional distances over gene ontology," *BMC bioinformatics*, vol. 9, no. 1, p. 50, 2008.

[55] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in wordnet," in *ECAI*, vol. 16, 2004, p. 1089.

[56] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 343–344.

[57] Z. Zhou, Y. Wang, and J. Gu, "A new model of information content for semantic similarity in wordnet," in *Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on*, vol. 3. IEEE, 2008, pp. 85–89.

[58] M. Mistry and P. Pavlidis, "Gene ontology term overlap as a measure of gene functional similarity," *BMC bioinformatics*, vol. 9, no. 1, p. 327, 2008.

[59] B. Li, J. Z. Wang, F. A. Feltus, J. Zhou, and F. Luo, "Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins," in *Proceedings of BIOCOMP10*, 2010, pp. 166–172.

[60] D. Sánchez, M. Batet, and D. Isern, "Ontology-based information content computation," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 297–303, 2011.

[61] G. K. Mazandu and N. J. Mulder, "A topology-based metric for measuring term similarity in the gene ontology," *Advances in bioinformatics*, vol. 2012, pp. 975 783–975 783, 2012.

[62] D. Sánchez and M. Batet, "A new model to compute the information content of concepts from taxonomic knowledge," *International*

bioRxiv preprint doi: https://doi.org/10.1101/459107; this version posted November 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Journal on Semantic Web and Information Systems (IJSWIS), vol. 8, no. 2, pp. 34–50, 2012.

[63] X. Song, L. Li, P. K. Srimani, P. S. Yu, and J. Z. Wang, "Measure the semantic similarity of go terms using aggregate information content," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 11, no. 3, pp. 468–476, 2014.

[64] S.-B. Zhang and J.-H. Lai, "Semantic similarity measurement between gene ontology terms based on exclusively inherited shared information," *Gene*, vol. 558, no. 1, pp. 108–117, 2015.

[65] A. Adhikari, S. Singh, A. Dutta, and B. Dutta, "A novel information theoretic approach for finding semantic similarity in wordnet," in *TENCON 2015-2015 IEEE Region 10 Conference*. IEEE, 2015, pp. 1–6.

[66] J. J. Lastra-Díaz and A. García-Serrano, "A new family of information content models with an experimental survey on wordnet," *Knowledge-Based Systems*, vol. 89, pp. 509–526, 2015.

[67] X. Wu, E. Pang, K. Lin, and Z.-M. Pei, "Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge-and ic-based hybrid method," *PloS one*, vol. 8, no. 5, p. e66745, 2013.

[68] L. Liu, X. Dai, C. Du, H. Wang, and J. Lu, "A new hybrid semantic similarity computation method based on gene ontology," in *Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on*. IEEE, 2014, pp. 849–853.

[69] Y. Xu, M. Guo, W. Shi, X. Liu, and C. Wang, "A novel insight into gene ontology semantic similarity," *Genomics*, vol. 101, no. 6, pp. 368–375, 2013.

[70] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, and P. Xuan, "Measuring gene functional similarity based on group-wise comparison of go terms," *Bioinformatics*, vol. 29, no. 11, pp. 1424–1432, 2013.

[71] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio, "Correlation between gene expression and go semantic similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 330–338, 2005.

[72] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between gene ontology terms," *Data & knowledge engineering*, vol. 61, no. 1, pp. 137–152, 2007.

[73] F. Azuaje, H. Wang, and O. Bodenreider, "Ontology-driven similarity approaches to supporting gene functional assessment," in *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*, 2005, pp. 9–10.

[74] W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea, "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, no. 6959, pp. 686–691, 2003.

[75] P. Hu, G. Bader, D. A. Wigle, and A. Emili, "Computational prediction of cancer-gene function," *Nature Reviews Cancer*, vol. 7, no. 1, pp. 23–34, 2007.

[76] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for go based protein semantic similarity: a systematic evaluation," in *BMC bioinformatics*, vol. 9, no. 5. BioMed Central, 2008, p. S4.

[77] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi *et al.*, "The biogrid interaction database: 2011 update," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D698–D704, 2011.

[78] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.

[79] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, "Human protein reference database2009 update," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.

[80] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. Ghanbarian, S. Kerrien, J. Khadake *et al.*, "The intact molecular interaction database in 2010," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D525–D531, 2010.

[81] A. Ceol, A. C. Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni, "Mint, the molecular interaction database: 2009 update," *Nucleic acids research*, p. gkp983, 2009.

[82] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue, "Bindthe biomolecular interaction network database," *Nucleic acids research*, vol. 29, no. 1, pp. 242–245, 2001.

[83] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes *et al.*, "The mips mammalian protein–protein interaction database," *Bioinformatics*, vol. 21, no. 6, pp. 832–834, 2005.

[84] S. Benabderrahmane, M. Smail-Tabbone, O. Poch, A. Napoli, and M.-D. Devignes, "Intelligo: a new vector-based semantic similarity measure including annotation origin," *BMC bioinformatics*, vol. 11, no. 1, p. 588, 2010.

[85] X. Yu, A. Wallqvist, and J. Reifman, "Inferring high-confidence human protein-protein interactions," *BMC bioinformatics*, vol. 13, no. 1, p. 79, 2012.

[86] S. Razick, G. Magklaras, and I. M. Donaldson, "irefindex: a consolidated protein interaction database with provenance," *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

[87] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org

[88] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "Gosemsim: an r package for measuring semantic similarity among go terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.

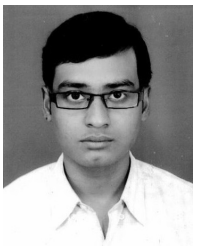
[89] M. Carlson, "Go.db: A set of annotation maps describing the entire. gene ontology. 2013," *R package version*, vol. 3, no. 2, 2013.

[90] M. Carlson, S. Falcon, H. Pages, and N. Li, "org.hs.eg.db: Genome wide annotation for human," 2013.

[91] ——, "org.sc.sgd.db: Genome wide annotation for yeast," *R package version*, vol. 2, no. 1, 2014.

[92] V. Carey and H. Redestig, "Roc: Utilities for roc, with uarray focus. r package version 1.16. 0," 2008.

[93] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: visualizing classifier performance in r," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.



Madhusudan Paul received the M.Tech degree in computer science and engineering from Pondicherry University in 2010. He is working towards the PhD degree in computer science and engineering from Indian Institute of Technology Guwahati, India. At present, he is an assistant professor at Visva-Bharati, Santiniketan, West Bengal, India. His current research interests include complex networks and systems biology.

Ashish Anand is an associate professor at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, India. His current research interests include Machine Learning and its application in computational biology, NLP, Clinical Text Mining, and Deep Learning. Web-site: http://iitg.ac.in/anand.ashish/