PROTEIN FUNCTION PREDICTION USING SEMANTIC SIMILARITY METRICS AND RANDOM WALK ALGORITHM

Ilinka Ivanoska
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, Macedonia

Kire Trivodaliev
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, Macedonia

Slobodan Kalajdziski
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, Macedonia

ABSTRACT

Most protein function prediction methods that have been proposed, are based on sequence or structure protein similarity and do not take into consideration the semantic similarity extracted from protein knowledge databases such as Gene Ontology. In this paper we present an approach for protein function prediction using semantic similarity metrics and the whole network topology of a protein interaction network by using a "semantic driven" random walk with restart. Different semantic similarity metrics are explored and future results should show the relevance of different semantic similarity metrics on protein function prediction using random walk with restart. To achieve the final goal of protein function prediction, the best semantic similarity metric should be used.

I. INTRODUCTION

Today, most of the methods for determining protein function by protein similarity are based on protein sequence or structure. However, the big drawback of these methods for protein similarity is that structure or sequence similarity is not directly related to function, since proteins with significant structure or sequence similarity can have different functions. Furthermore, proteins with different ancestors and no significant sequence similarity can have the same function, due to evolution.

One of the most important challenges of molecular biology is finding a method for extracting protein function and protein similarity knowledge, consisted in the great amount of protein and genome data in well-known protein databases. An important breakthrough in protein annotation is the creation of the Gene Ontology (GO) [1], the most famous bio-ontology that is a structured and controlled vocabulary for describing gene and protein products. The Gene Ontology defines a set of terms to which any given protein may be annotated and is structured as a directed acyclic graph (DAG). Therefore, proteins can be compared according to the Gene Ontology.

This type of comparison is called semantic similarity (SS), and it is based on the structure of the bio-ontology and the relations between its terms, focusing on the semantic similarity between the terms themselves. Generally, semantic similarity is a specific type of similarity that gives meaning similarity between two concepts in an ontology. This type of similarity is the subject of research in artificial intelligence, cognitive sciences, psychology, natural language processing etc.

Most semantic similarity metrics are firstly defined for WordNet (an English vocabulary database) [2], and afterwards used for protein semantic similarity. Protein semantic similarity is independent of homology and can help overcome some of the issues of sequence and structure similarity-based approaches. However, it is still not clear which is the best way to calculate semantic similarity considering the current bio-ontologies, but several metrics have been proposed to calculate protein semantic similarity in the context of the Gene Ontology (GO) [3],[10].

A protein interaction network (PIN) consists of nodes representing proteins, and edges representing interactions between proteins. Such networks are stochastic as edges are weighted with the probability of interaction. There is more information in a PIN compared to sequence or structure alone. A network provides a global view of the context of each gene/protein. Hence, our computational function prediction is characterized by the use of a protein's interaction context within the network to predict its functions. The main idea behind our function prediction technique is that function inference using only local network analysis but without the examination of global patterns is not general enough to cover

all possible annotation trends that emerge in a PIN. Therefore we use Random Walks to extract affinity neighborhoods which take into account the whole network topology. We additionally alter this by incorporating semantic similarity within the random walks and making the whole prediction semantic driven.

The aim of this paper is to present our work in progress for evaluating the metrics and presenting a new system for protein function prediction with the use of semantic similarity and random walk. For a given protein the system can determine similar proteins based on functional similarity, and we should see impact of semantic similarity metrics on determining protein function.

In section 2 we present a systematic analysis and an overview of the existing semantic similarity metrics that will be used, while section 3 will give the proposed system architecture for protein function prediction based on the semantic similarity metrics and the whole network topology using the semantic driven random walk with restart.

II. Overview of Semantic similarity metrics

Several approaches are available to quantify semantic similarity between terms or annotated entities in an ontology represented as a DAG such as GO. There are essentially two types of methods for comparing terms in a graph-structured ontology such as GO: node-based, in which the main data sources are the nodes and their properties; and edge-based, which use the edges and their types as the data source.

A. Node-based metrics

Node-based approaches rely on comparing the properties of the terms involved, which can be related to the terms themselves, their ancestors, or their descendants. One concept commonly used in these approaches is information content (IC), which gives a measure how specific and informative a term is. The IC of a term *c* can be quantified as the negative log likelihood - *log p*(*c*), where *p*(*c*) is the probability of occurrence of *c* in a specific knowledgebase, being normally estimated by its frequency of annotation. Alternatively, the IC can also be calculated from the number of children a term has in the GO structure, although this approach is less commonly used.

The concept of IC can be applied to the common ancestors two terms have, to quantify the information they share and thus measure their semantic similarity. There are two main approaches for doing this: the most informative common ancestor (MICA), in which only the common ancestor with the highest IC is considered [4]; and the disjoint common ancestors (DCA), in which all disjoint common ancestors (the common ancestors that do not subsume any other common ancestor) are considered [9].

Approaches based on IC are less sensitive to the issues of variable semantic distance and variable node density than edge-based metrics [4], because the IC gives a metric of a term's specificity that is independent of its depth in the ontology (the IC of a term is dependent on its children but not on its parents). The use of the IC also makes sense from a probabilistic point of view: it is more probable (and less meaningful) that two gene products share a commonly used term than an uncommonly used term.

Other node-based approaches include looking at the number of shared annotations, that is, the number of proteins annotated with both terms, computing the number of shared ancestors across the GO structure, and using other types of information such as node depth and node link density.

The most common semantic similarity measures used with GO have been Resnik's, Lin's, and Jiang and Conrath's, which are node-based metrics relying on IC [4],[5],[6]. They were originally developed for the WordNet, and then applied to GO. Resnik measures similarity between two terms as simply the IC of their most informative common ancestor (MICA):

$$sim_{Res}(c_1, c_2) = IC(c_{MICA}) \quad (1)$$

While this metric is effective in determining the information shared by two terms, it does not consider how distant the terms are from their common ancestor. To take that distance into account, Lin's and Jiang and Conrath's metrics relate the IC of the MICA to the IC of the terms being compared:

$$sim_{Lin}(c_1, c_2) = \frac{2 * IC(c_{MICA})}{IC(c_1) + IC(c_2)} \quad (2)$$

$$sim_{JC}(c_1, c_2) = \frac{}{1 - IC(c_1) + IC(c_2) - 2 * IC(c_{MICA})} \quad (3)$$

However, being relative metrics, sim_{Lin} and sim_{JC} are displaced from the graph. This means that these metrics are proportional to the IC differences between the terms and their common ancestor, independently of the absolute IC of the ancestor.

To overcome this limitation, in [8] a relevance similarity metric is proposed, which is based on Lin's metric, but uses the probability of annotation of the MICA as a weighting factor to provide graph placement.

$$sim_{Rel}(c_1, c_2) = sim_{Lin}(c_1, c_2) * (1 - p(c_A)) \quad (4)$$

A constraint all of these metrics share is that they look only at a single common ancestor (the MICA) despite the fact that GO terms can have several DCA. To avoid this, the GraSM approach proposed in [9], can be applied to any of the metrics previously described, and where the IC of the MICA is replaced by the average IC of all DCA.

B. *Edge-based metrics*

Edge-based approaches are based mainly on counting the number of edges in the graph path between two terms. The most common technique is the distance, that selects either the shortest path or the average of all paths, when more than one path exists. This technique gives a metric of the distance between two terms, which can be easily converted into a similarity metric. The common path technique calculates the similarity directly by the length of the path from the lowest common ancestor of the two terms to the root node.

While these approaches are intuitive, they are based on two assumptions in bio-ontologies: (1) nodes and edges are uniformly distributed, and (2) edges at the same level in the ontology correspond to the same semantic distance between terms. However, terms at the same depth do not necessarily have the same specificity, and edges at the same level do not necessarily represent the same semantic distance, so the issues caused by the mentioned assumptions are not solved by proposed methods.

Within the edge-based approaches, Pekar and Staab proposed a metric based on the length of the longest path between two terms' lowest common ancestor and the root (maximum common ancestor depth), and on the length of the longest path between each of the terms and that common ancestor [7]. It is given by the expression

$$sim_{PS}(c_1, c_2) = \frac{\delta(c_a, root)}{\delta(c_a, root) + \delta(c_1, c_a) + \delta(c_2, c_a)} \quad (5)$$

where $\delta(c_1,c_2)$ is the length in number of edges of the longest distance between term c_1 and term c_2.

There are other proposed edge-based metrics in [11], [12], based on a maximum common ancestor depth metric, but weighted each edge to reflect depth, introducing a distance to the nearest leaf node and the distance to the lowest common ancestor to take term specificity into account, etc.

C. *Hybrid metric*

Hybrid metric in which each edge is given a weight according to the type of relationship was developed in [13]. For a given term c_1 and its ancestor c_a, the semantic contribution of c_a to c_1, is defined as the product of all edge weights in the "best" path from c_a to c_1, where the "best" path is the one that maximizes the product. Semantic similarity between two terms is then calculated by summing the semantic contributions of all common ancestors to each of the terms and dividing by the total semantic contribution of each term's ancestors to that term.

III. PROTEIN FUNCTION PREDICTION SYSTEM ARCHITECTURE AND SEMANTIC RANDOM WALK WITH RESTART

Our approach divides function prediction into two steps: extraction of neighbourhood profile, and prediction based on the computed neighbourhood (Figure 1) [14]. We summarize the functional network context of a target protein in the neighbourhood extraction step. We compute the steady state distribution of a *Random Walk with Restarts (RWR)* from the protein. The steady state is then transformed into a functional profile. In the second step, we employ a prediction method for the function of a target protein based on its neighbourhood profile.

We summarize a protein's neighborhood by computing the steady state distribution of a *Random Walk with Restarts (RWR)*. We simulate the trajectory of a random walker that starts from the target protein and moves to its neighbors with a probability proportional to the weight of each connecting edge. We keep the random walker close to the original node in order to explore its local neighborhood, by allowing transitions to the original node with a probability of c, the restart probability.

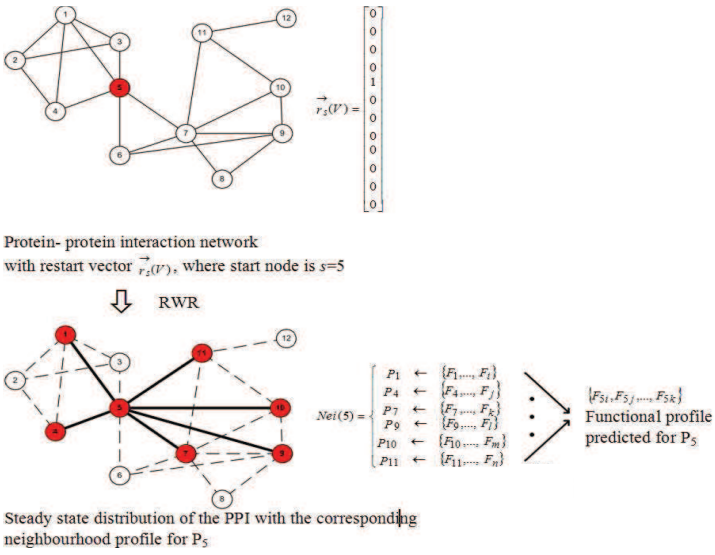

Figure 1: Function prediction process: extraction of neighbourhood profile, and prediction based on the computed neighbourhood

Let $G = (V;E)$ be the graph representing a protein-protein interaction network, where V is the set of nodes (proteins), and E is the set of weighted undirected edges, where the weight shows the probability of interaction (or functional association) between protein pairs. We define the proximity of a node v to a start node s, $p_s(v)$, as the steady state

probability that a random walk starting at node s will end at node v.

Random walk method simulates a random walker that starts on a source node, *s* (or a set of source nodes simultaneously). At every time tick, the walker chooses randomly among the available edges (based on edge weights), or goes back to node s with probability *c*. The restart probability *c* enforces a restriction on how far we want the random walker to get away from the start node *s*. In other words, if *c* is close to 1, the affinity vector reflects the local structure around *s*, and as *c* gets close to 0, a more global view is observed.

The probability $p_s(v)^{(t)}$, describes the probability of finding the random walker at node *v* at time *t*. The steady state probability $p_s(v)$ gives a measure of proximity to node *s*, and can be computed efficiently using iterative matrix operations. Figure 2. shows the iterative algorithm, which provably converges. The number of iterations to converge is closely related to the restart probability *c*. As *c* gets smaller the diameter of the observed neighborhood increases, thus the number of iterations to converge gets larger. The convergence check requires the L_1-norm between consecutive $\vec{p}_s(V)$ s to be less than a small threshold, e.g., 10^{-12}.

A possible interpretation of the neighbourhood profile is an affinity vector of the target node to all other nodes based solely on the network structure.

Input: the interaction network $G = (V;E)$;

a start node *s*;

restart probability *c*;

Output: the proximity vector $\vec{p}_s(V)$;

Let $\vec{r}_s(V)$ be the restart vector with 0 for all its entries except a 1 for the entry denoted by node *s*;

Let **A** be the column normalized adjacency matrix defined by *E* (adjacency semantic similarity matrix);

Initialize $\vec{p}_s(V) := \vec{r}_s(V)$;

while ($\vec{p}_s(V)$ has not converged):

$$\vec{p}_s(V) := (1 - c)A\,\vec{p}_s(V) + c\,\vec{r}_s(V);$$

Figure 2: The iterative algorithm to compute the proximity of all the nodes in the graph to a given start node *s*

We enrich our method and make it semantic driven by including the semantic similarity between the proteins within the step of neighbourhood extraction. Namely, in the algorithm shown on Figure 2. instead of the column normalized adjacency matrix we use the column normalized adjacency semantic similarity matrix S. Each element s_{ij} of S is computed by summing and then normalizing the value of the adjacency matrix element a_{ij} and the semantic similarity score between proteins *i* and *j*. We argue that by using the semantic driven random walk the prediction results will show improvement over the pure topology driven approach.

After extracting the neighborhood profile we set up a strategy for annotating the query protein with the adequate functions according to the functions of the proteins in his neighborhood. The simplest and most intuitive approach would be that each function is ranked by its frequency of appearance as an annotation for the proteins in the neighborhood. This rank is calculated by (6) and is then normalized in the range from 0 to 1.

$$f(j)_{j\in F} = \sum_{i\in K} z_{ij} \tag{6}$$

where F is the set of functions present in the cluster K, and

$$z_{ij} = \begin{cases} 1, \text{if protein i from K has function j from F} \\ 0, \text{otherwise} \end{cases} \tag{7}$$

In our proposed approach the previously explained Resnik's, Lin's, JC's, Rel, GraSM, and Pecar and Staab, and the hybrid metric will be used in the normalized adjacency semantic similarity matrix, and therefore, evaluated with the random walk protein function prediction algorithm. Expected results will show that semantic "driven" random walk with restart improves the general non-semantic approach.

REFERENCES

[1] The Gene Ontology Consortium: http://www.geneontology.org/, Gene Ontology Documentation [Accessed March 2012].

[2] WordNet, Princeton University: http://wordnet.princeton.edu/ [Accessed March 2012].

[3] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, Francisco M Couto, "Metrics for GO based protein semantic similarity: a systematic evaluation", *BMC Bioinformatics* 2008, 9(Suppl 5):S4, 2008.

[4] P. Resnik, "Using information content to evaluate semantic similarity", *Proceedings of the IJCAI05*, pg. 448 – 453, 1995.

[5] D. Lin., "An information-theoretic definition of similarity", *Proceedings of the 15th Int. Conf. on Machine Learning*, 1998.

[6] J. Jiang, D.W Conrath, "Semantic Similarity based on corpus and lexical taxonomy", *Proc. Of 10th Int. Conf. COLING*, 1997.

[7] V. Pekar, S. Staab, "Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision", *Proceedings of the 19th international conference on Computational linguistics*, Morristown, NJ, USA: Association for Computational Linguistics, pp. 1–7, 2002.

[8] Schlicker A., Domingues F., Rahnenfuhrer J., Lengauer T., "A new measure for functional similarity of gene products based on Gene Ontology", *BMC Bioinformatics* 2006, 7:302, 2006.

[9] Couto F., Silva M., Coutinho P., "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors", *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management* New York, NY, USA: ACM Press; 2005:343-344, 2005.

[10] Catia Pesquita, "Improving Semantic Similarity for Proteins based on the Gene Ontology", Master Thesis, University of Lisbon, Portugal, 2007.

[11] H. Wu, Z. Su, F. Mao, V. Olman, Y. Xu, "Prediction of functional modules based on comparative genome analysis and gene ontology application", Nucleic Acids Res. 33: 2822–2837, 2005.

[12] J. Cheng, M. Cline, J. Martin, D. Finkelstein, T. Awad, "A knowledge-based clustering algorithm driven by gene ontology", Journal of Biopharmaceutical Statistics 14: 687–700, 2004.

[13] Wang J.Z., Du Z., Payattakool R., Yu P.S., Chen C.F., "A new method to measure the semantic similarity of GO term", *Bioinformatics* 2007, 23(10):1274-1281, 2007.

[14] K. Trivodaliev, I. Chingovska, S. Kalajdziski, D.Davcev, "Protein Function Prediction Based on Neighborhood Profiles", Proceedings of the ICT Innovations, Springer-Verlaang, Macedonia, 2009.