Investigation and Review of Efficient Method for Multiple Protein Network's Pairwise Alignment

Smita Upendra Gumaste, Jyoti Rao

Abstract- Since from last decade, there is rapid growth in the availability of data over the protein-protein interaction (PPI) networks considering the various species like human, fly, bacteria, yeast and worm. As we know that, one of the highly impacted approach for protein networks is that their comparative analysis which has already gain many researchers attention specially for the predicting the network structure, protein function as well as interaction. The major challenge for execution of this approach is to present robust algorithm for multiple network alignment. In this review paper, we are first presenting the literature review over the network alignment problems and querying problems. In the literature we are also discussing different PPI networks and their alignment problems. Further our main aim is to investigate the algorithm which is presented for efficient, fast with more accuracy pairwise alignment of multiple protein networks. Here we considering the proposed approach is work with novel representation of multiple protein networks those are having linear size. From the experiment and results observations, we found that this approach is more efficient and fast as compared to previous studies for multiple protein networks..

Index Terms— Protein-protein interactions, pairwise alignment, yeast two-hybrid, data representation, search methods.

I. INTRODUCTION

Basically as in general, in the problem of network alignment we have to find out network reasons those are conserved in their interaction patterns as well as their sequence across two or more species. Whereas basic problem is difficult and generalizing subgraph isomorphism, there are many methods were presented to overcome this problem. One heuristic approach for the problem creates a merged representation of the networks being compared, called a network alignment graph, facilitating the search for conserved subnet works. In a network alignment graph, the nodes represent sets of proteins, one from each species, and the edges represent conserved PPIs across the investigated species. Protein-protein interactions (PPI) are of central importance for virtually every process in a living cell. Information about these interactions can improve our understanding of diseases and provide the basis for new therapeutic approaches. The main aim of system biology is to find out how the proteins from the cell are interacting with each other [2].

Procedures such as yeast two-hybrid and protein co-immunoprecipitation are routinely employed nowadays to generate large-scale protein interaction networks for human and most model species. Key to interpreting these data is the inference of cellular machineries.

Manuscript received on March, 2013.

Smita Upendra Gumaste, BE computer science and engineering ME pursuing, India.

Jyoti Rao, Assistant professor Computer engineering dept D Y Patil Institute of engineering and technology pimpri pune ME Computer engineering completed and Phd pursuing., India.

As in other biological domains, a comparative approach provides a powerful basis for addressing this challenge, calling for algorithms for protein network alignment [3].

Recently there are many algorithms proposed by various researchers in order to overcome the network alignment problem. But, its extension to more than a few (3) network proved difficult due to the exponential growth of the alignment graph with the number of species. Hence this becomes challenge to researchers again, and they started work over network alignment problem in case of multiple protein networks [4] [5]. To overcome this problem of multiple networks alignment few more methods presented by researchers like an algorithm was suggested to overcome this difficulty, proposing the idea of imitating progressive sequence alignment techniques. Very recently, Dutkowsky and Tiuryn proposed another framework for efficient alignment of multiple networks; however this approach was applied to date to three networks only.

Here in this paper, we are presenting and investigating the recently presented approach for pair wise alignment of multiple protein networks which resulted into most accurate and fast. This approach is basically depends over the novel representation of the network data. In following sections, III we will present the literature survey over network alignment problems and querying. Section IV, we are discussing the proposed algorithm, Section V; we are presenting the work and results over proposed approach.

II. NETWORK ALIGNMENTS AND QUERYING

PPI data present a valuable resource for this task. . Comparative analysis is used to tackle these problems, and improve the accuracy of the predictions. A fundamental problem in molecular biology is the identification of cellular machinery that are, protein pathways and complexes. But there is a considerable challenge to interpret it due to the high noise levels in the data and the fact that no good models are available to pathways and complexes. Main paradigm behind comparison of PPI networks is that evolutionary conservation implies functional significance. Conservation of protein sub networks measured both in terms of protein sequence similarity, and in terms of similarity interaction topology [7].

Some basic notations that appear in many previous works that find conserved pathways or complexes in the PPI networks of different organisms describes in this section.

1 Network Alignment

A PPI network is conveniently modeled by an undirected graph G (V;E), where V denotes the set of proteins, & (u; v) $\in$ E denotes an interaction between proteins u $\in$ V and v $\in$ V

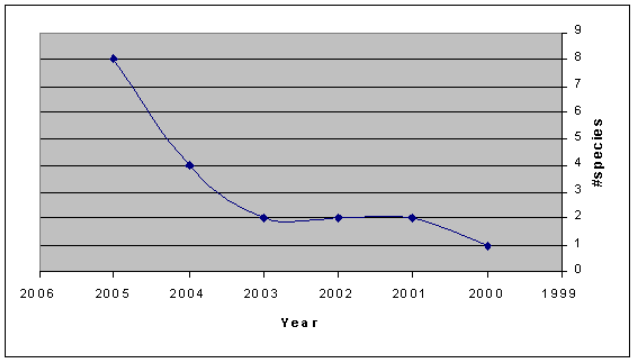

Figure 1: This graph shows the amount of species that their PPI network has been measured.

This graph consists of edges representing conserved interactions between the species and nodes representing sets of k sequence-similar proteins (one per species), the network alignment problem: Given k different PPI networks belonging to different species, to find conserved sub networks within these networks. In order to find these conserved sub networks an alignment graph is built. Illustration of such alignment is shown in Figure 2. This concept was first introduced and used by Ogata et al & Kelley et al.

A heuristic approach is required here since the problem of finding conserved sub networks in a group of networks is NP-Hard, because we can reduce it to sub graph-isomorphism known as NP-Hard. Creating an alignment graph from a set of k original networks is one heuristic that enables us to search in all k PPI networks simultaneously. Other heuristics or approximation methods are applicable as well [6].

2 Network Querying Problem Definition

Given a PPI network G, and a sub network S, we wish to find sub networks in G that are similar to S. Similarity in terms of sequence or topological both are measured. Allow the insertion of proteins into the matched sub network, or deletion of vertices from the query sub network S.

The network querying problem can be reduced to a network alignment problem, shown by Kelley etal, simply by aligning the sub network S with the network G. Also, more general formulations are possible; Network queries can be used to identify conserved functional modules across multiple species, as will be described in the following sections [7].

3 Protein Similarity

In order to build an alignment graph we need to define similarity measure between proteins. First, let us define Homology of proteins (Figure 3 illustrates the speciation and duplication events, and the described below protein relations):

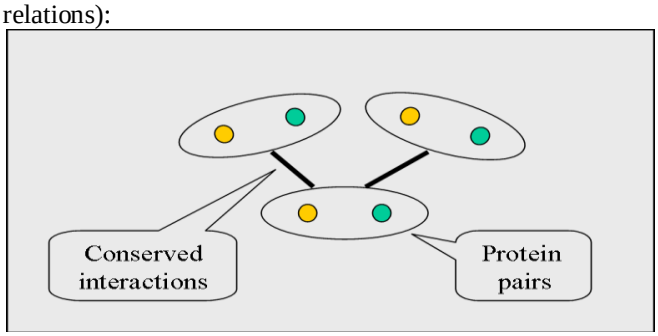

Figure 2: This figure illustrates an alignment graph of two species. Nodes are constructed of pairs of proteins, one per species, which present a high level of sequence-similarity.

Edges represent interactions between proteins in the original networks which are conserved, meaning they exist in a high level of confidence in both original networks.

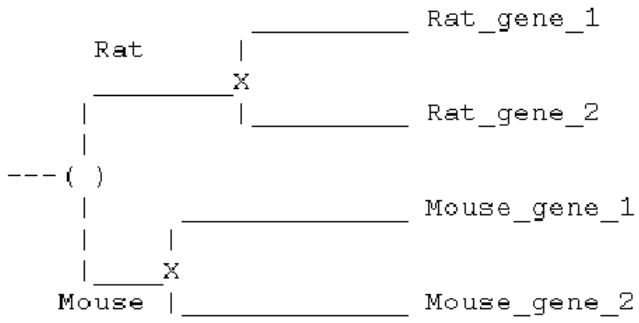

Figure 3: This figure show a gene that diverged after a speciation to a mouse gene and a rat gene.

Within the mouse and the rat species the gene has been duplicated to two different genes rat gene 1 and rat gene 2 in the rat, and mouse gene 1 mouse gene 2 and in the mouse. Each pair of genes is homologous. Each pair of genes that consists of a rat gene and a mouse gene is orthologous, and each pair that consists of genes in the same species is paralogous [8].

Homologous proteins - This is often detected by checking the sequence similarity between these proteins. Two proteins that have common ancestry. The proteins can be either from the same species, or from different species (either orthologous or paralogous).

Orthologous proteins - In a speciation event one species evolves into a different species (anagenesis) or one species diverges to become two or more species (cladogenesis). Two proteins from different species that diverged after a speciation event.

Paralogous proteins - Two proteins from an equivalent species that diverged when a duplication event, within which a part of the ordination is duplicated. We define similar proteins as potentially homologous proteins, i.e. proteins sequences maintain certain degree of similarity.

III. 4 METHODS

1 Data representation

Given k protein-protein interaction networks, we have a tendency to represent them employing a k-layer graph, that we have a tendency to decision a stratified alignment graph. Every layer corresponds to a species and contains the corresponding network. Further edges connect proteins from completely different layers if they're sequence similar. Formally, layer i has a set V_i of vertices and a set E_i of edges. For exposition purposes, assume that $|V_i| = n$ for all i. additionally, we have a set of inter-layer denoted by E_H

Let $G_H = (\cup_i V_i, E_H)$ denote the graph restricted to the inter-layer edges. Let δ be

The largest degree in G_H The relation between associate degree alignment graph and a bedded alignment graph ought to be clear whereas within the former each set of probably orthologous proteins is described by a vertex within the latter such a collection is described by a subgraph of

size k which has a vertex from every of the layers.

We decision such a subgraph a k-spine. Key to the recursive approach given below is that the assumption that a k-spine comparable to a collection of really orthologous proteins should be connected and, hence, admits a spanning tree. Thus, we are able to establish all potential vertex sets inducement k-spines by craving for trees instead [1] [9].

A collection of (connected) k-spines induces a candidate preserved subnet-work. We have a tendency to score it employing a chance magnitude relation score as delineate. The score evaluates the t of the protein-protein interactions inside this subnetwork to a preserved subnetwork model versus the possibility that they arise indiscriminately. The preserved subnetwork model assumes that every combine of proteins from constant species within the subnetwork ought to act, severally of all different pairs, with high chance the random model assumes that every species network was chosen uniformly indiscriminately from the gathering of all graphs with constant vertex degrees because the ones ascertained. This random model induces a probability of occurrence p_{uv} for each edge (u, v) of the graph. To accommodate for information on the reliability of interactions, the interaction status of every vertex pair is treated as a noisy observation, and its reliability is combined into the likelihood score. Overall, for a subnetwork with vertex set U the likelihood ratio score factors over the vertex pairs in it: $\mathcal{L}(U) = \prod_{(u,v)\in U\times U} w(u,v)$ where $w(v,v) = 0$ and for $u \neq v$,

$$w(u,v) = \log \frac{\beta Pr(O_{uv}|T_{uv}) + (1-\beta)Pr(O_{uv}|F_{uv})}{p_{uv}Pr(O_{uv}|T_{uv}) + (1-p_{uv})Pr(O_{uv}|F_{uv})},$$

Here O_{uv} denotes the set of experimental observations on the interaction status of u and v, T_{uv} denotes the event that u and v truly interact, and F_{uv} denotes the event the u and v do not interact. The computation of $Pr(O_{uv}|T_{uv})$ and $Pr(O_{uv}|F_{uv})$ is based on the reliability assigned to the interaction between u and v [1].
This notion of a conserved subnetwork is extended easily to a layered alignment graph. If we considered every k-spine to be a (super-) node in a graph, then an m-node subgraph is a subgraph of m k-spines, with a dense interconnection of PPI edges. Formally, define an m-subnet as a collection U of k multi-sets $U_i = \{u_i[1], \ldots, u_i[m]\}$ with the following properties:

- For all $1 \le i \le k$ and $1 \le j \le m$, $u_i[j] \in V_i$.
- For all $1 \le j \le m$, the set $U[j] = \{u_1[j], u_2[j], \ldots, u_k[j]\}$ is a k-spine.

The score $\mathcal{S}(U)$ of the m-subnet is given by $\mathcal{S}(U) = \sum_{i=1}^{k} \mathcal{L}(U_i)$.

2 The search algorithm

The main algorithmic task is to appear for top rating m-subnets, for a hard and fast m. This downside is computationally laborious even once there's solely one network and edge-weights area unit restricted to +1 for all edges, and 1 for all non-edges. Thus, we tend to resort to a greedy heuristic that starts from high weight seeds and expands them victimization native search. Such greedy heuristics are with success applied to go looking for preserved subnetwork in a very network alignment graph. There are unit 2 sub-tasks we like to tackle: (i) computing high weight seeds; and (ii) extending a seed. We offer algorithmic solutions for each task below.

Computing seeds: We start by computing d-subnets as seeds, where d << m. notably, even when d = 2, we do not know of any algorithm better than the naive approach, which involves looking at all pairs of k-spines. This $O(n^{dk}$ time algorithm is intractable for typical sized networks, so we consider two assumptions on the inter-layer edges that reduce the computational complexity while retaining sensitivity. The first assumption asserts that the k-spines of a seed support the same topology of inter-connections. This is motivated by the observation that proteins within the same pathway or complex are typically present or absent in the genome as a group [1]. Thus, we consider the following problem:

Problem 1. d-identical-spine-subnet: Compute a set of d k-spines with identical topologies and maximum score.

Theorem 1. The d-identical-spine-subnet problem admits an $O((n\delta)^d k 3^k)$ solution.

Proof. Recall that a d-subnet can be described as a collection U of size d multisets $U_1, U_2, \ldots U_k$. Let $(U_{i_1}, U_{i_2}) \in E_H$ iff $(u_{i_1}[j], u_{i_2}[j]) \in E_H$ for all $1 \le j \le d$. First, consider the case where each of the d k-spines is restricted to be a path (Figure 1). This implies that the d-subnet itself can be considered as a path $U_{i_1}, U_{i_2}, \ldots, U_{i_k}$. For a subset of species S, let S, let $\mathcal{S}(U, S)$ denote the score of the best d-subnet that uses only species in S, and consists of a path that ends with U. Let $s(U)$ be the species corresponding to U. To compute $\mathcal{S}(U, S)$, note that we only need to recurse using the predecessor of U in the path. Formally:

$$\mathcal{S}(U,S) = \begin{cases} \max\limits_{\substack{(U,W)\in E_H \\ s(W)\in S\setminus\{s(U)\}}} \mathcal{S}(W, S\setminus\{s(U)\}) + \mathcal{L}(U) & \text{if } |S| > 1 \\ \mathcal{L}(U) & \text{if } |S| = 1 \end{cases}$$

Thus, for paths, the overall complexity is $O((n\delta)^d k 2^k)$.
A similar recursion can be applied when searching for k-spines that are trees with identical topology. For a subset of species S, let $\mathcal{S}(U, S)$ denote the score of the best d-subnet that uses only the species in S, and consists of a tree rooted at U. Then for $|S| > 1$:

$$\mathcal{S}(U,S) = \max_{\substack{(U,W)\in E_H, S_1\subset S \\ s(U)\in S_1, s(W)\in S\setminus S_1}} \mathcal{S}(U,S_1) + \mathcal{S}(W, S\setminus S_1)$$

The overall complexity is $O((n\delta)^d k 3^k)$.

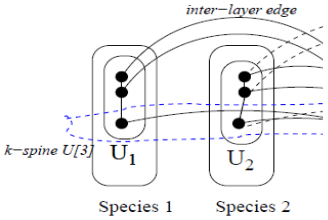

Fig.4. A seed defined by a d-identical-spine subnet, where

the k-spines are restricted to be paths with identical topology. The dashed line encloses one of the three k-spines. A second, slightly different assumption is based on the phylogeny (described as a rooted, binary tree T) of the investigated species. Consider a set of nodes a; b; c whose underlying species follow the phylogenetic triple (s(a); (s(b); s(c))).

We make the following phylogenetic assumption: if a; b; c are connected via inter-layer edges, then b and c must be connected. This implies that we can restrict our attention to k-spines that are guided by the phylogeny T in the following sense: any restriction of the k-spine to species that form a clade in T is a subtree of the k-spine. Note that two guided spines can have very different topologies (see Figure 5).

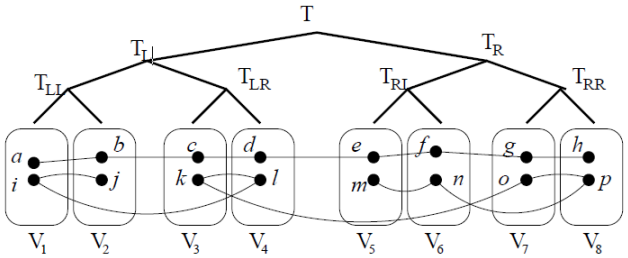

Fig. 5. Sketch of a 2-guided-spine-subnet. Note that while the paths of the two k-spines have different topologies, they are both guided by the underlying tree. Following the notation in the proof of Theorem 2, let $\text{let } U = \{a, j\}, W = \{h, m\},$ and consider two possible distant sets $X = \{d,k\}$ and $Y = \{e,p\}$. By definition, $T_{LL}(U \cup X) = \{a,j\}, T_{LR}(U \cup X) = \{d,k\}, T_{RL}(Y \cup W) = \{e,m\}, T_{RR}(Y \cup W) = \{h,p\}$. Hence, $\mathcal{S}(U,W,T) \geq \mathcal{S}(\{a,j\},\{d,k\},T_L) + \mathcal{S}(\{e,m\},\{h,p\},T_R) \geq \mathcal{S}(\{a,i\},\{b,j\},T_{LL}) + \mathcal{S}(\{c,k\},\{d,l\},T_{LR}) + \mathcal{S}(\{e,m\},\{f,n\},T_{RL}) + \mathcal{S}(\{g,o\},\{h,p\},T_{RR}) \geq \mathcal{L}(a,i) + \mathcal{L}(b,j) + \mathcal{L}(c,k) + \mathcal{L}(d,l) + \mathcal{L}(e,m) + \mathcal{L}(f,n) + \mathcal{L}(g,o) + \mathcal{L}(h,p).$

Problem 2. The d-guided-spine-subnet problem: Compute a set of d k-spines guided by the underlying phylogeny, with maximum score. Unfortunately, we do not know of any efficient algorithm better than the naïve $O(n^{kd})$ for this problem. However, we show a better solution for d-guided-paths, where the k-spines are restricted to be paths guided by the phylogeny.

Theorem 2. The d-guided-path-subnet problem can be solved in $O(k^3(n^3\delta)^d)$.

Proof. Consider a subtree T of the phylogeny with subtrees T_L, T_R respectively. Clearly, each of the d paths will have one end-point in T_L and the other in T_R However, the species topology of these paths is not identical. Therefore, we work with size d subsets U which are not restricted to be within a single species, but instead can span any species in T.

Let $\mathcal{S}(U,W,T)$ denote the best score of a d-guided-path-subnet restricted to a subtree T of the phylogeny such that $s(U) \subseteq T_L, s(W) \subseteq T_R$ are the end nodes. At the base of the recursion T consists of a single node and $\mathcal{S}(U,U,T) = \mathcal{L}(U).$

Otherwise, let $U = \langle u[1], u[2] \ldots u[d] \rangle \in T_L, \text{ and } W = \langle w[1], w[2] \ldots w[d] \rangle \in T_R.$ Denote the root of T by root(T).

For a node $u, \text{s.t. } s(u) \in T.$ define its distant set $\mathcal{D}_T(u) = \{x | \text{LCA}_T(s(u), s(x)) =$ oot(T)g, where LCAT (a; b) is the least common ancestor of a and b in T. Extend this to d elements by defining $\mathcal{D}_T(U) = \{X | \text{LCA}_T(s(u[j]), s(x[j])) = \text{root}(T) \forall j\}$ The key idea to note is that if $X \in \mathcal{D}_T(U)$, then for all $j\ s(x[j]) \in T_L, s(u[j]) \in T_R$ or $s(x[j]) \in T_R, s(u[j]) \in T_L$. Define $T_L(U \cup X)$ $(T_R(U \cup X))$ as the set of all vertices in $U \cup X$ with species in T_L (T_R). Then,

$$\mathcal{S}(U,W,T) = \max_{\substack{X \in D_{T_L}(U) \\ Y \in D_{T_R}(W) \\ (X,Y) \in E_H}} \left(\mathcal{S}(T_{LL}(U \cup X), T_{LR}(U \cup X), T_L) + \mathcal{S}(T_{RL}(Y \cup W), T_{RR}(Y \cup W), T_R)\right)$$

For an example see Figure 5. For the running time, note that there are $k^2 n^{2d}$ cells in the table $\mathcal{S}$. For each cell, there are kn^d choices for the set X and for each there are d choices for a set Y s.t. $(X,Y) \in E_H$. The total time is therefore $O(k^3(n^3\delta)^d)$.

In fact, we can improve the running time to $O((k^2 n^2 \delta)^d)$ (the proof will appear in the full version of the paper), but this is still not practical for reasonable values of n [1].

Extending a seed: The next phase of the algorithm is performing an iterative expansion of the seed by adding, in each iteration, the k-spine that contributes the most to the score. Let us denote by $H = (V', E')$ the current seed, and by $\mathcal{S}(v, S)$ the score of the best partial extension of H by a sub tree that is rooted at vertex v and visits the species in S. Further denote by s(v) the species corresponding to vertex v, and let $W(v) = \sum_{u \in V'} w(u,v)$. Then $\mathcal{S}(v,S)$ can be computed using the following recursive relation:

$$\mathcal{S}(v,S) = \begin{cases} \max_{\substack{(v,w) \in E_H, S_1 \subset S \\ s(v) \in S_1, s(w) \in S \setminus S_1}} \mathcal{S}(v,S_1) + \mathcal{S}(w, S \setminus S_1) & \text{if } |S| > 1 \\ \mathcal{L}(v) & \text{if } |S| = 1 \end{cases}$$

The overall complexity is $O(n\delta k 3^k)$.

There are two speedups one can introduce to this basic extension scheme. The first is to constrain k-spines to paths (rather than trees), obtaining an $O(n\delta k 2^k)$ time algorithm. The second is to set in advance the order of the species along the tree, eliminating the 3^k factor. We term this variant restricted order as opposed to the previous relaxed order variant [1].

IV. WORK DONE

We have done the following work.
Our algorithm, called HopeMap, can be described as follows.

- Obtain and preprocess the PPI network data from PPI network databases, such as DIP. Find all protein pairs that are interacting with each other in a species.
- We have taken the 6 different species as input from the DIP datasets for finding the protein pairs that are interacting with each other in species.
- We have shown this protein network in graphical format for each species separately. In this graph protein id's are taken as a node and interacting proteins are connected by edges.
- We will Find highly similar protein sequences across the species. We have used homolog clustering to identify homolog groups across different species based on all-versus-all BLAST scores or ortholog annotations. Existing tools, such as KO groups and INPARANOID have been used for this purpose. Once homolog groups are identified across the species, a network alignment graph is built based on these groups. The nodes in the graph represent sets of proteins, ideally one from each species, in the same homolog group, and edges represent conserved protein-protein interactions across the compared species. One way of adding the edges between two node pairs (a_1,b_1) and (a_2,b_2) is when both (a_1,a_2) and (b_1,b_2) are directly interacting with each other in their corresponding PPI networks. Other rules for adding edges can be incorporated, such as those introduced in NetworkBLAST.
- Then WE will identify conserved protein interaction regions in the alignment graph. The major algorithm is based on strongly connected-components (clusters) in the alignment graph. A strongly connected component of a graph is a maximal set of vertices in which each vertex is reachable from another. We have used the Depth-first-search algorithm to find the strongly connected components.

V. CONCLUSION

As we stated above, during these paper we have presented the new approach for network alignment by considering the multiple networks alignment problem. This approach is based on the novel representation not only single but also the multiple protein-protein interaction networks as well as the orthology relations between their proteins. In the literature of this paper we have presented the study over network alignment problem and network querying. These proposed methods will claim efficient and faster network alignment of multiple protein-protein networks. For the further work, we will implement the investigated approach here and will show its effectiveness through the extensive performance evaluation of proposed and existing cases.

REFERENCES

[1] "Fast and Accurate Alignment of Multiple Protein Networks", axim Kalaev1, Vineet Bafna2, and Roded Sharan1, 1 School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. fkalaevma,rodedg@post.tau.ac.il 2 CSE, University of California San Diego, USA. vbafna@cs.ucsd.edu.

[2] Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. Nature 422 (2003) 198{207 [3] Uetz, P., et al.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403 (2000) 623{627

[4] Ito, T., et al.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 98 (2001) 4569{4574

[5] Ho, Y., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415 (2002) 180{183.

[6] G.D. Bader, I. Donaldson, C. Wolting, B. Ouellette, T. Pawson, and C. Hogue. Bind–the biomolecular interaction network database. Nucleic Acids Research, pages 242–245, 2001.

[7] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. Genome Research, 16(3):428–435, 2006.

[8] R. Matthews, P. Vaglio, J. Reboul, H. Ge, B.P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence–based searches for conserved protein–protein interactions or "interologs". Genome Research, 11:2120–2126, 2001.

[9] "PAIRWISE ALIGNMENT OF INTERACTION NETWORKS BY FAST IDENTIFICATION OF MAXIMAL CONSERVED PATTERNS" WENHONG TIAN1,2†, NAGIZA F. SAMATOVA1,2, Pacific Symposium on Biocomputing 14:99-110 (2009).

[10] Sharan, R., S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker, *Conserved patterns of protein interaction in multiple species.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1974-9.

[11] Singh, R., J. Xu, and B. Berger, *Global alignment of multiple protein interaction networks.* Pac Symp Biocomput, 2008: p. 303-14.