Identifying Evolutionarily Conserved Protein Interaction Networks

Corban G. Rivera

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

T. M. Murali, Chair

Malcolm Potts

Liqing Zhang

May 23, 2005

Blacksburg, Virginia

Keywords: Conserved Networks, Species Hopping ,Protein-Protein Interaction, Orthologues

Identifying Evolutionarily Conserved Protein Interaction Networks

Corban G. Rivera

(ABSTRACT)

Our goal is to investigate protein networks conserved between different organisms. Given the protein interaction networks for two species and a list of homologous pairs of protein in the two species, we propose a model for measuring whether two subnetworks, one in each protein interaction network, are conserved. Our model separately measures the degree of conservation of the two subnetworks and the quality of the edges in each subnetwork. We propose an algorithm for finding pairs of networks, one in each protein interaction network, with high conservation and high quality. When applied to publicly-available protein-protein interaction data and gene sequences for baker's yeast and fruit fly, our algorithm finds many conserved networks with a high degree of functional enrichment. Using our method, we find many conserved protein interaction networks involved in functions such as DNA replication, protein folding, response to heat, protein serine/threonine phosphatase activity, kinase activity, and ATPase activity.

Acknowledgments

I would like to thank my advisor T. M. Murali for guidance.

Contents

List of Figures

Chapter 1

Introduction

Genome scale biological assays measure many facets of cellular state. Genome sequencing, protein-protein interaction assays, protein-DNA binding experiments, and DNA microarrays are some of the sources of high-throughput biological data. With the extent of whole genome data available, biologists have the capability to make inferences and hypothesis on a much broader scale than before. The field of systems biology has emerged to study the mechanisms controlling many intricate biological processes. Systems biologists construct mathematical models of biological systems and computational methods to organize, analyze, and reason about the influx of high-throughput biological data. The aim is to understand a cell not just as a collection of individual molecules but as a set of modules that behave coherently and interact with each other.

We want to find modules in protein-protein interaction networks. Such a module is a set of interacting proteins that perform a specific task in the cell. Many such protein interaction modules are likely to be conserved in many organisms, especially if they perform fundamental activities in the cell. Motivated by these hypotheses, in this thesis, we will address the question of finding evolutionarily conserved protein-protein interaction networks (PIN) among phylogenetically related species.

1.1 Protein-Protein Interaction

Many biological processes require the collaboration of groups of proteins,which act together as a complex. Protein complexes can be formed by covalent protein interactions,[1] ionic and hydrogen interactions,[2,3] and electrostatic interactions.[3,4] To detect these physically interacting proteins, biologists have developed high-throughput assay techniques. Most recently the two-hybrid technique has been employed on yeast[5–7] and fly.[8] The two-hybrid protein interaction detection mechanism[9] works by generating a signal if a pair of query proteins interact. Specifically, the signal generated in the two-hybrid assay is the transcription of an indicator gene. For this gene to be transcribed, the transcription factor that activates the gene must contain both a sequence binding and an activation domain. In the high-throughput two-hybrid experiment, every gene in the genome is cloned and augmented with a sequence binding domain. Another clone is made with each gene containing an activation domain. Clone pairs from the cross product of the activation and binding domain sets are systematically tested for resulting transcription. If the pair of clones bind to form a complete transcription factor, the indicator gene is transcribed. From the pairs of clones that activate transcription of the indicator gene, the set of interacting proteins is derived.

Co-immunoprecipitation is another method to discover interacting proteins.[10] The interaction detection mechanism works by isolating a bait protein and any proteins bound to the bait. The bait protein is cloned and augmented with a antibody binding tag. To isolate the bait from whole cell lysate, an antibody which is known to bind to the antibody tag on the bait is added. Next, a G-protein, known to bind to most antibodies, is used to extract the antibody, bait protein, and any proteins bound to the bait protein.[10] Subsequently, the purified protein complex is denatured into its component proteins for identification. The experiment yields a complex of two or more proteins containing the bait is derived. Co-immunoprecipitation has been applied on a genome wide scale to detect many protein complexes.[11]

While protein interactions have potential to provide many useful insights into fundamental biological questions, high-throughput biological assays to detect protein-protein interactions may find many interactions that do not take place in the cell.[12]

1.2 Genome Sequencing

The genome of an organism supplies all the information to create and sustain an organism. To date, biologists have sequenced more and 220 genomes[13] including bacteria, archaea, and higher eukaryotes.[14–17] The goal of genome sequencing projects is to identify the primary sequence of the entire genome.

To begin sequencing, the human genome project used a method called hierarchical shotgun sequencing. In hierarchical shotgun sequencing, the whole genome is initially cut with into fragments of about 150 million bases. Biologists insert the fragmented sequences into a bacterial artificial chromosome (BAC). Subsequently, the BAC is transfected into *E. coli* to be cloned. To allow sequencing on smaller fragments, biologists use shotgun sequencing to break the initial fragments into smaller fragments. By looking for sequence overlap at the ends of the fragment, computational methods align the smaller sequenced fragments. With the BAC library sequenced, the whole genome primary sequence is identified.[16,17]

1.3 A Survey of Evolutionary Conservation Studies

Biologists have found that organisms have intrinsic parent-child phylogenetic relationships.[18] Exploiting knowledge of evolutionary relationships, comparative genomics uses the wealth of DNA sequence data generated from genome sequencing projects to discover similarities between biological features of different organisms. Once more than one genome was sequenced, researchers developed methods to compare genomes. Using local alignment, algorithms were developed to locate regions in the genome with high sequence similarity.[19] Each DNA sequence alignment reveals evolutionary relationships between genes in different organisms.

1.3.1 Conserved Genes

Comparative genomics helps determine the evolutionary relationships between genes by comparing the gene sequences of related organisms.[20] Individual genes can be evolutionarily conserved.[21] Conserved genes contain similar sequence motifs.[22] We refer to a pair of genes with shared ancestry as a *homologous* pair of genes. The conserved motifs found in the sequence of genes provide a mechanism to systematically compare genes in the search for homologues. We refer to a pair of homologous genes from different organisms as an *orthologous* pair of genes. Local sequence alignment search tools for comparative genomics like BLAST help identify orthologous pairs of genes in different species.[23]

1.3.2 Conserved Interactions

The notion of a conserved interaction or interlog was first proposed by Walhout et al.[24] Using BLAST to find genes in different organisms with high sequence similarity, Walhout et al. constructed a set of potential orthologues. Given a pair of interacting proteins a and b in one organism and a pair of interacting proteins a' and b' in another Oona's, the quadruple $\{(a, b), (a', b')\}$ is an *interlog* if a and a' are orthologous and b and b' are orthologous. Knowing that a and b interact, Walhout et al. suggested that a' and b' might also interact. Yu et al.[25] continued the work by specifying a joint confidence score to be assigned to pairs of orthologous edges. Yu et al. compute the joint confidence e-value in an interlog $\{(a, b), (a', b')\}$ J as the geometric mean of the BLAST e-values of two homologous pairs (a, a') and $(b, b'$. If the geometric mean of the two e-values is less than 10^{-70}, Yu et al. say that interactions can be transferred. Yu et al. suggest 90,000 possible interactions to be transferred from yeast to worm. Yu et al. found that 45 suggested annotations overlapped with known protein-protein interactions.

1.3.3 Conserved Paths

By finding that protein pathways and complexes are typically either conserved or eliminated from genomes, Pellegrini et al.[26] inspired the search for conserved networks. In the study, annotations of known complexes or pathways were transferred to orthologous complexes or pathways in other organisms. Kelly et al.[27] extended the notion of conserved interactions to conserved paths. Kelly et al. convert two protein-interaction networks $N(V, E)$ and $N'(V', E')$ and a relation of homologous gene pairs θ into a combined protein interaction network $U(\theta, Z)$ such that $((a, b), (c, d)) \in Z$ if the length of the shortest path in N between proteins a and c is less than 3 and the length of the shortest path in N' between proteins b and d is less than 3 for all $(a, b) \in \theta$ and all $(c, d) \in \theta$.

Conserved paths consist of two paths $< a_1, a_2, \ldots, a_n >$ and $< a'_1, a'_2, \ldots, a'_n >$ where $a_i \in V$ and $a'_i \in V'$ and a_i is homologous to a'_i for $1 \leq i \leq n$. Gaps and mismatches are used to allow non-homologous genes to be included in the path. A gap occurs when a_i interacts with a_{i+1} but a'_i and a'_{i+1} do not directly interact. A mismatch occurs when a_i and a_{i+1} do not directly interact and a'_i and a'_{i+1} do not directly interact. A conserved path is a path P in $U(\theta, Z)$. A combined score $S(P)$ defines the confidence assigned to the nodes v and edges e in the path P.

$$S(P) = \sum_{v \in P} log_{10} \frac{p(v)}{p_{random}} + \sum_{e \in P} log_{10} \frac{q(e)}{q_{random}}$$

In this score, $p(v)$ denotes the probability of true homology for vertex $v \in \theta$. $q(e)$ represents the probability that the underlying interaction edges represent a protein interaction that takes place in the cell, and p_{random} and q_{random} are the expected values of $p(v)$ and $q(e)$ respectively in the combined protein interaction network $U(\theta, Z)$.

Kelly et al. use dynamic programming to search for conserved pathways in an acyclic combined interaction graph $U(\theta \times \theta, Z)$. Constraining the length of the path l, the highest scoring path in an acyclic graph can be found in linear time using dynamic programming. As $U(\theta \times \theta, Z)$ is not typically acyclic, many acyclic subgraphs of G are constructed. The results from the acyclic subgraphs are compared, and the top scoring paths are reported. Between yeast and bacteria, Kelly et al. find the mitogen-activated protein kinase (MAPK) signaling and ubiquitin ligation pathways.

Further work on conserved paths incorporated models of evolution. Koyuturk, Grama, and Szpankowski also construct a combined protein interaction network.[28] Like the previous model by Kelly et al., direct interactions in the combined protein interaction network improve the score. Likewise, the score is reduced by the occurrence of gaps. The novel feature in the model is the integration of evolutionary forces that result in gene duplication. With the understanding that gene duplication decreases the conservation of function, a penalty for gene duplication is incorporated into the score for conserved interactions. Given two protein-interaction networks $N(V, E)$ and $N'(V', E')$ and a relation of homologous gene pairs θ, a combined protein interaction network $U(\theta, Z)$ such that $((a, b), (c, d)) \in Z$ if $\pi_N(a, c) < 3$ and $\pi_{N'}(b, d) < 3$ for all $(a, b) \in \theta \times \theta$ and all $(c, d) \in \theta \times \theta$. To evaluate the evolutionarily conservation between a set of nodes from different organisms $P \in V$ and $Q \in V'$, Koyuturk, Grama, and Szpankowski construct the set $M \subseteq Z$ such that $((u, u'), (v, v') \in M$ if u and v interact with u' and v' respectively; the set of gaps $G \subseteq Z$ such that $((u, u'), (v, v') \in M$ if either u interacts with u' or v interacts with v'; and the set $D \in \theta$ such that $(a, b) \in D$ if $a, b \in V$ or $a, b \in V'$. With the sets M, G, and D defined, the score $S(P, Q)$

becomes.

$$S(P,Q) : \sum_{m \in M} \mu(m) - \sum_{g \in G} \nu(g) - \sum_{d \in D} \delta(d)$$

Different choice for the scoring functions μ, ν, and δ alter the conserved paths returned by the algorithm. Koyuturk, Grama, and Szpankowski search for conserved pathways of maximum weight in the combined network. For each node in the combined network, they repeatedly add the node to the conserved pathway such that the score is reduced the most. Koyuturk, Grama, and Szpankowski find a conserved portion of the DNA-depended transcription regulation pathway shared between mouse and human protein interaction networks. They also find the transforming growth factor beta receptor signaling pathway conserved between mouse and human protein interaction networks.

1.3.4 Conserved Complexes

Previous research has shown that dense subgraphs in protein interaction networks correspond to functionally coherent protein complexes. Researchers modeled the dense subgraphs by either cliques,[29] quasi-cliques[30] or other dense subgraphs with high average node degree.[31] Extending the technology developed by Kelly et al. to find conserved complexes, Sharan et al.[32] designed a method for finding evolutionarily conserved complexes. They begin by creating a model of a complex in a single protein-protein interaction network. For control, their probabilistic model contains a conserved complex model M_c and a null model M_n. Assuming a clique-like organization for conserved complexes, the conserved complex model assumes that all pairs of proteins in the complex interact with a high probability β. In contrast, the null model assumes that pairs of proteins in a complex interact with a probability equivalent to the probability that any two proteins in the network interact. For a subset of nodes O and a set of interactions among those nodes O_U in the network, a log likelihood ratio is used to generate a score $L(U)$ for O. The score under the conserved complex model for a set of nodes O represents the similarity of the set of nodes to the conserved complex model compared to the random model.

$$L(U) = log\frac{P(O_U|M_c)}{P(O_U|M_n)}$$

Since it is unknown if an observed protein-protein interaction takes place in the cell, the model uses conditional probabilities to score the likelihood of true interaction.

Subsequently, the conserved complex model is extended from a single protein interaction network to a combined protein interaction network. As described previously, the combined protein interaction network aligns two protein interaction networks $N(V, E)$, and $N'(V', E')$ and relation between homologous gene pairs $(a, b) \in \theta$. For a set of genes from each organism $U \subseteq V$ and $U' \subseteq V'$, the log likelihood ratio becomes the following.

$$L(U, U') = log\frac{P(O_U|M_c)}{P(O_U|M_n)} + log\frac{P(O_{U'}|M_c)}{P(O_{U'}|M_n)}$$

In a combined protein interaction network graph, nodes correspond to pairs of orthologous genes. As BLAST e-values are used to determine orthology between genes, it is unknown if a pair of putative orthologues are truly orthologous. Conditional probabilities are used to model the uncertainty of homology.

With this probabilistic model for conserved complexes, Sharan et al. define a complete weighted orthology graph to search for conserved complexes. In the combined protein interaction graph, nodes have a weight corresponding to the likelihood of orthology between the genes corresponding to that node. Edges in the complete weighted orthology graph have two weights associated with them. The two weights correspond to the probability of interaction for the two interaction edges of the interlog.

The search algorithm first constructs a set of high weight graphs of size at least three then the algorithm refines each such graph using an iterative process. The iterative process either removes low scoring nodes in the complex or adds high scoring neighbors. The process converges when neither removing nodes nor adding nodes increases the conserved complex score. Sharan et al. say that the highest scoring graphs are putative conserved complexes. Sharan et al. find 11 conserved pathways using the conserved complex model. Also, Sharan et al. suggest a functional annotation for a few bacterial proteins involved in a nuclear pore complex.

1.4 Contributions of this Thesis

Previous research confirms that protein interaction networks are conserved in different species.[33,34] Given the protein-protein interaction networks of two organisms, this thesis addresses the problem of finding subnetworks in each protein-protein interaction network that are evolutionarily conserved. In order to facilitate the computation of biologically significant conserved networks, we develop a formal model of conserved networks. Our model requires that conserved networks have two properties. First, the genes in the conserved networks must share a high degree of evolutionary conservation. Secondly, each protein-protein interaction in a conserved network must be of high confidence. Given our model for conserved networks, we propose an algorithm to search for conserved networks. The algorithm begins with a set of interlogs. For each interlog in the set, the algorithm iteratively increases the degree of conservation around the interlog by adding proteins. An interlog converges into a conserved network when adding proteins no longer increase the degree of conservation. At each step of the iteration, the algorithm maintains a pair of networks G and G' and their conservation score. The algorithm keeps one networks, say G fixed. Using orthology relations, the algorithm identifies nodes in the other PIN that could potentially be added to G'. The algorithm adds a node to G' only if the conservation score reduces. If there is such a node, the algorithm continues by keeping G' fixed and expanding G. Thus, our thesis provides a systematic method for finding conserved protein interaction networks, confirming the results of previous research.[33,34]

Our conserved network model has several advantages over the proposed models by Kelly et al.,[27] Koyuturk, Grama, and Szpankowski,[28] and Sharan et al.[32] Our conserved network model does not assume a pattern of interaction for proteins in conserved networks. Also, our conserved network model allows evolutionary conserved proteins to interact through more than one intermediate protein. We detail the comparison in section 3.6

We find many conserved pathways and complexes detected from previous research like kinase cascades, the DNA replication factor C complex, and a protein folding complex. We also find many conserved networks with functions not found using prior models like actin cytoskeleton organization and biogenesis, GTPase activity, and hydrolase activity.

Chapter 2

Mathematical Background

The chapter introduces the mathematical notation used throughout this thesis. As we model our problem with graphs, we describe the properties of a graph here. A graph provides a way to reason about objects and relationships between those objects. The objects are referred to as nodes, and the relationships between nodes are referred to as edges. A basic understanding of graph theory is crucial to understanding this thesis.

2.1 A Formal Graph Model

A *graph* G is defined as a tuple $G(V, E)$ containing a set of nodes V and a set of edges E. Each edge $e = (a, b) \in E$ represents an undirected relationship between the nodes $a \in V$ and $b \in V$. Each edge e can have an associated weight. A *subgraph* $G'(V', E')$ of graph G is a subset of edges $E' \in E$ and a subset of nodes $V' \in V$ such that $a \in V'$ and $b \in V'$ for all $(a, b) \in E'$. In an *unweighted* subgraph graph, all edges have weight 1. The weight of a subgraph G denoted w_G is the sum of the weights over all of the edges. Given a subset of nodes V' from a graph G, an *induced subgraph* is a graph $G'(V', E')$ such that $(a, b) \in E'$ if and only if $(a, b) \in E$ and $a \in V'$ and $b \in V'$. We say that a graph G is *bipartite* if the nodes of the graph can be partitioned into two sets such that no edge is incident on two nodes in the same set. A directed acyclic graph (DAG) is a graph with the property that for each node in the graph there does not exist a path that starts and ends at that node.

The *neighbor set* for a node v is $N(v) = \{v' \in V | (v, v') \in E\}$, the set of nodes connected to v by an edge in E. We say a *path* $P_G(a, b)$ exists between nodes $a \in V$ and $b \in V$ in graph G if there exists an integer $k \geq 1$ and a set of edges $\{(a_i, a_{i+1}),\ 1 \leq i \leq k\}$ such that $(a_i, a_{i+1}) \in E$ for each $1 \leq i \leq k$ and $a_1 = a$ and $a_k = b$. We say that nodes $a \in V$ and $b \in V$ are *connected* if there exists a path $P_G(a, b)$; the weight of $P_G(a, b)$ is the total weight of the edges in $P_G(a, b)$. We use $|\pi_G(a, b)|$ to denote the weight of $\pi_G(a, b)$. The *shortest path* $\pi_G(a, b)$ in a graph G between nodes $a \in V$ and $b \in V$ is a path of least weight connecting a and b. A *spanning tree* of graph G is a subgraph $G'(V, E')$ such that there exists exactly one path connecting $P_G(a, b)$ in G' for all $a \in V$ and $b \in V$. Given a graph G, the *minimum spanning tree* (MST) is the spanning tree of G with least weight.

2.2 Steiner Tree Problem

In the algorithm presented in this thesis, we will often need to find a subgraph of small weight that connects a set of nodes. Formally, given a graph $G(V, E)$ and set of nodes in the graph $W \subseteq V$, the problem is to find a subgraph $T(X, Y)$ such that $W \subseteq X$ and the weight of T is the smallest among all subgraphs of G that contains W. This problem is called the Steiner tree problem.[35–37] Since the Steiner tree problem is NP-Hard,[38] we use the following algorithm given by Kou et al.[39] to find an approximate solution.

1. Construct the complete weighted graph $G'(W, E')$, where the weight of an edge $(a, b) \in E'$ is $|\pi_G(a, b)|$, the weight of the shortest path in G between nodes a and b
2. Construct a minimum spanning tree $T(W, A)$ of G'[40]
3. Construct $G^*(N^*, E^*)$ as follows: for each $(a, b) \in A$, add the shortest path $\pi_G(a, b)$ to G^*

The algorithm outputs the graph G^*. Kou et al.[39] prove that the weight of G^* is at most two times the weight of the minimum weight Steiner tree connecting the nodes in W. Solutions to the steiner tree problem help to identify minimal, connected subgraphs constrained by the requirement of including a subset of the nodes.

2.3 Hypergeometric Distribution

In this thesis, we often need to assign a score to the likelihood of randomly selecting a given number of marked elements from a universe U of elements. Let $T \subseteq U$ be the set of marked elements. Suppose we have a computational procedure that selects a set of elements $U' \subseteq U$ and that U' contains a subset T' of T. Let u, u', t, and t' denote the sizes of the sets U, U', T, and T' respectively. We are interested in computing the probability that this event is statistically significant. The null hypothesis H_0 is that if we select u' items from a set of u items uniformly at random without replacement, our random sample will contain at least t' samples from the set T. The alternative hypothesis H_1 is that this event cannot happen at random. We accept H_1 if and only if the probability of H_0 is less than a user specified threshold. The number of ways of choosing t' elements from a set of t elements is $\binom{t}{t'}$. The number of ways of choosing the remaining elements of T is $\binom{u-t}{u'-t'}$. The total number of ways of choosing u' elements from a set of u elements is $\binom{u}{u'}$. Therefore, the probability of this event is

$$H(u, t, u', t') = \frac{\binom{t}{t'}\binom{u-t}{u'-t'}}{\binom{u}{u'}}$$

Since H_0 is the probability that the random sample contains t' or more samples from T, to calculate the probability that H_0 is true, we sum $H(u, t, u', i)$ for $t' \leq i \leq t$.

$$Pr(H_0\ is\ true) = \omega(u, t, u', t') = \sum_{i=t'}^{min(t,u')} \frac{\binom{t}{i}\binom{u-t}{u'-i}}{\binom{u}{u'}}$$

We will often perform many such tests simultaneously. In this case, we need to correct for the possibility that one of the tests might be true by random chance. We use the Bonferroni correction[41] in this thesis to correct for testing n hypotheses simultaneously, we multiply the probability of each event by n and accept the hypothesis only if the resulting probability is less than the user-specified threshold..

Chapter 3

Methods

In this chapter, we present a formal model for conserved PINs. We also describe the algorithm for detecting conserved networks. For each of the conserved network model parameters, we describe how we select their values. We also make a comparison between our conserved network model and previous models.[27, 28, 32]

3.1 Protein-Protein Interaction Networks

A natural representation for the set of protein-protein interactions in an organism is as a graph $G(V, E)$. Formally, V is the set of genes in the genome of the organism. For each observed protein interaction between genes $a \in V$ and $b \in V$, we include the edge (a, b) in E.

In this thesis, we use the weight of an edge to represent our confidence in the fact that corresponding interaction is a true interaction. Many methods have been developed to assign the reliability of protein-protein interactions.[12,42] In this thesis, we use the method developed by Goldberg and Roth.[42]

Many researchers have observed that the degree distribution of protein interaction networks is well described by the power law distribution.[43,44] Other researchers[42] have also observed that protein interaction networks have the small world property. There are many informal way of defining this property including the fact that the average distance between all pairs of nodes in the network is small and that the neighbors of a node are themselves connected. Goldberg and Roth use the second property to access confidence in an interaction; if the nodes incident on an edge have more common neighbors than would be expected by chance, then they assign a high confidence to that edge. It is natural to use the hypergeometric distribution to calculate the true significance of the observed number of common neighbors of the node incident on an edge. Let the edge e connect nodes a and b. Following Goldberg and Roth, we set the weight of e to be

$$\rho_e : \quad \omega(|N(a) \cup N(b)|, |N(a)|, |N(b)|, |N(a) \cap N(b)|)$$

ρ_e measures the probability that if we select the neighbors of b from the set $N(a) \cup N(b)$, we will select $N(a) \cap N(b)$ or more neighbors from the marked set $N(a)$, This number is small if a and b share many common neighbors. In this thesis, we delete all edges with weight less than 0.05. We then consider the graph to be unweighted. Specifically, we set the weights of all remaining edges to be 1.

3.2 A Formal Model for Conserved Networks

In order to facilitate the computation of biologically significant conserved networks, a formal definition is necessary. Our model requires that conserved networks have two properties. First, the genes in the conserved networks must share a high degree of evolutionary conservation. Secondly, conserved networks must contain edges in which we are very confident.

Let $S_1(V_1, E_1)$ and $S_2(V_2, E_2)$ be protein interaction networks in different organisms. H is a bipartite graph where each edge $(a, b) \in V_1 \times V_2$, where a is homologous to b, specifies the set of homologous pairs of proteins. An edge $e(a, b) \in H$ has a weight B_e equal to the DNA-sequence similarity between genes a and b. Specifically, we denote b_e as the BLAST e-value between genes a and b.

We use two parameters κ and λ to filter our protein-protein interaction networks. Given a protein interaction network $S_1(V_1, E_1)$, we compute edge weights as described in Section 3.1 given a parameter $\kappa > 0$, we remove edges $e \in E_1$ such that $\rho_e > \kappa$. We apply a similar operation to S_2. The parameter κ establishes a minimum confidence threshold for protein interaction edges. Given a set of homologous pairs of proteins H and a parameter λ, we remove edges of $e \in H$ such that $b_e > \lambda$. The parameter λ sets a minimum degree of confidence between putative orthologues a and b.

Definition 1: Conserved Network

We define a *conserved network* as a triple $C(T_1, T_2, O)$ where $T_1(P_1, Q_1)$ and $T_2(P_2, Q_2)$ are connected subgraphs of S_1' and S_2' respectively and $O \subseteq H'$ such that $(a, b) \in O$ if and only if $a \in P_1'$

and $b \in P_2'$. Thus, O is a subset of H induced by the nodes in P_1 and P_2. Therefore, if O contains more relations than would be expected if we selected P_1 and P_2 randomly from T_1 and T_2, then the evolutionary conservation between T_1 and T_2 is high. Given a universe of $|V_1 \times V_2|$ pairs of proteins, of which the $|H|$ orthologous pairs are marked, what is the probability that if we select $|P_1 \times P_2|$ pairs uniformly at random from the universe, we will obtain $|O|$ or more protein pairs that are orthologous? It is natural to use the hypergeometric distribution to determine if O contains more relations than would be expected by chance. Therefore, we define the *conservation score* $\phi(P_1, P_2)$ to be

$$\phi(P_1, P_2) = \omega(|V_1 \times V_2|, |H|, |P_1 \times P_2|, |O|)$$

where $\omega()$ is the function for calculating the hypergeometric probability introduced in section 2.3. Our goal is to find conserved networks with low conservation scores. We can now formally state the problem we want to solve as follows:

Given two protein-protein interaction networks $S_1(V_1, E_1)$ and $S_2(V_2, E_2)$, a relation H between homologous pairs of genes, and parameters $\alpha, \kappa, \lambda > 0$ find all conserved networks $C(T_1(P_1, Q_1), T_2(P_2, Q_2), O)$ that satisfy the following property:

i $\phi(P_1, P_2) \leq \alpha$

ii for each $e \in Q_1 \cup Q_2, \quad \rho(e) \leq \kappa$

iii for each $(a, b) \in O, \quad B_{a,b} \leq \lambda$

We show that to exhaustively search for all conserved networks with a conservation score less than a threshold is NP-complete. A biclique is a bipartite graph such that there exists an edge between every pair of nodes in different partitions. Given a bipartite graph G and positive integer k, does there exist a biclique with at least k edges. Given that the maximum edge biclique problem is NP-complete,[45] we show that deciding if any conserved interaction networks exist between two organisms can be reduced to the maximum edge biclique problem in polynomial time. Let the protein-protein interaction networks be fully connected. We denote the orthology graph as the set of all proteins from both protein interaction networks with orthology relationships between the proteins. We have that conserved networks have a conservation scores less than α. The conservation score is minimum for a set of proteins from each protein interaction network that is fully connected by orthology relationships. To determine if there exists a conserved network with conservation score less than α, we must find the conserved network with the least conservation score. We have that finding the conserved network with least conservation score can be solved by finding the maximum edge biclique in a bipartite orthology graph.

3.3 Species Hopping: An Algorithm for Finding Conserved Networks

Since finding networks with small conservation scores is computationally hard, we have developed a novel ”species hopping” heuristic for finding such networks. The algorithm is not guaranteed to find all networks with a conservation score less than α. The algorithm starts with a basis set, a conserved network with a small set of nodes. Iteratively, the algorithm expands the conserved network to reduce the conservation score. At each step of the iteration, the algorithm maintains a pair of networks G and G' and their conservation score. The algorithm keeps one networks, say G fixed. Using orthology relations, the algorithm identifies nodes in the other PIN that could potentially be added to G'.

Henceforth, we assume that S_1 and S_2 only contain edges with weights less than κ and that H only contains relations with BLAST e-value less than λ. Therefore, all conserved networks constructed by the algorithm we describe will satisfy properties (ii) and (iii).

Basis Set We say that a pair of proteins in a PIN are closely interacting if they directly interact or they share a common neighbor. By capturing closely interacting homologous pairs of genes, the species hopping algorithm is seeded with basic units of conservation. The elements of the basis set closely resemble interlogs.[24]

Given protein interaction networks $S_1(V_1, E_1)$ and $S_2(V_2, E_2)$ and a set of homologous pairs of protein H, we construct a *basis set* B of conserved networks where each conserved network $C(T_1, T_2, O)$ has the following properties (i) O contains 2 pairs of nodes $(a, a') \in H$ and $(b, b') \in H$, (ii) $T_1 = \pi_{S_1}(a, b)$ and $T_2 = \pi_{S_2}(a', b')$, and (iii) $\pi_{S_1}(a, b) < 3$ and $\pi_{S_2}(a', b') < 3$. B consists of all such conserved networks

Inductive Step After iteration k of the algorithm, we have a conserved network $C^k(T_1^k, T_2^k, O^k)$. In iteration $k+1$, we construct a new conserved network $C^{k+1}(T_1^{k+1}, T_2^{k+1}, O^{k+1})$ such that $\phi(C^{k+1}) < \phi(C^k)$. If k is odd then $T_1^k = T_1^{k+1}$ and $T_2^k \subseteq T_2^{k+1}$. If k is even then $T_2^k = T_2^{k+1}$

$T_1^k \subseteq T_1^{k+1}$. In other words, in iteration $k+1$, we keep either T_1^{k+1} or T_2^{k+1} fixed and "expand" the other graph. Without loss of generality, we assume that k is odd in the following discussion. We use two operations to construct T_2^{k+1} from T_2^k. First, we use the cross-over operation to find a set $P_2' \subseteq V_2$ of nodes. Each node in P_2' is adjacent to a node in T_1^{k+1} in the bipartite graph H. We then use the expansion operation to find a small subgraph containing P_2' in S_2. The process is shown in figure 3.1.

The *cross-over* operation defines the process of how to step between species. Given a set of homology pairs H and a conserved network $C(T_1(P_1, Q_1), T_2(P_2, Q_2), O)$, the cross-over operation $\eta_H(T_1)$ returns the set P_2'. The cross-over operation $\eta_H(P_1)$ selects the element a_1 such that $\phi(P_1, P_2 \cup \{a_1\} < \phi(P_1, P_2 \cup \{a\})$ and $\phi(P_1, P_2 \cup \{a_1\} < \phi(P_1, P_2)$ for all $(a, b) \in H$ and $b \in P_1$. The cross-over operation returns $P_2' = P_2 \cup \{a_1\}$.

As we would like a small subgraph connecting the nodes of P_2 in S_2, the *expansion* operation connects a set of nodes in a graph. Given a graph $S_2(V_2, E_2)$ and a set of nodes P_2, the expansion operation $\zeta_{S_2}(P_2)$ uses the algorithm by Kou et al.[39] described in section 2.2 to construct the Steiner tree.

Therefore we construct P_2^{k+1} by applying the cross-over followed by the expansion operation of P_2^k, ie $P_2^{k+1} = \zeta_{S_2}(\eta_H(P_1^k))$. After computing P_2^{k+1}, we compute the induced subgraph $T_2^{k+1}(P_2^{k+1}, Q_2)$. We add edges $e(a, b) \in H$ to O iff $a \in P_1^{k+1}$ and $b \in P_2^{k+1}$. The conserved network $C(T_1^{k+1}, T_2^{k+1}(P_2^{k+1}, Q_2), O)$ is prepared for the next inductive step.

Stopping Condition The iterative process stops when subsequent cross-over and expansion operations no longer reduce the conservation score. Specifically, if there does not exist an element a_1 during the cross-over operation, the species hopping process has converged. Given a converged network $C(T_1(P_1, Q_1), T_2(P_2, Q_2), O)$, we add converged network to the set N if $\alpha > \phi(P_1, P_2)$ The species hopping algorithm returns the set N.

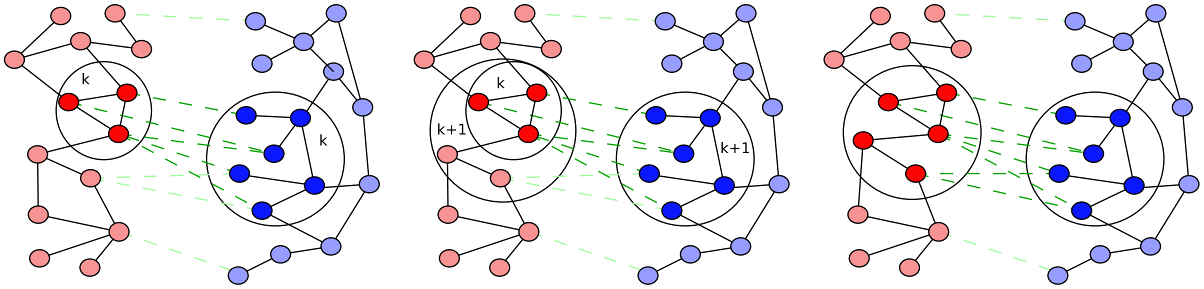

Figure 3.1: An example of a species hopping inductive step from k to $k + 1$. Networks connected by solid lines denote protein-protein interaction networks. Dashed edges denotes a homologous relationship between proteins. (Left) Conserved networks at the end of iteration k. (Middle) Cross-over and expansion operations are used to construct a new conserved network during k+1. (Right) End of iteration $k + 1$ In this example the conservation score is calculated as $\omega(14 * 16, 9, 5 * 6, 7)$.

3.4 Conserved Network Functional Enrichment

In this section, we discuss how we post-process computed conserved networks to characterize them in terms of the functions of the constituent proteins. If the number of genes in a conserved network that share a particular function is more than would be expected by chance, we say that the network is *enriched* in the function. Given a protein-protein interaction network $S(V, E)$ and a subgraph $T(P, Q)$ of S , we compute the functions enriched in T as follows: Let F_g be the set of functional annotations for gene g, and let $F_P = \bigcup_{g \in P} F_g$ the set of functions annotating all proteins in P. For each function $f \in F_A$, we compute the functional enrichment score

$$\phi_{f,T} = \omega(V, G, P, W)$$

where G is the set of proteins in V annotated by f and W is the set of proteins in P annotated by f, and $\omega()$ is the function that computes the hypergeometric probability (see Section 2.3). We say f is *enriched* if $\phi_{f,T}$ is at most a user-defined p-value. In this thesis, we use a p-value of 1×10^{-4}. Since we perform the test for enrichment for each function in F_P. we perform multiple hypothesis correction. We use the Bonferroni correction (described in Section 2.3) to account for testing multiple hypotheses simultaneously.

3.5 Value Selection for Model Parameters

For the three model parameters α, κ, and λ, we describe the method used to select their values. The value of α is determined empirically. We observe a correlation between the conservation score of a conserved network and the score of the most enriched function in the conserved network. Our hypothesis is that a good correlation between the conservation score and the score of the most enriched function indicates the ability to find conserved networks using our model. For a sample of 283 conserved networks with conservation score $\delta < 10^{-40}$, we find a correlation of 0.85. The value of the correlation indicates that 70.8% of the variation in the score of the most enriched function for a conserved network is explained by the conservation score see figure (3.2). To gain a desired level of maximum functional enrichment in conserved networks, we obtain a corresponding value for α. For the results presented in this thesis, we have selected the value 10^{-40} for α.

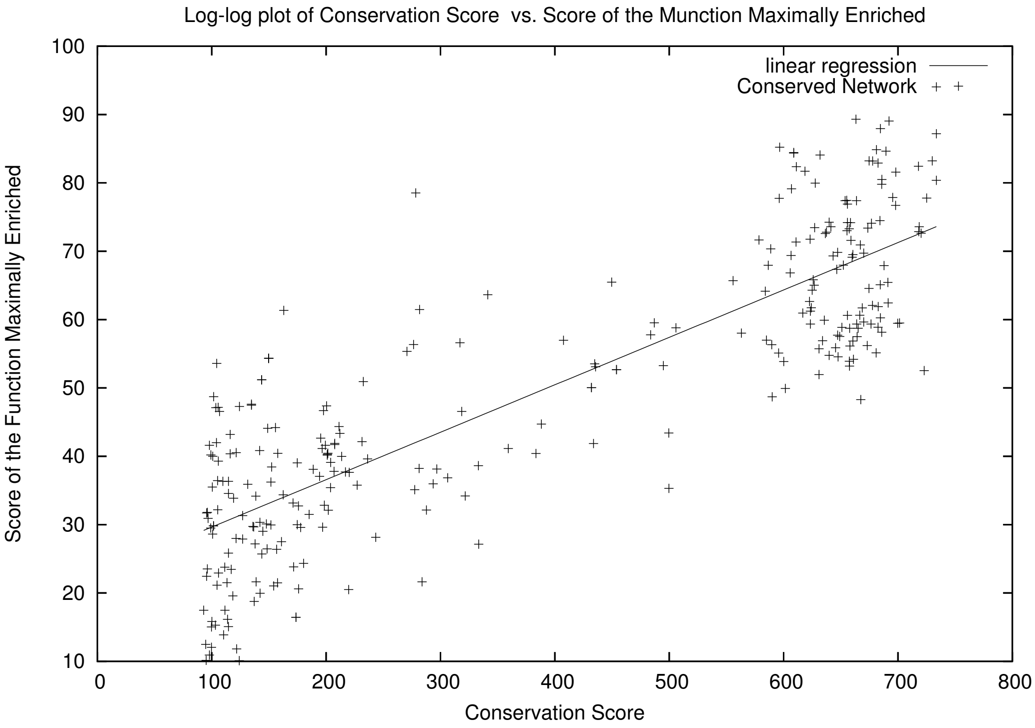

Figure 3.2: We find a positively correlated relationship between the conservation score and the maximum functional enrichment. For each conserved network, we plot the log of the conservation score of the network on the x-axis and the log of the score of the most enriched function in the conserved network on the y-axis.

We select the value for the model parameter λ such that between genes sequence similarity is rarely found by random chance. A BLAST e-value of 10^{-5} corresponds to a Bonferroni corrected e-value of 0.01. We also have that genes with high sequence similarity are likely to be orthologous.[46] Therefore, to obtain a high confidence set of homologous pairs of genes, we assign 10^{-20} to λ. We select the value for the model parameter κ to ensure that we consider only reliable protein interactions. Goldberg and Roth[42] generate a probability to represent the reliability of each protein-protein interaction as described in section 3.1. To select high confidence protein-protein interactions, we assign κ to 0.05.

3.6 Comparison to Existing Methods

Our conserved protein interaction subnetwork model is more general and flexible in interaction patterns for proteins in conserved subnetworks. Wagner estimated that nearly half of all the protein-protein interactions get replaced every 300 million years.[47] The evolutionary distance between species is described by the amount of time since their speciation from the least common ancestor. With greater evolutionary distance between organisms, proteins have a greater change in their patterns of interaction. Consequently, when looking for conserved networks between species with greater evolutionary distance, it is necessary to allow proteins in conserved networks to have a more diverse pattern of interaction.

Our conserved protein interaction subnetwork model provides more generality than the conserved protein interaction subnetwork models proposed of Kelly et al.,[27] Koyuturk, Grama, and Szpankowski,[28] and Sharan et al.[32] The improved sensitivity of our model arises from the lack of presumption about conserved protein interaction network topology. The conserved protein interaction subnetwork model of Kelly et al. assumes a linear structure for conserved protein interaction networks. The conserved protein interaction model of Sharan et al. assumes a clique-like interaction structure between all the genes in the conserved protein interaction subnetwork. Our conserved network model subsumes both these graph topologies naturally.

Since we do not construct a combined protein-protein interaction network, our conserved network model allows a more flexible pattern of interaction for proteins in conserved networks. The previous methods for finding conserved subnetworks of,[27] Koyuturk, Grama, and Szpankowski,[28] and Sharan et al[32] construct a combined protein interaction network. The construction of the combined network has the property that evolutionarily conserved interactions requires proteins to interact directly or through a single intermediate protein. Previous methods defined gaps and mismatches to account for variation of protein interaction patterns in conserved networks. Provided the degree of evolutionary conservation is high, our conserved protein interaction network model allows interaction between proteins with homologous counterparts through more than one intermediate protein.

Chapter 4

Results and Discussion

The analysis of conserved networks requires the integration of a broad base of whole-genome biological data. The Database of Interacting Proteins (DIP)[48] provided the protein-protein interaction data for yeast and fly. Whole genome yeast two-hybrid, co-immunoprecipation, and other protein-protein interaction experiments contribute their findings to the DIP database. Of 14271 yeast and 20947 fly protein-protein interactions, we find 8485 and 14646 respectively with clustering coefficient less than 0.05. The FlyBase Consortium[49] and the Saccharomyces Genome Database[50] provided the gene sequences for fly and yeast respectively. Running BLAST on all yeast genes against all fly genes, we find 64433 homologous pairs of proteins, and 18943 with e-value less than 10^{-20}. 5525 have proteins in both of the protein interaction networks. With 21954 basis sets, we find 940 conserved networks with an evolutionary conservation score less than 10^{-40}. The Gene Ontology Consortium[51] provided the functional annotations for proteins.

4.1 Orthologous Genes Share Function

To determine whether we could find conserved networks that share function, we first assessed whether orthologous genes share function. Since the functions in GO are related by parent-child relationships, we can represent these relationships in a directed acyclic graph (DAG) as described in section 2.1. If two genes do not share a common function, we measure the distance between these functions in the underlying DAG. We consider the DAG to be an undirected network so as to capture the situation when two functions are related by a common child or parent. We denote this graph by L. Let F_a denote the set of functions in GO annotating a gene a. For each pair (a, b) of homologous genes, we define the functional distance $fd(a, b)$ to be the smallest distance between all pairs of functions annotating a and b. $fd(a, b) = \min_{f \in F_a, g \in F_b} |\pi_L(f, g)|$, ie for all the annotations given to a each gene in the homologous pair (a, b), we find the pair of annotations with the smallest distance in L We perform the analysis separately for each GO category. Additionally, we exclude GO terms referring to unknown functional annotations. We compute the shortest distance between any two nodes in the graph computed using breath first search.

We find a large number of homologous pairs between yeast and fly that share a GO function (see figure 4.1). Also, many homologues share functions that are close to each other in L. If $fd(a, b) = 0$, then a and b have at least one shared function. If $fd(a, b) = 1$, then either a or b has a function that is either a specialization or generalization of a function performed by its orthologue. The large number of homologues with high functional similarity suggests that we might find conserved networks that share function. Of 18943 orthologues between yeast and fly, we have that 8372 share a molecular function, 13589 share a biological process, and 2060 share a cellular component.

We find that many homologous pairs of genes share similar functions. For pairs of homologous genes with a BLAST e-value less than 10^{-20}, we calculate the distance in the GO DAG between nearest functions for each pair of orthologous genes. For each distance, we plot the ratio of the number of homologues to the total number of homologous pairs of genes such that both genes have functional annotations in the GO category.

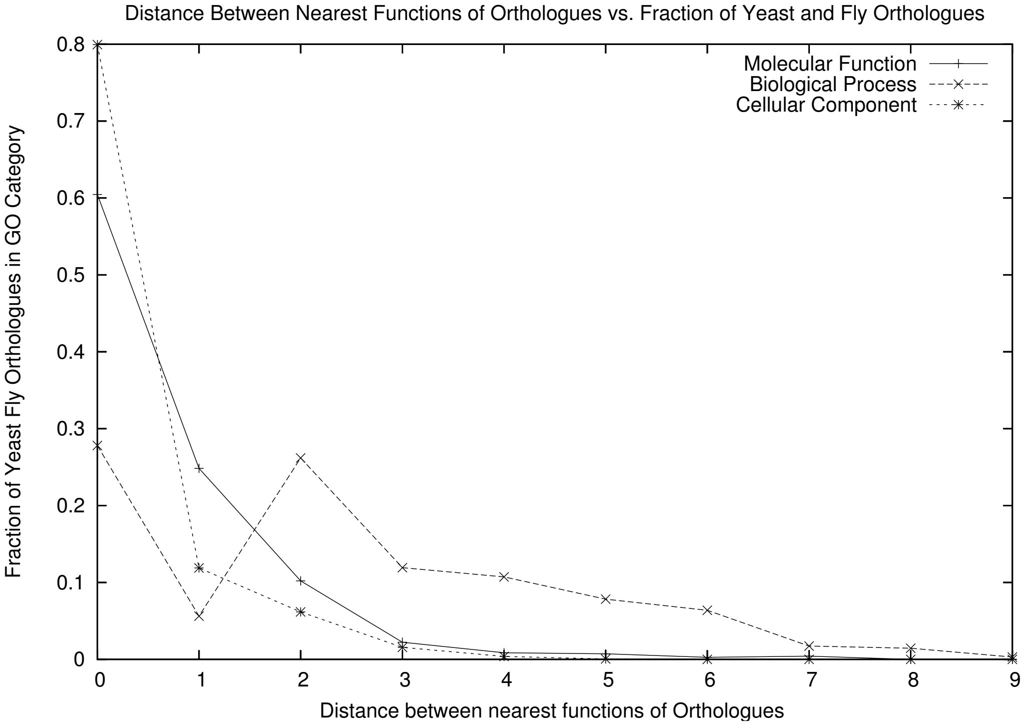

Figure 4.1: We find that many homologous pairs of genes share similar functions.

4.2 Conservation Score of Conserved Networks

To validate the significance of conserved networks found by the species hopping algorithm, we run the species hopping algorithm on randomly constructed protein-protein interaction data. Like Sharan et al.,[52] we construct a randomized protein-protein interaction network from $G(V, E)$ by crossing the edges in E. Formally, we select two edges (u, u') and (v, v') uniformly from E. We add the edges (u, v') and (v, u') to E and remove (u, u') and (v, v') from E. Additionally, we require that no self-loops are created as a result of the operation. We repeat the crossing procedure $|E|$ times. The procedure maintains node degree while randomizing the network. As the edges are selected uniformly at random, we have that the expected value for the number of crossings for each edge is 2.

We obtain 20 sets of conserved networks from running the species hopping algorithm on 20 different pairs of randomized protein-protein interaction networks. To facilitate comparison of conservation scores, we take the reciprocal of the conservation score. For each set of conserved networks, we compute the mean and standard deviation of the conservation score over all conserved networks in the set. Over these 20 sets of conserved networks, the summary statistics included a mean conservation score of 308.043 and a standard deviation of 222.44. When we run the species hopping algorithm on the original protein-protein interaction data for yeast and fly, we have a mean conservation score of 4220.35 and standard deviation of 1656.08. We observe that the mean conservation score of conserved networks found using original protein interaction data is at least an order of magnitude greater than the mean conservation scores of conserved networks found using randomized protein-protein interaction data.

4.3 Conserved Protein Interaction Networks between Yeast and Fly

Our algorithm constructs 21954 basis sets. We find 940 conserved networks with an evolutionary conservation score less than 10^{-40}. For comparison, we compute the functions enriched in T_1 (with respect to S_1), T_2 (with respect to S_2), and $T_1 \cup T_2$ (with respect to $S_1 \cup S_2$). By calculating the functions enriched in both individual protein interaction subnetworks and the whole conserved network, we can determine if the individual conserved protein interaction subnetworks share a similar function. We can also use the conserved network to find enriched functions that are not found in individual protein interaction subnetworks. Using the conserved network model, we find the DNA replication factor C complex (function enrichment 2.4×10^{-18}) The complex is also found by Sharan et al.[52] using the conserved complex model (see figure 4.2).

The conserved networks with the most significant conservation scores are collections of genes with kinase activity (4.3). It is well known that kinase catalytic domains are evolutionarily conserved.[53] In conserved networks of kinases with conservation scores less than 2.6×10^{-319}, we find a functional enrichment score of 1.43×10^{-41} for protein amino acid phosphorylation. Our result provide additional evidence for the conservation of the kinase catalytic domain.

Using the species hopping algorithm, we find many conserved networks with highly significant

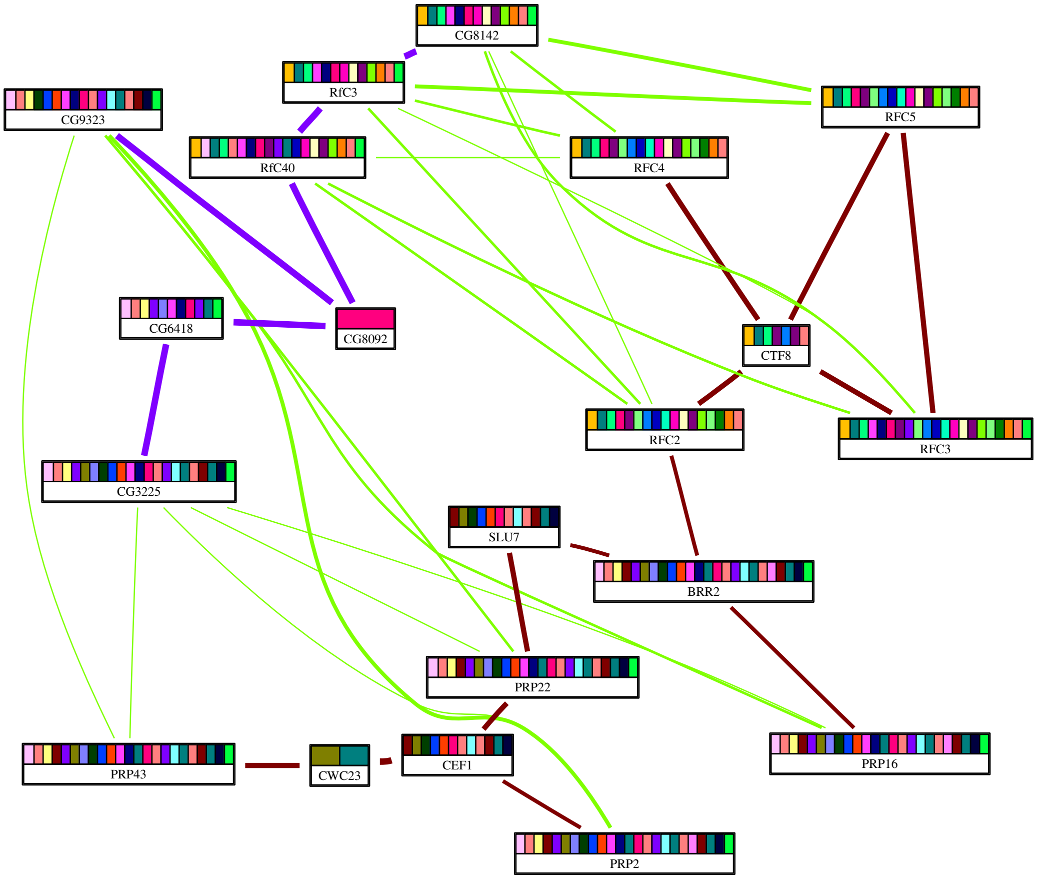

Figure 4.2: Using the conserved network model, we find the DNA replication factor C complex (function enrichment 2.4×10^{-18}). Proteins are denoted by boxes. The colors above the boxes represent the enriched functions for the protein. Dark edges represent protein-protein interactions. Light edges represent relationships between homologous pairs of proteins.

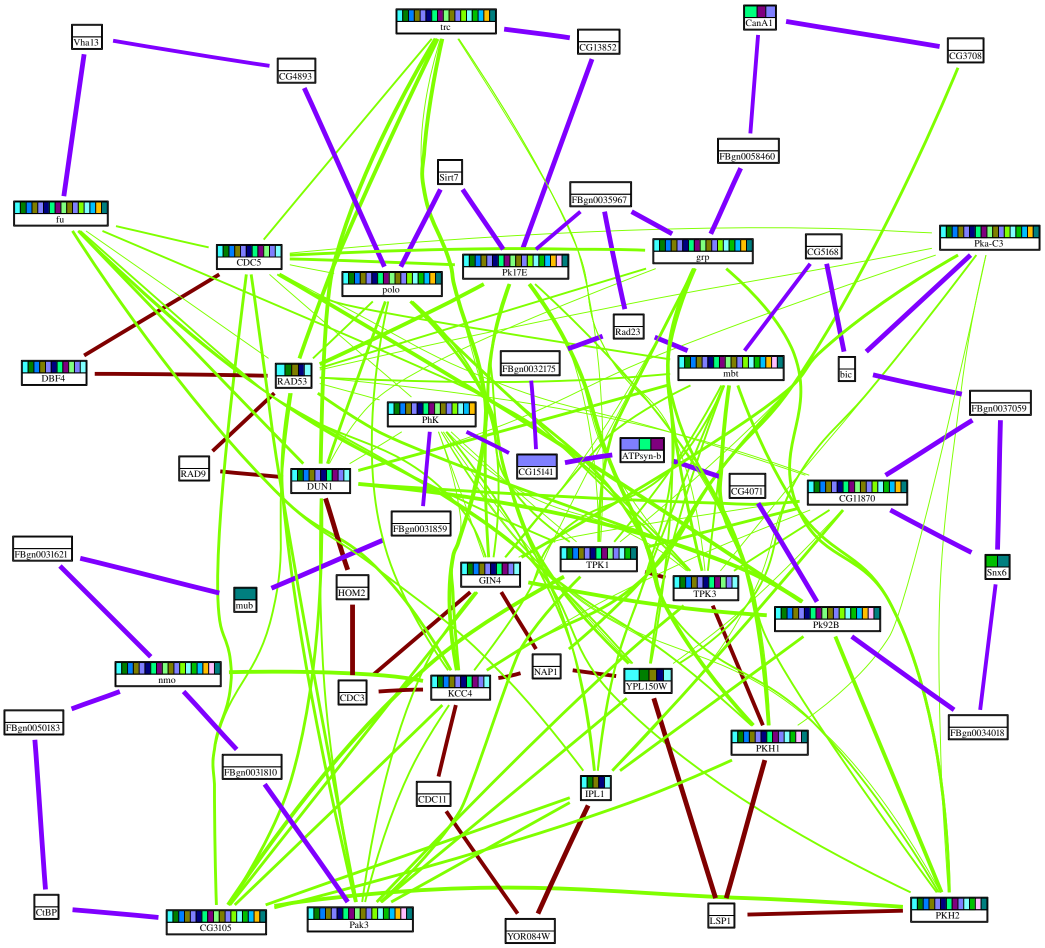

Figure 4.3: We find a large degree of evolutionary conservation in the kinase catalytic domain. The kinase catalytic domain is known to be conserved from prior research.

conservation score and functional enrichment. A few examples are presented here to illustrate observations (4.4). The remaining conserved networks are available online at:

http://bioinformatics.cs.vt.edu/∼cgrivera/hop/

4.4 Future Work

We would like to extend the conserved network model and species hopping algorithm to accommodate greater than two species. As we find many putative proteins in conserved networks, we would like to use conserved networks to transfer functional annotations to unannotated proteins. With many new high-throughput protein-protein interaction data sets, we would like to find conserved protein interaction networks between other organisms. In addition to protein-protein interaction data, we could incorporate gene co-expression data. By combining protein-protein interaction and synthetic lethality data, we could use species hopping to find functionally parallel pathways within the same organism.

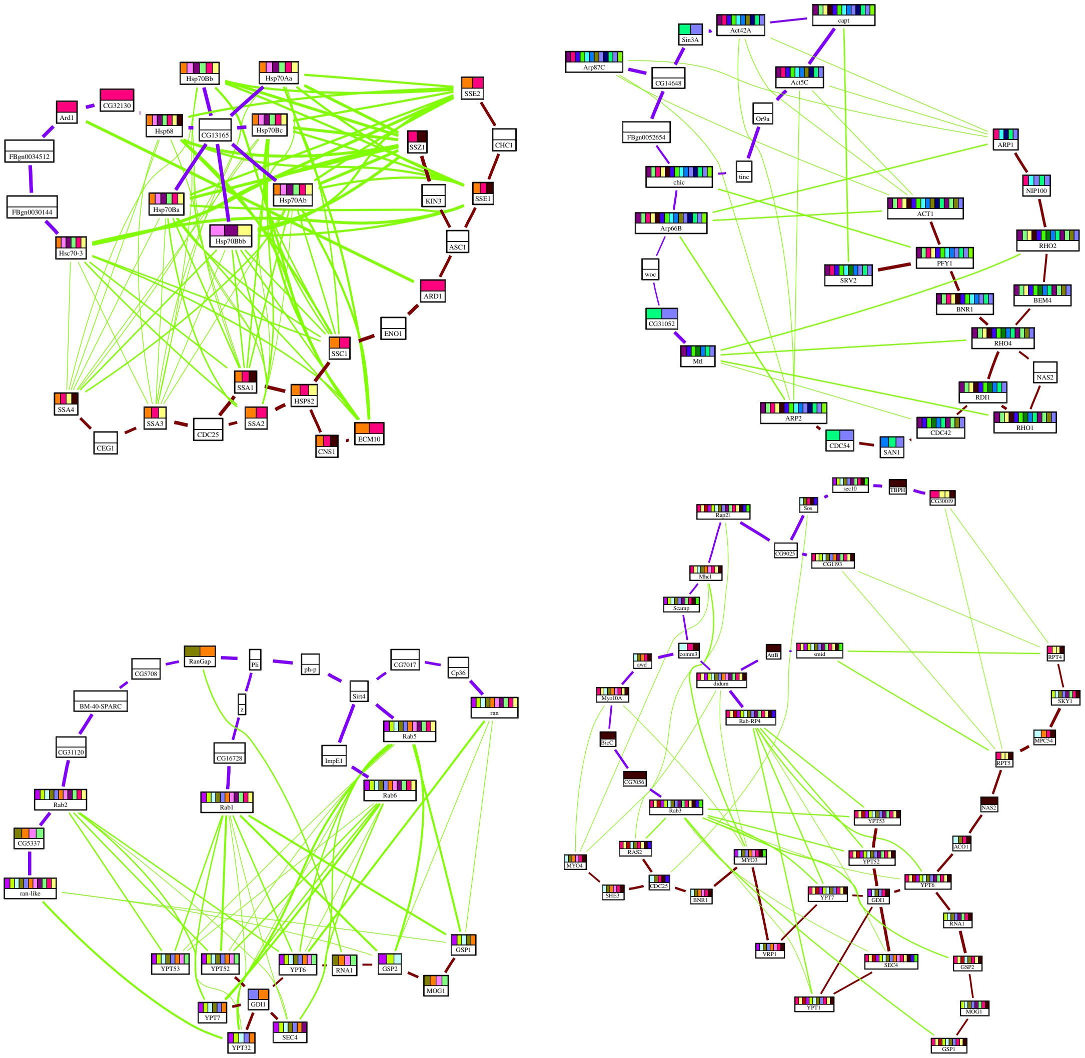

Figure 4.4: Examples of evolutionarily conserved networks. *Top Left* We find a conserved network for protein folding (functional enrichment 7.4×10^{-24}). Protein folding is a function that is known to be conserved by evolution. *Top Right* The conserved network involved in actin cytoskeleton organization and biogenesis (functional enrichment 3.7×10^{-14}). This conserved network demonstrates the ability to detect conserved networks with a high degree of change in protein interaction patterns *Lower Left* We find a conserved network involved in GTPase activity (functional enrichment 4.3×10^{-18}). There are few references to suggest that GTPase activity is conserved, but this study suggests that the function is evolutionarily conserved. *Lower Right* A conserved network involved in hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides (functional enrichment 8.4×10^{-14}). The conserved network illustrates the detect conserved networks with non specific topologies.

Bibliography

1. Nocek, J. M., Zhou, J. S., Forest, S. D., Priyadarshy, S., Beratan, D. N., Onuchic, J. N., and Hoffman, B. M. Theory and Practice of Electron Transfer within Protein-Protein Complexes: Application to the Multidomain Binding of Cytochrome c by Cytochrome c Peroxidase. Chem. Rev. 96:2459–2489, 1996.

2. Loo, J. A. Studying noncovalent protein complexes by electrospray ionization mass spectrometry. Mass Spectrometry Reviews 16(1):1–23, December, 1998.

3. Xu, D., Tsai, C. J., and Nussinov, R. Hydrogen bonds and salt bridges across protein protein interfaces. Protein Engineering 10(9):999–1012, 1997.

4. Norel, R., Sheinerman, F., Petrey, D., and Honig, B. Electrostatic contributions to protein protein interactions: Fast energetic filters for docking and their physical basis. Protein Science 10(2147-2161), 2001.

5. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M.

A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature 403(6770):623–627, February, 2000.

6. Walhout, A. J. and Vidal, M. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. Methods 24(3):297–306, July, 2001.

7. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A 97(3):1143–1147, February, 2000.

8. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. A protein interaction map of drosophila melanogaster. Science 302(5651):1727–1736, December, 2003.

9. Fields, S. and Sternglanz, R. The two-hybrid system: an assay for protein-protein interactions. Trends Genet 10(8):286–292, August, 1994.

10. Ransone, L. J. Detection of protein-protein interactions by coimmunoprecipitation and dimerization. Methods Enzymol 254:491–497, 1995.

11. Gavin, A. C., Bsche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–147, January, 2002.

12. Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22(1):78–85, January, 2004.

13. Bernal, A., Ear, U., and Kyrpides, N. Genomes online database (gold): a monitor of genome projects world-wide. Nucleic Acids Res 29(1):126–127, January, 2001.

14. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., and Qiu, L. The c. elegans genome sequencing project: a beginning. Nature 356(6364):37–41, March, 1992.

15. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y.-H. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor Miklos, G. L., Abril, J. F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M.,

Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. d., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidn-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z.-Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R.-F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. The genome sequence of drosophila melanogaster. Science 287(5461):2185–2195, March, 2000.

16. Waterston, R. H., Lander, E. S., and Sulston, J. E. On the sequencing of the human genome. Proc Natl Acad Sci U S A 99(6):3712–3716, March, 2002.

17. Morgan, M. J. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921, February, 2001.

18. Dacks, J. B. and Doolittle, W. F. Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. Cell 107(4):419–425, November, 2001.

19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. J Mol Biol 215(3):403–410, October, 1990.

20. Graves, J. A. M. A. Background and overview of comparative genomics. ILAR J 39(2-3):48–65, 1998.

21. Ghosh, S., May, M. J., and Kopp, E. B. Nf-kappa b and rel proteins: evolutionarily conserved mediators of immune responses. Annu Rev Immunol 16:225–260, 1998.

22. Siomi, H., Matunis, M. J., Michael, W. M., and Dreyfuss, G. The pre-mrna binding k protein contains a novel evolutionarily conserved motif. Nucleic Acids Res 21(5):1193–1198, March, 1993.

23. Li, W., Pio, F., Pawowski, K., and Godzik, A. Saturated blast: an automated multiple intermediate sequence search used to detect distant homology. Bioinformatics 16(12):1105–1110, December, 2000.

24. Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. Protein interaction mapping in c. elegans using proteins involved in vulval development. Science 287(5450):116–122, January, 2000.

25. Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. Genome Res 14(6):1107–1118, June, 2004.

26. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96(8):4285–4288, April, 1999.

27. Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. R., and Ideker, T. Pathblast: a tool for alignment of protein interaction networks. Nucleic Acids Res 32(Web Server issue), July, 2004.

28. Koyuturk, M., Grama, A., and Szpankowski, W. Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution. RECOMB 1, 2005.

29. Spirin, V. and Mirny, L. A. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A 100(21):12123–12128, October, 2003.

30. Bader, G. D. and Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4(1), January, 2003.

31. Przulj, N., Wigle, D. A., and Jurisica, I. Functional topology in a network of protein interactions. Bioinformatics 20(3):340–348, February, 2004.

32. Sharan, R., Ideker, T., Kelley, B. P., Shamir, R., and Karp, R. M. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, 282–289 (ACM Press, New York, NY, USA, 2004).

33. Vespignani, A. Evolution thinks modular. Nat Genet 35(2):118–119, 2003.

34. Brody, T. The interactive fly: gene networks, development and the internet. Trends Genet 15(8):333–334, August, 1999.

35. Berman, P. and Ramaiyer, V. Improved approximations for the steiner tree problem. In *SODA '92: Proceedings of the third annual ACM-SIAM symposium on Discrete algorithms*, 325–334 (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992).

36. Garg, N., Konjevod, G., and Ravi, R. A polylogarithmic approximation algorithm for the group steiner tree problem. J. Algorithms 37(1):66–84, October, 2000.

37. Kahng, A. and Robins, G. A new class of iterative steiner tree heuristics with good performance. Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on 11(7):893–902, 1992.

38. Garey, M. R. and Johnson, D. S. The rectilinear steiner tree problem in np complete. SIAM Journal of Applied Mathematics 32:826–834, 1977.

39. Kou, L., Markowsky, G., and Berman, L. A fast algorithm for steiner trees. Acta Informatica (Historical Archive) 15(2):141–145, June, 1981.

40. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. Introduction to Algorithms, Second Edition. The MIT Press, , September, 2001.

41. Simes, R. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1, 1986.

42. Goldberg, D. S. and Roth, F. P. Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci U S A 100(8):4372–4376, April, 2003.

43. Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. Modeling of protein interaction networks, , August, 2001.

44. Barabsi, A.-L. and Albert, R. Emergence of scaling in random networks. Science 286(5439):509–512, October, 1999.

45. Peeters, R. The maximum edge biclique problem is np-complete. Discrete Appl. Math. 131(3):651–654, September, 2003.

46. McInerney, J. O. and Wolfe, K. H. Genomic analysis methods. Microbiology Today 26(1):157–159, 1999.

47. Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18(7):1283–1292, July, 2001.

48. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. Dip: the database of interacting proteins. Nucleic Acids Res 28(1):289–291, January, 2000.

49. Drysdale, R. A. and Crosby, M. A. Flybase: genes and gene models. Nucleic Acids Res 33(Database issue), January, 2005.

50. Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R. K., and Botstein, D. Genetic and physical maps of saccharomyces cerevisiae. Nature 387(6632 Suppl):67–73, May, 1997.

51. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25(1):25–29, May, 2000.

52. Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. From the cover: Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A 102(6):1974–1979, February, 2005.

53. Hanks, S. K., Quinn, A. M., and Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241(4861):42–52, July, 1988.

Vita

Corban G. Rivera attended North Carolina State University between 1999 and 2003. During 2003, Corban studied for a semester at Lunds University in Sweden. Corban graduated first in his class from North Carolina State University with his Bachelor of Science in Computer Science and minor in Mathematics. From 2003 to 2005, Corban attended Virginia Polytechnic Institute and State University. He pursued a Master of Science in Computer Science with the Bioinformatics option.