Prediction of Protein Function Using Graph Container and Message Passing

Hongbo Zhou  
Department of Computer Science  
Southern Illinois University, IL,62901  
Email: hongboz@siu.edu

Qiang Cheng  
Department of Computer Science  
Southern Illinois University, IL,62901  
Email: qcheng@cs.siu.edu

Mehdi Zargham  
Department of Computer Science  
Southern Illinois University, IL,62901  
Email: mehdi@cs.siu.edu

Abstract—We introduce a novel parameter called container flux, which is used to measure the information sharing capacity between two distinct nodes in a graph. Other useful information, both from gene expression or protein interaction test, can be readily incorporated into our container flux. Also, contain flux is a proper candidate for describing the multi-path communication effect caused by large amount of short pathways in real protein networks. In our test, we find there are usually 3–15 pathways consisted of less than 4 edges, between two different proteins in the network built from Baker’s Yeast. We also formulate a new equation for protein function prediction by integrating the container flux as an information sharing component. Based on the scale-free characteristic of protein interaction network, we propose that these proteins of high degree would most likely to be the exemplars for difference clusters. By further exploration, we reveal an interesting consistency between the global optimization of our prediction equation and the exemplar guided clustering problems. We adopt a log-space version of sum-product algorithm, a well-established algorithm called affinity propagation, to approximately solve our optimization problem. At the end, we assign every member of the same cluster with the exemplar’s functions. The exemplar-representing assumption is strongly supported by our preliminary experimental results.

I. INTRODUCTION

A central challenge of the post-genomic era is to determine gene and protein functions. This problem is fundamental to understanding molecular and biological processes that sustain health or cause disease, to develop new drug targets and reliable diagnostics. Proteins are responsible for many cellular activities. However, biological functions are still unknown for a large proportion of sequenced proteins. For example, about one-fourth of the proteins in the best-studied model organism Saccharomyces Cerevisiae remain uncharacterized.

Early effort of using computational methods to annotate protein functions collects a set of characterizing features for each protein and infers the function annotation rules using machine-learning algorithms, e.g., by using gene homology. These methods are useful at a small scale for predicting functions for each individual protein. Recently, novel high-throughput technologies have created large-scale protein-protein interaction (PPI) measurements across human and most model species. A number of computational methods have been designed to predict protein functions based on these PPI data, which are commonly represented as networks, with nodes representing proteins and edges the detected interactions. Despite the growing effort on functional annotation of proteins via the genomic-scale network of interactions, there are still a number of open problems. In this paper, we consider a new approach to functional annotations based on the PPI measurements. Our contributions include the following:

• We propose a new concept called information flux of containers on graphs. It can automatically calculate the information content shared by the nodes of the network. It considers not only the local neighborhoods but also the topology of the network.

• We formulate a new payoff function based on the concept of information flux. Optimizing such a payoff function can automatically find the cellular modules in the PPI network, which propagate the labels from the nodes with known labels to the nodes with unknown labels. The resultant modules have maximal consistency with respect to the given labels.

• We identify an interesting consistency between our formulated payoff function and the objective function used in clustering problems. Based on this connection, we adopt an affinity message passing algorithm to efficiently optimize the payoff function.

This paper is organized in the following way. Section II presents our formulation of the functional annotation problem. Related work is reviewed in Section III. Section IV introduces the new concept of information flux based on containers of a graph. Section V solves the formulated optimization problem by revealing a connection to data clustering problems. Experimental results are given in Section VI and conclusions are drawn in Section VII.

II. PROBLEM FORMULATION

We formulate the problem of functional annotation into a problem of propagating labels from the classified nodes to the unclassified nodes. The quality of the information flow can be measured in terms of consistency of the labels within a neighborhood and sharing of information content between two nodes in a global topology. This quality measure will overcome the drawbacks of the existing computational protein annotation methods in the literature, which either use only local neighborhood labels or cluster the whole network into subnetworks taking into account merely the global topology. Specifically, we formulate the problem based on both local neighborhoods and the concept of the information fluxes of
containers of a graph, which will be elaborated in Section IV. The local neighborhood provides a measure of consistency according to the principle of “guilt by association.” That is, neighboring proteins tend to share similar functions. This principle has been widely used in direct schemes as reviewed in Section III. The global topology will be considered using the summation of information fluxes over all containers of the network. And the consistency of both local measures and the global measures will be combined together additively. We present the payoff function of Equation (1) in Section V after introducing the concept of information flux.

III. RELATED WORK

The computational approaches using large-scale networks of interactions can be generally grouped into three types: direct annotation schemes and module-assisted annotation schemes. Direct schemes annotate functions for each unclassified protein based on its local connections in the network, and module-assisted schemes first identify modules of related proteins and then annotate each module with the same functions for all its members.

Direct schemes can be generally categorized into several groups: neighborhood based methods, graph-theoretic methods, and probabilistic methods. Majority algorithm (MA) method is to predict for a given protein up to three functions that are most common among its neighbors by Schwikowski et al. (2001). The MA method is simple and effective (Chua et al., 2006); however, it treats all associations with equal weight and it considers only the local neighborhood while taking no consideration of the full topology of the network. An improved version of the MA method is to compute a statistical score for each function f within the n-neighborhood of a protein by normalizing the squared difference between the number of proteins with function f and the expectation of this number based on the frequency among the network’s proteins (Hishigaki et al., 2001). Although this improved method assigns different weights to different functions, proteins at different distances are still treated in the same way. To overcome this problem, Chua et al. (2006, 2007) devise a functional similarity score assigning weights to neighboring proteins according to their distances from the target protein. The one- and two-neighborhoods are used in assigning the weights. These methods mainly focus on the local neighborhoods, and thus do not take full consideration of the global topology of the network.

Graph-theoretic methods consist of cut-based methods and flow-based methods. The protein functional annotation problem can be formulated as a minimum multiway cut problem by assigning a function to each unclassified protein to maximize the number of edges that connect proteins assigned with the same function. Vazquez et al. (2003) designed a payoff function consisting of two additive terms: the total number of edges incident on two unlabeled proteins and with the same functions; and the number of edges connecting unlabeled proteins with labeled proteins. The formulation generalizes the minimum multiway cut problem and it is heuristically solved using simulated annealing. With a similar approach, Karaoz et al. (2004) handles one function at a time by partitioning the nodes into two sets and assigning one set with a state of +1 to denote that the current node has the function in question, and the other with -1 otherwise. Nabieva et al. (2005) also formulate the annotation problem as a multiway cut problem and they propose to solve this problem using an integer programming technique. In the same paper, Nabieva et al. (2005) also propose a flow-based approach. They treat each classified protein as a source of functional flow and spread the flow into the network. They also handle one function at a time. Integer programming techniques for optimization can work well only for small-scale problems and can hardly scale to large-scale problems. Probabilistic methods rely on a Markovian assumption that the function of a protein is independent of all other proteins given the functions of its immediate neighbors. Deng et al. (2003) use a Markov random field (MRF) model with the probability of the network conditional on the functional labeling proportional to an exponential function of linearly combined numbers of \((l, l')\) -pairs in the network, where \(l, l' = 0\) or \(1\). The labels of the unclassified proteins are inferred using Gibbs sampling techniques. It has been noted that the method by Karaoz et al. (2004) is a special case of the MRF model. Letovsky and Kasif (2003) also propose an essentially similar MRF model. These methods heavily depend on the Markovian assumption with only immediate neighbors considered. This can be a problem because other than the immediate neighbors, level-2 neighbors also play key roles in providing information on the labels of the proteins (Chua et al., 2006).

Module-assisted schemes have also been proposed. Graph clustering methods are based on binary interactions. Enright et al. (2002) have proposed a Markov clustering (MCL) algorithm, which simulates flow on the PPI graph and computing its successive powers to increase the contrast between regions with high flow and with low flow. Krogan et al. (2006) have used the MCL for complex detection in PPI data, and showed that the high-flow regions correspond to protein complexes which are separated by regions of no flow. Spirin and Mirny (2003) proposed two algorithms: one is based on superparamagnetic clustering (SPC) (Blatt et al., 1996) and the other is based on a Monte Carlo algorithm. Both methods require the user to provide the size of the sought clusters. King et al. (2004) have proposed the restricted neighborhood search clustering (RNSC) algorithm. RNSC defines a cost function to evaluate the partitioning. Starting from a random clustering, RNSC re-partition the graph to minimize the cost. Only local neighborhood information is used by RNSC. Dunn et al. (2005) have designed an edge-betweenness clustering for PPI graph, which is based on Girvan-Newman edge-betweenness algorithm (Girvan and Newman, 2002) for detecting communities in complex systems. A “community structure” in Girvan and Newman algorithm is a subset of nodes within which the connections are dense while the edges among communities are less dense.

It is noticed that direct schemes mainly use the local
information with little consideration of global topology; while module-assisted schemes mainly use the global network topology. It is important to combine both local and global information.

IV. GRAPH CONTAINER AND INFORMATION FLUX IN PROTEIN INTERACTION NETWORK

The prediction is a process of propagating the known functions (labels) to the unclassified proteins. Therefore, for two arbitrary nodes in a network, a key issue is in what capacity they can share their information and how to measure such kind of capacity. For a given protein-protein interaction (PPI) network, this capacity would play a fundamental role in propagating the protein’s functions. For example, two directly connected neighbors would have a high probability of sharing one or more functions. As we will see, most of the former studies Deng, et al.(2003) vary in the description of typical structures for sharing information among nodes. Local methods such as the MA measure such sharing capacity at nearest-neighbor (NN) layer. Chua, et al.(2005) states the advantages of exploiting the availability of the level-two neighbors. Markov random field methods (MRF) adopt every locally-connected component as an information sharing structure. Further studies of Chua, et al.(2007), point out that the function similarity (FS) weighted average method yields a better performance in most cases. The graph container supplies an uniform way of interpreting all of these variations.

The concept of containers in graph is widely studied (Hsu, 2004), which provides an effective method to study the information delay and vulnerability of various networks. The basic idea of a graph container is to use a general and abstract structure to substitute actually complex connected edges. Much mathematical work can be done upon this abstraction, e.g., to investigate the distribution of edge capacity. Here is a formal definition of graph container.

First let us consider an unweighted and undirected graph $G(V, E)$, where $V$ stands for the set of vertices and $E$ is the set of edges. A pathway $p$ is a sequence of distinct vertices, in which two vertices must be neighbors if they are next to each other in the pathway. For two distinct nodes $x, y \in V$, a container $C(x, y)$ is defined as a set of pathways between $x$ and $y$. Simply, $C(x, y)$ stands for all possible information pathways for two nodes $x$ and $y$. We can also define the width of a container as $W_C(x, y) = |C(x, y)|$, where $|S|$ stands for the cardinality of set $S$. For each pathway $p \in C(x, y)$, we denote the length $L(p)$ as the number of distinct internal nodes consisted of $p$ (minus 1).

By comparing to the physical transflux process of water flow in a piping system, we define a new quantity, information flux, for each container $C(x, y)$ as

$$IF_{C(x, y)} = \sum_{p=1}^{W_C(x, y)} \frac{1}{L(p)}.$$

Some illustrations of the container and information flux are given in Fig.2.

For a weighted and undirected graph, some small modifications can yield similar definitions of weighted container and weighted information flux. Let $G(V, E, w(E))$ be a weighted and undirected graph, where the $w(E)$ is a vector of weights which has the same size as $E$. Let the weight of a pathway $w(p)$ as the sum of weights of distinct edges consisted of $p$. Accordingly, the weighted information flux is

$$WIF_{C(x, y)} = \sum_{p=1}^{W_C(x, y)} \frac{w(p)}{L(p)}.$$

Fig. 1. A part of the Baker’s Yeast PPI network. Note that some nodes have extremely large degrees while most nodes only admit a small number of degrees. This scale-free effect indicates the existence of some exemplar nodes which form a skeleton of the network.

Fig. 2. Illustration of the graph containers in both unweighted and weighted graph. it also gives out the calculation of $IF$ and $WIF$. By comparing these values of $WIF$, we can get a better understanding of the multi-path communication effect caused by large amount of short pathways in protein network.
For an unweighted (or, weighted) and undirected graph, we note that two nodes would share much information if their container has a larger value of (weighted) information flux. For an unweighted graph, some level-3 neighbors would have a larger flux than the corresponding level-2 neighbors as illustrated in Fig 2. We point out that this would be always the case among the large-degree nodes in scale-free networks. In other words, these large-degree nodes would make up of the skeleton of the network by sharing much more information (see Fig 1.). More dramatic scenario would happen in weighted graphs, as the neighborhood structure is no longer a primary consideration in terms of the information flux.

It is a computationally prohibiting task to determine all pathways between two distinct nodes in a large-scale network, which turns out to be an NP-Complete problem (Cook, 1971). Note that what we actually need are only short pathways, because what we compare is the sum of the reciprocal values of the pathways’ lengths. We implement a quick search algorithm which can determine all short-paths (usually, with edges less than 4) between two arbitrary nodes, and it can retrieve all paths of lengths less than 4 in about 3 seconds (CPU 2.2G, Memory 2 G, Windows Vista) in a network of 4,770 proteins and 61,702 links.

V. SOLVE THE FUNCTION PREDICTION PROBLEM WITH INFORMATION FLUX

The original optimization equation for protein function prediction can be found in Freschi (2007). We formulate this problem by integrating information flux as a main information sharing component. Specifically, we need to maximize the following pay-off function

$$E = \frac{1}{|\Phi_k|} \sum_{\Phi_k} \sum_i h_i(\Phi_k) + \sum_{i,j} WIF_C(i,j) \delta(\Phi_k(i), \Phi_k(j))$$

where $\Phi_k$ is the $k$-th possible function, $\Phi_k(i)$ is the $k$-th function assigned to the protein $i$, $|\Phi_k|$ is the cardinality of set $\Phi_k$, $h_i(\Phi_k)$ is the number of neighboring proteins of the protein $i$ that are annotated with the function $\Phi_k$, $\delta$ is the discrete delta function. Here we point out that the graph container in the second term represents the local structure. This formulation is different from the formulations in (Przulj et al., 2004; Deng et al., 2003). In Freschi et al. (2006), the network parameter is non-zero only when $i, j$ are direct neighbor. In further work of Chua et al. (2007), they consider more than one layer of neighbors. Again, the MRF method, which utilizes the Markovian property of the original PPI networks, is a hybrid method regarding the graph’s structure parameter. Actually, MRF treats each fully-connected component as a single-step node, and thus can consider more than barely one or two layers of nodes.

Our aim is to maximize Equation (1). As reviewed by Sharan R., et al. (2007), several methods have been used in the literature including simulated annealing, deterministic approximation algorithm, integer programming, page ranking, etc. Here we would adopt a totally different optimization method of message passing. Message passing is essentially a distributed recursive procedure which is proved as a "Bethe free energy approximation" for marginal probability computation in various constraint satisfaction problems (CSP). Some introductory materials for CSP problems and message passing can be found in Murphy et al. (1999). The variation of belief propagation includes sum-product, max-product etc. To make use of max-product method, a first thought is to reshape equation (1) to a new form of product. This can be re-written as

$$e^E = e^{\left(\frac{1}{|\Phi_k|} \sum_{\Phi_k} h_i(\Phi_k) + \sum_{i,j} WIF_C(i,j) \delta(\Phi_k(i), \Phi_k(j))\right)}$$

$$= \Pi_{i} e^{\sum_{\Phi_k} |\Phi_k| h(i) \delta(\Phi_k(i), \Phi_k(j))}.$$  

This is a standard form for max-product algorithm. However, we would like to deal with Equation (1) directly. Fortunately, the max-sum algorithm can help out with a couple of simpler recursive rules by updating messages in the log-space of the original max-product messages (Frey et al., 2007).

Before we get into the details of the updating rules, we would like to elaborate more about the connection between maximizing (1) and the clustering problems. At first sight, there seems no correlation between these two problems. A classic data clustering problem can be formulated as maximizing the following equation:

$$E = \sum_i Sim(i, E(i)),$$

where $Sim(i, j)$ is the similarity between the data points $i$ and $j$, and $E(i)$ is the exemplar point of the cluster where the point $i$ belongs to. As pointed out by Frey et al. (2007), the similarity function can be $l_2$ distance or other measures of similarity. More materials on exemplar point and clustering can be found in Frey et al. (2007). By using the following similarity measure,

$$Sim(i, j = E(i)) = \frac{1}{|\Phi_k|} \sum_{\Phi_k} h_{i,j}(\Phi_k)$$

$$+ \frac{1}{|\Phi_k|} \sum_{\Phi_k} WIF_C(i,j) \delta(\Phi_k(i), \Phi_k(j)),$$

we can recover our original formulation of equation (1), where $h_{i,j} = min(h_i(\Phi_k(i)) \cap \Phi_k(j)), h_j(\Phi_k(i) \cap \Phi_k(j))$ and $\cap$ is the set intersect operation. In this paper, we assume all members in one cluster will be assigned to the same functions as its exemplar node! this will be referred as Exemplar-representing assumption), and thus $\delta(\Phi_k(i), \Phi_k(j)) = 1$ because $j = E(i)$.

Now Equation (6) includes only local operations,

$$Sim(i, j = E(i)) = \frac{1}{|\Phi_k|} \sum_{\Phi_k} h_{i,j}(\Phi_k) + WIF_C(i,j).$$

For a given PPI network and corresponding protein annotations, we can pre-calculate $Sim(i, j)$ for each pair of proteins $i, j$. By using similar updating rules of Frey et al. (2007) for maximizing Equation (4), we can get approximately optimal results for clustering. The subsequent work is to assign the exemplar’s functions to all members of the same cluster. In our
cross-validation test, we set all exemplars as known annotated proteins. This is used to compare and verify our Exemplar-representing assumption. In the general case of unannotated exemplars, we suggest some local majority voting method to determine the functions of the exemplar, then annotate the other members of the same cluster.

VI. Experimental Design and Results

A. Physical Interaction Network and Functional Annotations

We use the baker’s yeast, which is commonly used in the literature. The PPI network is built from the dataset of *Saccharomyces cerevisiae* in BioGrid (Stark C., 2006). This version (2.0.38) includes 5064 proteins and 77806 physical interaction links. The corresponding annotation files are from Munish Information Center for Protein Sequences (MIPS) (Ruepp et al., 2004), which supplies 4770 known proteins. We also use the MIPS functional hierarchy, and consider the 72 MIPS biological processes that comprise the second level of hierarchy. After removing the unannotated proteins from the interaction network, there are still about 61702 links between these 4770 proteins. There are comprehensive studies, Karaoz et al. (2004) and Chua et al. (2007), regarding the weight of the PPI. In this paper we incorporate them directly by adopting Chua et al. (2007).

B. Cross-Validation Evaluation Based on Random Sampling

We make use of the cross-validation evaluation method (Chua et al., 2007). We adopt a modified version of random-seed sampling to avoid the bias caused by manually partition. Each time, we uniformly select $P$ proteins from the set of total $N$ proteins. To avoid isolated proteins, we impose a restriction that each seed should have one or more neighbors in the interaction network. Then, we use affinity message passing method to get the partition of different classes, as well as the exemplar for each class. Finally, every member will be assigned with the same functions as the exemplar of the same cluster.

The performance measurement is always a disputed issue. Generally, we cannot assert which one is overwhelming of another. As adopted by Vazque et al. (2003) and Nabieva et al. (2005), we will measure the success as the fraction of times the top prediction for each protein is correct. This measure corresponds to computing TPs and FPs for the rightmost proteins in every experiment. We evaluate our method by a TPs and FPs curve. The successful prediction ratio is approximately 0.45 for unweighted network, and 0.55 for weighted network. The results are plotted in Fig. 3 and Fig. 4.

VII. Conclusion

We introduce a novel parameter of container flux to measure the capacity sharing ability between two distinct nodes in network, as well as a quick search method to find all short-paths in protein-protein interaction graph. The availability of graph container is justified by the existence of large number of short pathways in real protein-protein interaction networks. Additionally, as stated by (Chua et al., 2007), other information, both from gene expression or interaction test, can be interpreted as certain kind of weights, and thus readily be incorporated into our container flux. This is another advantage of using container flux. As well, container flux can be used to discuss many other problems, such as information delay and attack vulnerability of certain networks.

Another contribution of our work is the exemplar-representing assumption, which is derived from the scale-free characteristic of the protein interaction network. This assumption is assured by our analysis of the consistency between the prediction optimization equation and the exemplar guided clustering problems, and further supported by our experimental results. Further theory analysis and experimental verification will be added regarding the inner cluster structures, as well as its relations with the neighborhood distribution of annotations. The supplementary materials are available at: http://www.cs.siu.edu/~hbzhou/ProteinPrediction.

REFERENCES

