De-noise Biological Network from Heterogeneous Sources via Link Propagation

Nan Du*, Jing Gao†, Vishrawas Gopalakrishnan‡, Aidong Zhang§
Computer Science and Engineering Department
State University of New York at Buffalo,
Buffalo, U.S.A,
$\{nandu^*, jing^\dagger, vishrawa^\ddagger, azhang^\S\}$*@buffalo.edu*

***Abstract*—Lots of recent bioinformatics works have focused on the inference of various types of biological networks, such as gene coexpression networks, protein-protein interaction networks, signal transduction networks, etc. Unfortunately, these raw biological network data often contain much noise, especially the false positive predictions which in many cases hinder accurate reconstruction of biological networks. In addition, since the labeled data is scarce and expensive, we hope that the knowledge from other domains can help handle this lack of labeled data problem. In order to construct a more robust and reliable biological network, we propose a novel link propagation based algorithm to de-noise false positives from the target biological network through propagating information from few labeled samples and a set of auxiliary domain networks. While comparing with many current state-of-the-art algorithms, our proposed approach has shown good performance in de-noising biological network.**

***Keywords*-biological network de-noising; link propagation; heterogeneous sources.**

I. Introduction

Through the analysis of biological network, such as protein interactions network, gene coexpression network and metabolic network, we can better understand how molecules interact with each other at different systems-level to carry out certain biological functions as individuals or modules. However, these raw biological network data contain much noise which may come from sample contaminations, experimental design or measurement errors and thus in many cases hindering accurate extraction of domain knowledge.

Also, it is hard to predict the real interactions between molecules only through the target domain information itself. A promising solution is to predict it by integrating various types of available proteomics and genomics data sources. Link propagation to infer biological networks of multiple species based on genome-wide data and evolutionary information is proposed by [1]. This method propagates labels among the neighbors in the network based on the assumption that “if two pairs of nodes are similar to each other, then these two pairs have similar link strengths”. The successful application has proved that an integration of heterogeneous types of biological data can increase predictive coverage and reduce false positive predictions.

Although [1] did a similar work, their framework’s goal is different from ours. While they reconstruct several biological networks simultaneously by exchanging and propagating information with each other, we propagate the label information from auxiliary networks to de-noise the target network. In addition, instead of propagating the information between each pairwise network, our proposed method directly maps the information from other auxiliary networks to the target network, and then the information is propagated locally.

Summarizing, our work makes two main contributions: (i) Provide a novel link propagation algorithm for combining auxiliary networks’ information with target network’s information aiming at removing false positive links from the target network. (ii) Instead of using only the biological information as [1], we propose a novel method to measure the similarity between the links (interactions between genes), that uses both the biological information from the molecule itself and the topological information from the link position.

The rest of the paper is organized as follows: in the next section, we describe the data set and present the setting of the problem. The proposed approach is presented in Section III. Extensive experimental results are shown in Section IV. Finally, we conclude our work in Section V.

II. Data Set and Problem Setting

In this section, we describe the data set used in this paper and present the problem setting.

A. Data Set

The data set used in this paper is from Kashima et al [1]. This data set consists of three different species - *C.elegans*, *H.plylori* and *S.cerevisiae*. Each species has a metabolic network, where each node is an enzyme and there is a link between two nodes if this pair of enzymes is associated to the nodes that catalyze successive reactions in a known metabolic pathway [2]. These proven links’ information would be used as the ground truth to train and evaluate our approach. Table I summarizes the metabolic networks’ information. Moreover, the data set has the cross-species similarity matrices which are computed by the normalized Smith-Waterman scores of pairwise protein sequence and the intra-species similarity matrix is computed from the gene expression data [1], [2].

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

Table I: Summary of the Metabolic Networks Data

	S.cerevisiae	H.pylori	C.elegans
# nodes	722	291	532
# edges	2323	492	2892

B. Problem Setting

We are considering the problem of de-noising the target biological network by mapping the labeled information from other auxiliary networks to the target network and then propagating the labeled information locally. Our framework aims at removing the false positive edges from the target network. Hence, instead of predicting new links between molecules, we are aiming at removing unreliable existing links from the target network. Therefore our task is summarized as follows:

Input:

- An adjacency matrix of target network $G = (V, E)$, where V is a set of molecules and E is a set of interactions between these molecules.
- The $|V| \times |V|$ node similarity matrix S which measures the pairwise molecular similarity in the network G.
- M auxiliary networks, where each auxiliary network's adjacency matrix is defined as $G^i = (V^i, E^i)$ $(1 \le i \le M)$. Note that in the auxiliary network G^i $(1 \le i \le M)$, only labeled links are represented.
- The molecules similarity matrix $S^i : |V^i| \times |V|$ measure the nodes similarity between i-th auxiliary network and the target network.
- The ground truth (initial labeling) vector Y comes from the metabolic network mentioned above is a $|E| \times 1$ vector, and $Y = (Y_L, Y_U)$ consists of labeled links $Y_L = Y_1, ..., Y_L$ and unlabeled links $Y_U = Y_{L+1}, ..., Y_{L+U}$.

Output:

- A $(1 \times |E|)$ link strength vector F refers to the link strength in the target network which represents the likelihood of whether this link should exist or not.

III. Method

In this section, we present our method for solving the problem of de-noising biological network. We begin by overviewing the regularization framework in Section III-A, and the methods of calculating the link similarity matrix of the target network and the regularizer for each link are shown in the Section III-B and Section III-C respectively.

A. Overview of the Regularization Framework

Here we develop a regularization framework for denoising the target network. Our regularization framework is defined as:

$$\Omega(F) = \frac{1}{2} \sum_{ij} W_{ij} \left(\frac{F_i}{D_{ii}} - \frac{F_j}{D_{jj}}\right)^2 + \frac{\mu}{2} \sum_{i} (F_i - R_i * Y_i)^2, \tag{1}$$

where $\mu > 0$ is the constant that balances the first and the second term. Having formulated the problem as above, we want to minimize this objective function as:

$$F^* = \arg \min_{F^* \in F} \Omega(F). \tag{2}$$

In addition, F_i $(1 \le i \le |E|)$ is the link strength of the i-th link in the target network, and W_{ij} is the similarity between the i-th and the j-th link in the target network which is defined in the Section III-B. The first term $\frac{1}{2}\sum_{ij} W_{ij}(\frac{F_i}{D_{ii}} - \frac{F_j}{D_{jj}})^2$ indicates that the two link strength values F_i and F_j should be close to each other if the similarity W_{ij} is large. The second term $\frac{\mu}{2}\sum_i (F_i - R_i * Y_i)^2$ is the loss function that fits the predictions to the ground truth, where Y_i is the i-th link's label and R_i is a regularizer of the i-th link which balances the influence of labels from different sources. The method to calculate R_i is shown in the Section III-C. Differentiating this objective function with respect to F, we have:

$$\frac{\partial \Omega}{\partial F} = F^* - LF^* + \mu(F^* - RY), \tag{3}$$

where L is the stochastic matrix defined as $L = D^{-1}W$, D is the diagonal degree matrix of W which has $D_{ii} = \sum_j W_{ij}$ and R is the a diagonal matrix which includes the weight for each labeled link. To be more specific, L represents the transition probability of jumping in one step from one node v_i to the other node v_j and is proportional to the edge weight w_{ij}. It is given by $L_{ij} = \frac{W_{ij}}{D_i}$ [3]. Note that, L is similar to the widely used normalized graph laplacian matrix $S = D^{-1/2} W D^{-1/2}$ that $L = D^{-1/2} S D^{1/2}$ [4].

Since L is graph laplacian matrix, it is positive semi-definite. Therefore, LF^* is convex in function F. Moreover, since $F^* - RY$ is convex in function F and μ is a constant, $\Omega(F)$ is a non-negative-weighted sum of convex functions, which means $\Omega(F)$ is convex. In addition, it could be easily proven that $\Omega(F)$ is strictly convex. Therefore, the objective function can be minimized with alternating optimization [5]. After transforming this function, we have

$$F^* - \frac{1}{1+\mu} LF^* - \frac{\mu}{1+\mu} RY = 0. \tag{4}$$

Assume $\alpha = \frac{1}{1+\mu}$ and $\beta = \frac{\mu}{1+\mu}$ (note that $\alpha + \beta = 1$), then the close solution F^* be computed as follows

$$F^* = \beta(I - \alpha L)^{-1} RY. \tag{5}$$

B. The similarity between links

In order to construct the pairwise link similarity matrix W mentioned in the regularization framework, we first need to define a similarity metric. Currently, there are mainly two ways to measure the similarity between links. The first one is using nodes' information. For example, [6] calculates the pairwise link similarity based on the sequence similarities between proteins. The second one uses the topological information. However, using these methods alone it is hard to find out the true similarity between links. In order to measure the pairwise link similarity in a more reliable and comprehensive way, we use both the biological and topological information.

Our method is based on the assumption that if two links have both similar node pairs information and similar link topological position, these two links should have similar link strength which judges this edge's existence.

1) Node information based similarity: The domain information of molecules include genomic or experimental data about genes or proteins of the target species; for example, gene ontology similarity [7], protein sequence similarity [8] and gene expression similarity [9]. In this article, we define the node based similarity between links $e_1 = (a, b)$ and $e_2 = (c, d)$ as

$$W^n_{(e_1,e_2)} = Max(S_{ac} * S_{bd}, S_{ad} * S_{bc}), \tag{6}$$

where S is the intra-node or inter-node similarity matrix mentioned above. Since the links in our case are undirected, we use the maximum mapping case for a pair of links. If an edge pair has high similarity it means their nodes have close structure, function, or evolution process.

2) Link topology based similarity: Based on the assumption that the links having similar topological position should have similar link strength, we want to design a metric that reflects the link's position similarity effectively. Therefore we use the edge betweenness similarity which utilizes the shortest-path edge betweenness metric introduce by [10]. This method generalizes Freeman's betweenness centrality to links and defines the edge betweenness of a link as the number of shortest paths between pairs of vertices that run along it, and it can be thought as signals traveling through a network. The edge betweenness similarity is defined as:

$$S_{bw}(v, w) = \frac{SP_{vw}}{SP_{max}}, \tag{7}$$

where SP_{vw} is the number of shortest paths passing through edge vw, and SP_{max} is the maximum number of shortest paths passing through a link in the graph. Scores are again normalized to the range [0 1] using min-max normalization. [10] has shown that edge betweenness is effective to reflect the links' position in the network in which nodes are joined together in tightly-knit groups between which there are only looser connections.

Since the edge betweenness from different categories are different, the edge betweenness is an effective metric to measure the edges' topological similarity. Finally, we define the edge topology-based similarity between edges $e_1 = (a, b)$ and $e_2 = (c, d)$ as

$$W^t_{(e_1,e_2)} = K(S_{bw}(a, b), S_{bw}(c, d)), \tag{8}$$

where K is a Gaussian RBF (radial basis function) kernel

$$K(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}), \tag{9}$$

which is used to measure the distance between betweenness value from two edges and S_{bw} is defined at Eq. 7.

3) Similarity metric: Finally, based on the node information and topological information mentioned above, we define the edge similarity between edge $e_1 = (a, b)$ and $e_2 = (c, d)$ as:

$$W_{(e_1,e_2)} = W^n_{(e_1,e_2)} * W^t_{(e_1,e_2)}, \tag{10}$$

where W^n is the node information based similarity defined in Eq. 6 and W^t is the link topology based similarity defined in Eq. 8. Using these information simultaneously, one can predict confidently the similarity between two links - if two links have high similarity then that means these links are similar to each on both gene features and topological position and should have the same label (exist or not).

C. Link Propagation via Auxiliary Networks

1) Labeled information mapping: Although inspired by [1] which uses the link propagation on multiple data sources, there are three main points that make our method different. First, the goal of their method is to reconstruct several biological networks simultaneously which is different with our goal - propagating the label information from auxiliary networks to de-noise the target network. Second, our proposed method does not need any negative examples (non-interacting pairs of molecules). Actually, there are no "gold standard" negative examples, because the pairs tested and found not to interact are almost never reported [11]. Third, instead of calculating the similarity of each pair of edges between target network and each auxiliary network, we directly map the auxiliary networks' labeled information to the target network, thus these inter-information is propagated in the target network as a intra-information, which need less running time.

Instead of propagating the labeled information from other auxiliary networks, we directly map the labeled information from the auxiliary networks to the target network, which makes these external information local. In our method, we need to assign two set of weights which are for each auxiliary network and each link in the target network, respectively. First of all, we would assign each auxiliary network a weight, which represents the prediction power of this auxiliary network. Intuitively, this weight measures the agreement that this auxiliary network agrees with the target network. Once the weights for each auxiliary network has been received, we would maintain a set of weights over each unlabeled link at the target network, which depends on the the support from the auxiliary networks. The Figure 1 shows the principle of mapping the labeled links from auxiliary networks to the target network.

The information that can be used for inferring biological networks is, evolutionary information about the conservation of protein interactions, called *interolog*. The assumption of *interolog* is that, if protein u interacts with protein v in one kind of species, then their orthologous protein u^* and v^* in other species are likely to interact with each other [1], [12].

The pseudo code in Algorithm 1 shows how to measure each auxiliary network's weight. The outer for-loop between line 2 and line 15 assigns each auxiliary network a weight based on the agreement with labeled information in the target

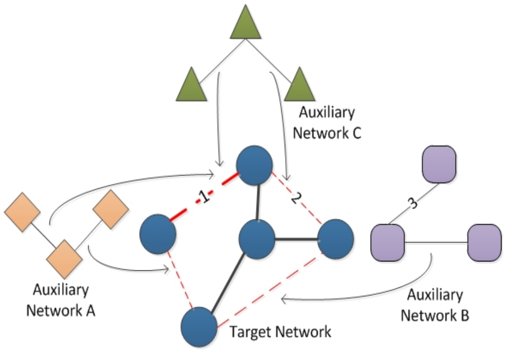

Figure 1: The structure of mapping the external labeled information to the target network

network. The for-loop between line 4 and 14 sequentially goes through each labeled link in the current auxiliary network; if this link could be mapped to a labeled link in the target network, the matching number increases. In line 17, $\sum_{j=1}^{m} count_j$ is a normalization factor. To prevent some mapped links from having higher weight than the original labeled link in the target network, $\sum_{j=1}^{m} |E^j|$ is used to constrain the weight of the mapped links between [0, 1].

Algorithm 1 Assign each auxiliary network a weight
Input: Target adjacent matrix $G = (V, E)$ and auxiliary network $G^1, ..., G^M$, the target links' labeled situation vector Y, and the inter-nodes similarity matrices $S^1, ..., S^M$
Output: The $M \times 1$ vector ASW which records the weight for each auxiliary network

```
 1: begin
 2: for $i = 1$ to $M$ do
 3:   $count_i \leftarrow 0$
 4:   for $j = 1$ to $|G^i|$ do
 5:     for $k = 1$ to $|G^i|$ do
 6:       if $G^i_{jk} == 1$ then
 7:         $u \leftarrow \arg\max_x S^i_{jx}$
 8:         $v \leftarrow \arg\max_x S^i_{kx}$
 9:         if Link $uv$ is labeled in $Y$ then
10:           $++count_i$
11:         end if
12:       end if
13:     end for
14:   end for
15: end for
16: for $i = 1$ to $M$ do
17:   $ASW_i \leftarrow \frac{count_i}{\sum_{j=1}^{M} count_j \times \sum_{j=1}^{M} |E^j|}$
18: end for
19: end
```

The pseudo code in Algorithm 2 shows how to map the labeled information from auxiliary networks to the target network. The outer for-loop between line 2 and line 17 assigns each unlabeled link in the target network a weight based on the number of mappings from the auxiliary networks. The for-loop between 3 and 16 sequentially goes through each labeled link in the current auxiliary network, if one link is mapped to an unlabeled link in the target network, this unlabeled link's weight would increase. As line 9 shown, the final weight of each unlabeled link is a weighted vote from each labeled link from the auxiliary networks, and the way of evaluating ASW_i is shown in the Algorithm 1. Once the weight for this unlabeled link has been received, we would update the labeled situation of it in Y, as line 10. Observe that, to shorten the gap between the links' weight, we assign the new link weight as the square root of the old one in line 19.

Algorithm 2 Mapping the labeled information
Input: Target adjacent matrix $G = (V, E)$ and auxiliary network $G^1, ..., G^M$, the target links' labeled situation vector Y, and the inter-nodes similarity matrices $S^1, ..., S^M$
Output: The $|E| \times 1$ target links' link strength vector Y and a $|E| \times |E|$ diagonal weight matrix R

```
 1: begin
 2: for $i = 1$ to $M$ do
 3:   for $j = 1$ to $|G^i|$ do
 4:     $u \leftarrow \arg\max_x S^i_{jx}$
 5:     for $k = 1$ to $|G^i|$ do
 6:       $v \leftarrow \arg\max_x S^i_{kx}$
 7:       if Edge $uv$ is included in $G$ and not labeled in $Y$ then
 8:         Find $t$ which is the index of edge $uv$ in $Y$
 9:         $R_{tt} \leftarrow R_{t,t} + ASW_i$
10:         Updated Edge $uv$'s relative value in $Y$ as 1
11:       else if Edge $uv$ is labeled in $Y$ then
12:         Find $t$ which is the index of edge $uv$ in $Y$
13:         $R_{tt} \leftarrow 1$
14:       end if
15:     end for
16:   end for
17: end for
18: for $i = 1$ to $|E|$ do
19:   $R_{tt} \leftarrow R_{tt}^{-1/2}$
20: end for
21: end
```

D. Time Complexity

The proposed method has time complexity $O(n^3)$ where n is the number of links in the target network which is equal to E. Since our goal is to de-noise the false positive links from the target network $|E| \ll |V| \times |V|$. So the actual running time of the proposed method crucially depends on the number of links in the raw target network. Actually, it is widely believed and proven in some cases, that biological networks are scale free networks, with a few nodes densely connected to many others and most molecules interacting only with a few others. Thus, the actual running time of our proposed method is completely acceptable. Since the major part of the framework is the auxiliary labeled information mapping scheme and the proposed method is based on the link propagation, we name our approach **I**nformation **M**apping **L**ink **P**ropagation - **IMLP**.

IV. Experiments

In this section, we conduct extensive experiments to show the proposed method is effective in de-noising a target network and the relationship between performance and the auxiliary sources.

A. Baseline Methods

We compare the performance of our proposed approach to 6 relative methods which could be used for biology network de-noising problem. These baseline algorithms are from 4 different categories and exhibit a good diversity to cover the state-of-the-art methods and their details are discussed below.

1) Network Local Topology-based Approaches:

- Common Neighbors: Common neighbors is the most straightforward technique and is defined as the number of common neighbors shared by two nodes.
- Jaccard's Index: Jaccard's Index is a measure to evaluate the similarity between two neighbor sets, which can be viewed as the normalized number of common neighbors.

2) Network Global Topology-based Approaches:

- Katz Index: The Katz index is a weighted number of walks starting from a given vertex. The Katz index is measured as: $Katz_{vw} = \sum_{l=1}^{\infty} \beta^l.|paths_{vw}^l|$, where $paths_{vw}^l$ is the set of all length-l paths from v to w [13]. In our experiment, the parameter l is set as 3 and β is set as 0.3;
- Random-Walk: The random-walk is based on the idea that information propagated from source v will travel through randomly chosen intermediate visiting nodes to the target node w [13].

3) Kernel-based Approach: The diffusion kernel has been demonstrated to be a method for measuring the pairwise similarity between edges in many applications. In our experiment, diffusion kernel is modified as a special case for network de-noise. To be more specific, the diffusion kernel matrix as $G^{dk} = exp(\frac{-\sigma^2}{2\bar{L}})$, where $\bar{L}$ is the normalized laplacian. Note that G_{ij} is the link strength of the edge connecting node i and j, which is used to predict the existence of links [14]. In our experiment, we set the parameter $\sigma = 1$.

4) Link Propagation: The idea and method of Link Propagation has been mentioned in the previous section. In our experiment, the hyperparameters for Link Propagation are set as $\epsilon = 10^{-3}$ and $\sigma = 10^{-5}$ which is the same as the original setting in [1].

B. Experimental Results

In this section, we show the experimental results for our proposed algorithm. The first set of experiments compares the results of the proposed algorithm with the baseline algorithms. We demonstrate that our proposed method performs better than the baseline algorithms in most cases. The second set of experiments describes the effect of varying number of auxiliary sources on our proposed method. We demonstrate that the more auxiliary sources we uses for de-noising the target network, the better predictive performance we have. Throughout all of the experiments, we set $\sigma = 0.4$ (for Eq. 9) and $\mu = 1$ (for Eq. 4) for our algorithm. The performance of the algorithms are evaluated with the Area Under Curve (AUC), which summarizes the ROC curves. The perfect algorithm has an AUC measure of 1 and the random algorithm has an AUC measure of 0.5.

1) Performance Comparison: In this part, we compare the proposed method with 6 other algorithms mentioned above. We randomly selected 10^4 links from one species as the raw target network G (*S.cerevisiae*, *C.elegans* and *H.plylori*, respectively), and the rest two species' metabolic networks are utilized as auxiliary networks. It is noted that all the labeled links are added to this raw target but only part of them would be used as training data and the rest unlabeled links are used to evaluate the method's performance. Note that, throughout the following experiments, each result is the average over five runs. Moreover, we choose six different ratio 15%, 30%, 45% , 60%, 75% or 90% of the all proven links in the target network as a training set. The results of de-noising *S.cerevisiae*, *C.elegans* and *H.plylori* are shown in Figure 2(a), Figure 2(b) and Figure 2(c), respectively. It can be seen that the proposed method is better than the other methods in de-noising each target network except for the extreme case when only 15% labeled examples are available. It is expected that an increase in labeled examples rate should lead to increase in AUC performance, and this is corroborated in our experiment result figures.

There are mainly two reasons for that: First, as a semi-supervised learning method, the proposed algorithm propagates the information from both the labeled and the unlabeled data, while the other topology-based or kernel-based methods only use the information from the labeled data. Second, our new metric measures the similarity between edges and considers both the node information and the link's topological information, while the other baseline algorithms only use either of them.

2) Varying the Number of Sources: In this part, we demonstrate the relationship between the proposed method's performance and the number of auxiliary sources. Similar to the previous experiment, we randomly selected 10^4 links of all the links from one species as the raw target network G (*S.cerevisiae*, *C.elegans* and *H.plylori*, respectively), and then using different number of sources to de-noise it. To be more specific, we compare the situations where we have zero, one and two auxiliary sources. Note that each result is averaged over five runs. Similar to previous case, we choose six different ratio 15%, 30%, 45% , 60%, 75% or 90% of the all proven links in the target network as a training set. The results of de-noising *S.cerevisiae*, *C.elegans* and *H.plylori*

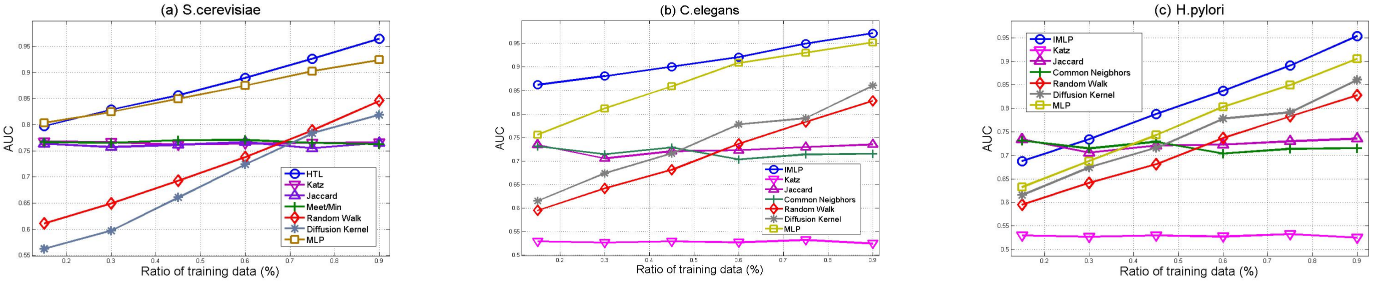

Figure 2: Performance Comparison with baseline methods: (a) *S.cerevisiae*, (b) *C.elegans* and (c) *H.plylori*.

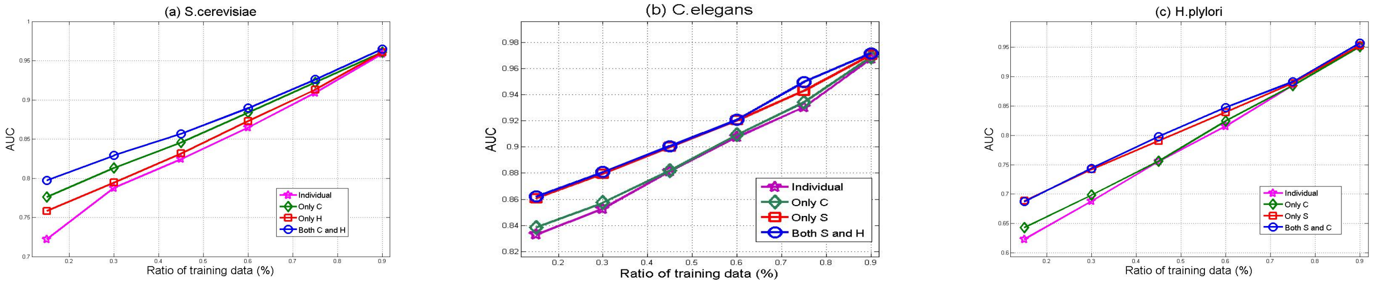

Figure 3: Performance Comparison with Different Number of Sources: (a) *S.cerevisiae*, (b) *C.elegans* and (c) *H.plylori*.

are shown as examples in Figure 3(a), Figure 3(b) and Figure 3(c), respectively. As the figures reveal, it can be observed that AUC is improved with the increasing of the number of the sources. These results show that each auxiliary network aid at removing the false positives at the target network, and more the auxiliary networks we use, better the performance.

V. Conclusion

In this paper, we proposed a novel link propagation framework method by mapping the label information from auxiliary to the target network, then propagate them as local information. Our aim is to remove false positives from our target biological network by transferring network connectivity knowledge from multiple auxiliary networks with the help of weighted mapping. The proposed method achieves significant performance improvement through the correspondence between auxiliary networks and target network. In order to compute the similarity of two links, we proposed a metric which effectively considers both the node-based information (domain knowledge) and topology structures of the links simultaneously. The experimental results show that the performance of removing false positive links in the target network can indeed be boosted with the help of the auxiliary network by transferring useful knowledge to the target network.

References

[1] H. Kashima, Y. Yamanishi, T. Kato, M. Sugiyama, and K. Tsuda, "Simultaneous inference of biological networks of multiple species from genome-wide data and evolutionary information: a semi-supervised approach." *Bioinformatics*, vol. 25, no. 22, pp. 2962–2968, 2009.

[2] M. Kanehisa, M. Araki, S. Goto, and et al., "Kegg for linking genomes to life and the environment," *Nucleic Acids Research*, vol. 36, no. Database issue, pp. D480–D484, 2008.

[3] M. Planck and U. V. Luxburg, "A tutorial on spectral clustering a tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. August, pp. 395–416, 2006.

[4] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch, "Learning with local and global consistency," *Advances in Neural Information Processing Systems 16 Proceedings of the 2003 Conference*, vol. 1, 2004.

[5] T. Hwang and R. Kuang, *A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery*, 2010, pp. 583–594. [Online]. Available: http://www.siam.org/proceedings/datamining/2010/dm10_051_hwangt.pdf

[6] H. Kashima, T. Kato, Y. Yamanishi, M. Sugiyama, and K. Tsuda, "Link propagation: A fast semi-supervised learning algorithm for link prediction," *Science*, pp. 1099–1110, 2009.

[7] Z. Lei and Y. Dai, "Assessing protein similarity with gene ontology and its use in subnuclear localization prediction," *BMC Bioinformatics*, vol. 7, no. 1, p. 491, 2006.

[8] Z. Zhang, A. A. Schaffer, W. Miller, T. L. Madden, D. J. Lipman, E. V. Koonin, and S. F. Altschul, "Protein sequence similarity searches using patterns as seeds," *Nucleic Acids Research*, vol. 26, no. 17, pp. 3986–3990, 1998.

[9] P. Kharchenko, D. Vitkup, and G. M. Church, "Filling gaps in a metabolic network using expression information." *Bioinformatics*, vol. 20 Suppl 1, no. Suppl 1, pp. i178–i185, 2004.

[10] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[11] J. Yu and F. Fotouhi, "Computational approaches for predicting protein-protein interactions: a survey." *Journal of Medical Systems*, vol. 30, no. 1, pp. 39–44, 2006.

[12] L. R. Matthews, P. Vaglio, and et al., "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or interologs," *Genome Research*, vol. 11, pp. 2120–2126, 2001.

[13] A. Zhang, *Protein Interaction Networks: Computational Analysis*, 1st ed., New York, NY, USA, 2009.

[14] A. J. Smola and R. Kondor, "Kernels and regularization on graphs," *Machine Learning*, vol. 2777, pp. 1–15, 2003.