PREDICTION OF PROTEIN-PROTEIN INTERACTIONS FROM SEQUENCE USING EVOLUTIONARY RELATIONS OF PROTEINS AND SPECIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF COMPUTER ENGINEERING DEPARTMENT
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TACETTİN DOĞACAN GÜNEY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2009

Approval of the thesis:

PREDICTION OF PROTEIN-PROTEIN INTERACTIONS FROM SEQUENCE USING EVOLUTIONARY RELATIONS OF PROTEINS AND SPECIES

submitted by **TACETTİN DOĞACAN GÜNEY** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Müslim Bozyiğit
Head of Department, **Computer Engineering** _____

Assist. Prof. Dr. Tolga Can
Supervisor, **Computer Engineering Department, METU** _____

Examining Committee Members:

Prof.Dr.Volkan Atalay
Computer Engineering Department, METU _____

Prof.Dr İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assist.Prof.Dr. Çağdaş Son
Biology Department, METU _____

Prof.Dr. Gerhard Wilhelm Weber
Institute of Applied Mathematics _____

Assist.Prof.Dr Tolga Can
Computer Engineering Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: TACETTİN DOĞACAN GÜNEY

Signature :

ABSTRACT

PREDICTION OF PROTEIN-PROTEIN INTERACTIONS FROM SEQUENCE USING EVOLUTIONARY RELATIONS OF PROTEINS AND SPECIES

Güney, Tacettin Doğacan
M.S., Department of Computer Engineering
Supervisor : Assist. Prof. Dr. Tolga Can

September 2009, 39 pages

Prediction of protein-protein interactions is an important part in understanding the biological processes in a living cell. There are completely sequenced organisms that do not yet have experimentally verified protein-protein interaction networks. For such organisms, we can not generally use a supervised method, where a portion of the protein-protein interaction network is used as training set. Furthermore, for newly-sequenced organisms, many other data sources, such as gene expression data and gene ontology annotations, that are used to identify protein-protein interaction networks may not be available. In this thesis work, our aim is to identify and cluster likely protein-protein interaction pairs using only sequence of proteins and evolutionary information. We use a protein's phylogenetic profile because the co-evolutionary pressure hypothesis suggests that proteins with similar phylogenetic profiles are likely to interact. We also divide phylogenetic profile into smaller profiles based on the evolutionary lines. These divided profiles are then used to score the similarity between all possible protein pairs. Since not all profile groups have the same number of elements, it is a difficult task to assess the similarity between such pairs. We show that many commonly used measures do not work well and that the end result greatly depends on the type of the similarity measure used. We also introduce a novel similarity measure. The resulting dense

putative interaction network contains many false-positive interactions, therefore we apply the Markov Clustering algorithm to cluster the protein-protein interaction network and filter out the weaker edges. The end result is a set of clusters where proteins within the clusters are likely to be functionally linked and to interact. While this method does not perform as well as supervised methods, it has the advantage of not requiring a training set and being able to work only using sequence data and evolutionary information. So it can be used as a first step in identifying protein-protein interactions in newly-sequenced organisms.

Keywords: phylogenetic profile, clustering, evolution, protein-protein interactions

ÖZ

PROTEİN PROTEİN ETKİLEŞİMLERİNİN SEKANS BİLGİSİNDEN PROTEİN VE TÜRLER ARASINDAKİ EVRİMSEL İLİŞKİLERİ KULLANARAK TAHMİN EDİLMESİ

Güney, Tacettin Doğacan
Yuksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Y. Doç. Dr. Tolga Can

September 2009, 39 sayfa

Protein-protein etkileşim tahmini, bir hücredeki biolojik süreçleri anlamanın en önemli adımlarından biridir. Pek çok sekansı bilinen organizmanın henüz deneylerle doğrulanmış protein-protein etkileşim ağları bulunmamaktadır. Bu tip organizmaların için, bir öğrenme verisine ihtiyaç duyan tipik makine öğrenme algoritmaları uygulanamaz. Ayrıca, yeni sekanslanmış organizmaların, genelde protein-protein etkileşimlerini tahmin etmek için kullanılan biyokimsay de bulunmayabilir. Bu tez çalışmasında, protein-protein etkileşim ikililerini sadece genomun sekans ve evrimsel bilgilerini kullanarak bulmak ve kümelemeye çalıştık. Evrimsel baskı hipotezinin benzer filogenetik profillere sahip proteinlerin etkileşim olasılığının yüksek olduğu önermesinden yararlanarak, tez çalışmasında proteinlerin filogenetik profillerini kullandık. Ayrıca bu filogenetik profilleri evrimsel çizgilere göre daha küçük gruplara böldük. Bu bölünmüş profilleri iki proteinin filogenetik olarak benzerliğine puan vermek için kullandık. Profilin içindeki gruplarda her zaman aynı sayıda eleman olmadığı için, bu grupların benzerliğine puan verebilecek yeni bir benzerlik fonksiyonu ürettik. Aynı zamanda sık kullanılan benzerlik fonksiyonlarının bu duruma uygun olmadığını gösterdik. Çıkan benzerlik puanlarını daha sonra Markov Kümeleme algoritmasına vererek sonuçları kümeledik. Sonuçta çıkan kümeler içinde kalan proteinlerin fonksiyonel olarak benzer özelliklere sahip olması ve

etkileşim içinde olması yüksek bir olasılıktır. Her ne kadar bu method diğer makine öğrenme algoritmaları kadar iyi çalışmasa da, öğrenme verisine ihtiyaç duymaması ve sadece sekans ve evrim bilgilerini kullanarak çalışmasının bir avantaj olduğu söylenebilir. Bu metod protein-protein etkileşimlerinin tanımlanmasında bir ilk adım olarak kullanılabilir.

Anahtar Kelimeler: filogenetik profil, kümeleme, evrim, protein-protein etkileşimleri

To my father

ACKNOWLEDGMENTS

I want to start by thanking my supervisor Asst. Prof. Dr. Tolga Can. His support and advice were invaluable during the course of this thesis.

I want to thank my friends (in no particular order), Furkan Kuru, Sertan Alkan, Enis Söztutar, Eren Halıcı, Gülhan Serhat, Mehmet Yılmaztürk and many others for their constant help and understanding.

I would also like to thank all the jury members for their insightful comments and suggestions on this thesis.

I would like to thank TÜBİTAK (Turkish Scientific and Technical Research Council) for the graduate scholarship during this study.

Finally, I would like to thank my family for all the encouragement they provided. Without them, this thesis would never be finished.

This thesis work is conducted as part of the TUBITAK Career Project # 106E128

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

Introduction

1.1 Problem Definition and Motivation

Protein-protein interactions (PPIs) are a vital part of biological processes in a cell, such as the structure of sub-cellular organelles, the transport machinery across the various biological membranes, packaging of chromatin, the network of sub-membrane filaments, muscle contraction, signal transduction and regulation of gene expression.

There are a number of methods to extract the protein-protein interaction network of an organism. Among these methods are Yeast Two Hybrid, Co-immunoprecipitation and Tandem affinity purification. Many of these methods are high-throughput methods, meaning that the existence (or non-existence) of many interactions can be determined at once. However, all of these experimental methods report a large number of false-positive and false-negative interactions, thus limiting their usage. Furthermore, there is usually a time delay between the sequencing of a new genome (of a new organism) and the appliance of a PPI detection method. So there are a large number of sequenced organisms out there that have been sequenced but without any information on its protein-protein interaction network.

Understanding the protein-protein interaction network of an organism enables us to understand the biological functions better. As mentioned above, protein-protein interactions play an important role in virtually all biological functions. Also aberrant protein-protein interactions are partly responsible for many diseases (such as cancer or Alzheimer's Disease). So, a better understanding of PPI networks can lead to better treatments.

Due to problems in experimental high-throughput PPI detection methods, algorithms have been developed to predict protein-protein interactions. However, most of these methods use

supervised learning methods, i.e., they take a ”gold standard” set of positive and negative interactions). But, the task of generating such sets is a difficult one due to inherent noise in every PPI detection method. Also, for newly-sequenced organisms such gold standard sets do not exist, limiting the usage of supervised algorithms.

In this thesis, our goal is to develop a unsupervised PPI prediction algorithm using only sequence based on co-evolution. For each protein of an organism, we generate its phylogenetic profile as a vector, then divide the resulting vector into smaller groups using a phylogenetic tree of the species used to construct its phylogenetic profile. So, similar organisms within the tree are grouped together. We then calculate the similarity between all protein pairs using the grouped vector and cluster the pairs. Because of co-evolution, the pairs more likely to interact will be clustered together while pairs more likely to not interact will also be clustered together. By examining the resulting clusters, valuable information can be gained about possible protein-protein interactions. We also show that the end result greatly depends on the type of distance measure used. We examine commonly used similarity measures and provided a novel similarity measure in this thesis.

1.2 Related Work

1.2.1 Supervised Methods for Protein-Protein Interaction Prediction

The Auto-Covariance (AC) method proposed by Guo *et al.* [7] generates a feature vector using several biochemical measures and the neighborhood of amino acids in a protein sequence by means of auto-covariance. A training set of positive and negative interaction sets are then used to train a Support Vector Machine (SVM). We used their positive and negative gold standard data sets to measure the success of our distance measures. Another method by Martin *et al.* [12] predicts the protein-protein interactions by training an SVM with signatures of protein pairs that are generated by encoding the variable length amino acids using their neighbours. Another technique proposed by Bock *et al.* [2] uses the primary structure of proteins together with the residual properties of amino acids such as charge, hydrophobicity and surface tension of a known database of protein interactions as training data for SVM. Other than being supervised methods, all of these methods use the physical and/or chemical properties of proteins and do not consider co-evolution.

1.2.2 Using Phylogenetic Profiles for Protein-Protein Interaction Prediction

The co-evolutionary pressure hypothesis suggests that during evolution two interacting proteins are likely to coevolve and interact in the evolved organism or disappear together, i.e, no orthologs of two proteins exist. Many researchers use this idea in protein-protein interaction prediction. The study by Jothi *et al.* [9] shows that proteins with similar phylogenetic profiles are likely to interact assuming that proteins in the same metabolic pathway or cellular system are co-inherited during evolution. Wu *et al.* [22] infer a confidence value between protein pairs based on the probability that a given arbitrary degree of similarity between two profiles would occur by chance, with no biological pressure. Pellegrini *et al.* [17] demonstrates that functions of protein-protein interactions can be detected by comparing the phylogenetic profiles and counting the numbers of bits changed (a basic form of similarity measure). They use binary phylogenetic profiles. While simple, this approach loses information as more fine-grained phylogenetic profiles can be generated. Bowers *et al.* [3] compute the probability of coevolution based on hypergeometric distribution. In other words, given two phylogenetic profiles they convert it into a probability value that represents their confidence on their co-evolution. They use this probability value in an integrative framework to derive functional association of proteins. Kim *et al.* [11] use a mutual information function based on the Shannon entropy to indicate the level of similarity between two phylogenetic profiles. Vert [21] developed a tree kernel that can use the phylogenetic tree of a genome along with the phylogenetic profiles of two proteins. This tree kernel is used to predict the functional class of a gene. Sato *et al.* [19] improve Pearson‘s correlation coefficient by using partial correlation coefficient on the matrix of Pearson‘s correlation coefficient values calculated between all proteins. Gonzales *et al.* [6] include the phenotype knowledge to phylogenetic profiles in order to extend the binary strings to continuous phenotypes and develop scoring functions to use them in pairs. Juan *et al.* [10] use an estimator of coevolution that takes the whole network of similarities between all of the pairs of proteins within a genome instead of relying on the individual tree similarity between two proteins, thereby also taking coevolutionary context into account. They use an iterative neighborhood extension method to construct links between protein pairs that are likely to interact. This is similar to a clustering approach but it is limited to only two iterative steps. We use a full clustering approach to cluster the protein pairs.

Another research in this area is the master's thesis work by Bahar Pamuk[15]. The author uses a number of clustering and filtering methods to reduce the dimensionality of phylogenetic profile vectors, then use SVM to predict the PPI networks.

1.3 Contributions

Our contributions in this thesis are:

1. We propose a new clustering approach where a distance value is calculated between all possible protein pairs in an organism. This distance value is defined by the similarity of the phylogenetic profiles of protein pairs.

2. We propose that by grouping similar organisms together in a phylogenetic profile, a more useful profile can be obtained. A grouped phylogenetic profile will feature less dimensions than a regular one. We show that when distance measures are used on high-dimensional phylogenetic profiles of protein pairs, results tend to converge, thus making protein-protein interaction prediction difficult.

3. We, furthermore, propose that by weighing the phylogenetic groups that are closer to the organism from an evolutionary perspective, the quality of clustering can be improved.

4. We propose a new distance measure that is more applicable than Pearson's correlation or Euclidean distance. We also show that the quality of the end result varies greatly depending on the distance measure employed.

1.4 Thesis Outline

This thesis is organized as follows: In Chapter 2, we provide the necessary background knowledge to understand the problem domain and the solutions. In Chapter 3, datasets are described and technical details of the proposed methods are given. In Chapter 4, experimental results which demonstrate the utility of the proposed methods are shown. In Chapter 5, the thesis is concluded with a summary and future directions.

CHAPTER 2

BACKGROUND

2.1 Amino Acids and Proteins

Proteins are essential parts of organisms and participate in virtually every process within cells. They are formed of linear chains of amino acids. A number of amino acids (as defined by a gene sequence) are joined by a peptide bond. An amino acid chain is commonly made of 200-300 amino acids.

The amino acid sequence of a protein can be represented by a string composed of letters each representing one of the 20 different kinds of amino acids. The amino acid sequence is not a complete representation as a protein's function or its interaction with other proteins are also defined by their structures (Figure 2.1), their physiochemical properties and locations in the living cell, among other things.

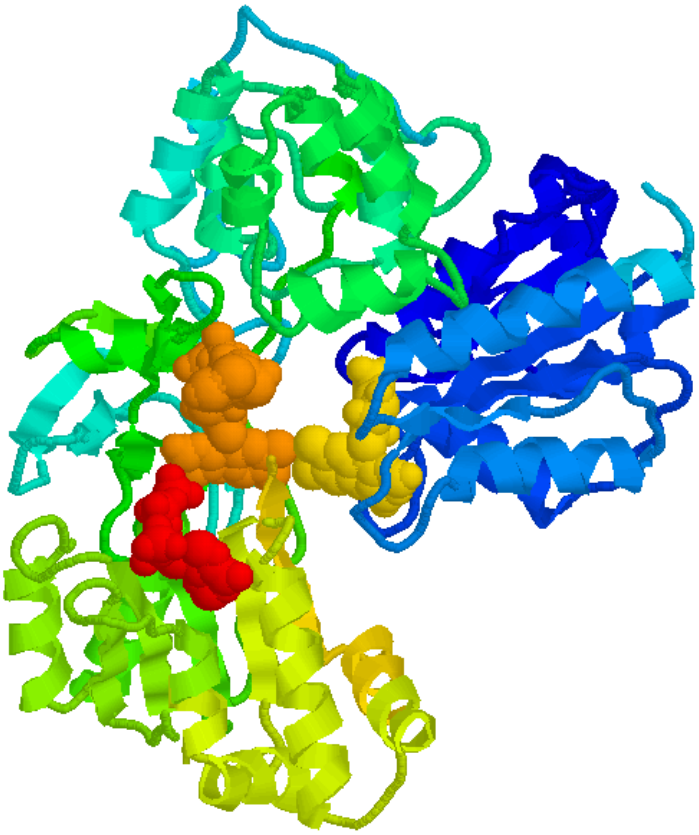

Figure 2.1: The complex 3D structure of a protein.

2.2 Pairwise and Multiple Protein Sequence Alignment

Making and alignment between the amino acid sequences of two proteins can give important insights into the functional relationship between them or help a researcher understand if two proteins are homologous or not. There are well-known algorithms for aligning a pair of proteins (pairwise alignment) based on dynamic programming. But generally, these approaches do not scale well to multiple protein sequence alignment. So for multiple alignment faster heuristics are used. These methods usually work by first aligning local sequences that are deemed more significant by the algorithm then using this alignment to align other proteins. Commonly used alignment tools are FASTA [16], BLAST [1], CLUSTAL W [20] and MUSCLE [5].

2.3 Protein-Protein Interaction Networks

Protein-protein interaction network is a graph (often undirected) that represent the interactions for all proteins in a genome (Figure 2.2). An edge between two protein nodes in the network indicate that the proteins interact in some way. Due to uncertainty in determining interactions between proteins, protein-protein interaction networks may be weighted where the weight of the edge indicate the confidence value for the interaction between proteins. Edges may also be labeled to indicate the type of interaction.

A common feature of protein-protein interaction networks is that they are very sparse; i.e. only a tiny portion of all possible edges is present in the network. For example, research by Hart et al. [8] indicate that a yeast's protein-protein interaction network is predicted to contain only 40 to 80 thousand interactions out of the possible 18 million edges.

Over the years, numerous organisms have been extensively researched and have publicly available high-quality protein-protein interaction networks available. The interactions in these organisms are usually validated by wet-lab experiments. There are also several high-quality databases of protein-protein interactions networks, such as MIPS and DIP. MIPS (Munich Information Center for Protein Sequences) [14] Mammalian Protein-Protein Interaction Database provide PPI networks for many mammals. All interactions in MIPS are confirmed by wet-lab experiments. DIP (Database of Interacting Proteins) [23] is another protein-protein interaction network database that catalogs experimentally determined interactions between proteins by using both wet-lab experiments and high-confidence PPI prediction methods.

2.4 Phylogenetic Profiles and Trees

Phylogenetic profile of a protein represents the existence or absence of the ortholog of a protein across a number of organisms. One way of generating a phylogenetic profile is to use a binary representation where 0 indicates absence and 1 indicating presence. In our thesis, BLAST is used to generate a protein's phylogenetic profile across 450 organisms. For every protein a BLAST search is performed and the negative log of E-value is used. E-value of the alignment of the query protein with a database protein indicates the similarity. We use a series of normalizations and quantizations on BLAST data before using it in our method. Methods

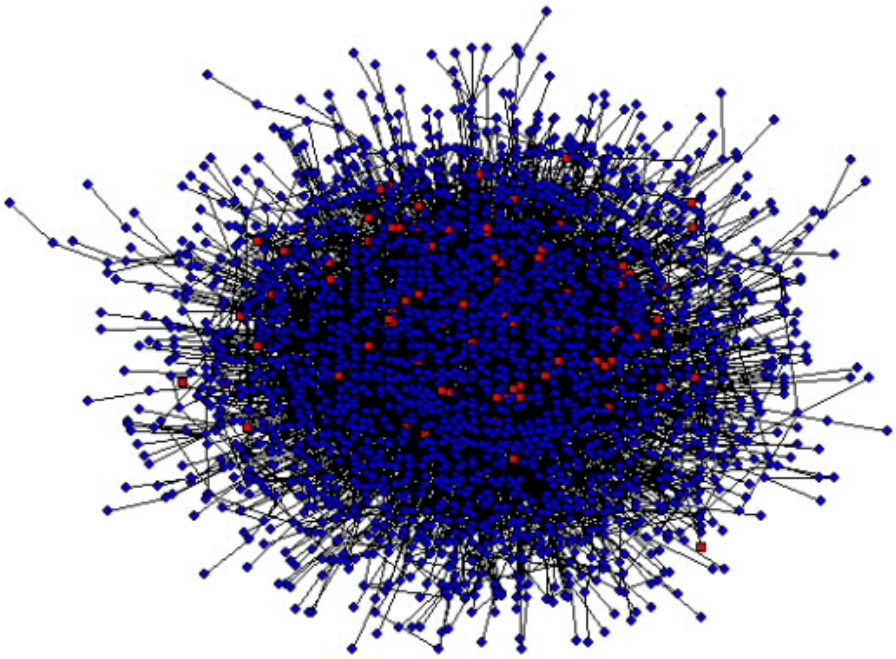

Figure 2.2: A Human Protein-Protein Interaction Network.

that generate binary phylogenetic profiles generally use BLAST E-values use 10^{-3} as a cutoff point.

Phylogenetic trees show the evolutionary links between organisms. In a phylogenetic tree, a parent node represents the common ancestor of the organisms in its child nodes. Observed organisms are in leaf nodes while the internal nodes can not be directly observed and thus are hypothetical groupings. A phylogenetic tree can also have a branch length indicating the evolutionary distance between nodes. In our dataset, evolutionary distances between taxonomy units are not present so we assume all branch lengths to be 1.

Phylogenetic trees are very useful for understanding and visualizing the evolutionary process.

Figure 2.3 shows a sample phylogenetic tree[1].

[1] http://nai.arc.nasa.gov/library/images/news_articles/big_274_3.jpg

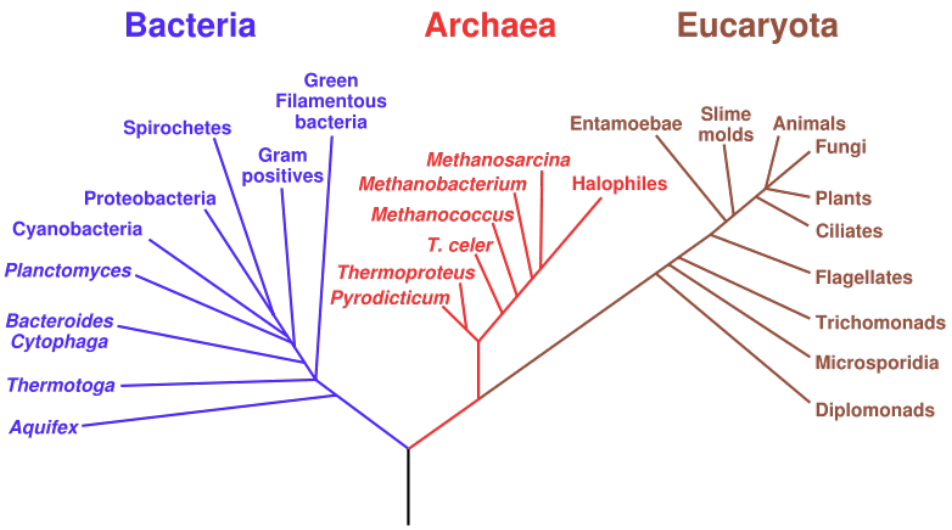

Figure 2.3: A sample phylogenetic tree.

2.5 Euclidean Distance

For two points (represented by n dimensional vectors **p** and **q**), the Euclidean distance between two points is defined as:

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

2.6 Pearson Product-Moment Correlation Coefficient

Pearson's Product-Moment Correlation Coefficient (sometimes referred to as Pearson's r, referred simply as Pearson's correlation in this thesis) is a measure to calculate the linear dependence between two variables. A value of +1 indicates that the two variables are perfectly linearly dependent on each other, i.e., as one increases the other also increases. A value of -1 indicates a negative correlation such that if the increase in one variable causes a decrease in the other. A value of 0 indicates that the two variables are linearly independent. Values in between show different levels of

positive or negative correlation. Pearson's correlation is defined as:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right) \quad ,$$

where

$$\bar{X} \text{ and } s_X$$

are the mean of X and the standard deviation of X, respectively.

Pearson's correlation values can be misleading if outliers are present. It is also undefined for cases where standard deviation of a variable is zero, even though biologically two profiles may be highly correlated.

2.7 Dijkstra's Shortest Path Algorithm

Finding the shortest path between all nodes in a graph is a common computer science problem. In our thesis, Dijkstra's shortest path algorithm is used to calculate the evolutionary distance between different nodes in a phylogenetic tree. The algorithm is defined as:

- Set all distances between all nodes to infinity.
- Mark all nodes as unvisited..
- Select an initial node and set its distance to 0.
- For all nodes still unvisited in graph, select the one with the smallest distance from the initial node:
    1. Consider all its unvisited neighbors and calculate their distance from the initial node.

2. If this distance is less than the distance stored in the edge, update the distance.
3. Mark the current node as visited.

Computing the evolutionary distance between groups allows us to assign weights to phylogenetic profile groups. From an evolutionary perspective, the existence or absence of a protein pair in organisms closer to each other in phylogenetic trees may be more relevant information than the existence of a protein pair between two organisms far from each other.

2.8 Markov Clustering

Clustering algorithms try to divide a graph to clusters so that generally, the number of edges between nodes in a cluster is high. Markov Clustering [4] is a fast graph clustering algorithm based on Markov chains. The basic idea behind Markov Clustering is that a random walk between nodes should frequently stay within the nodes of that cluster and seldom go to a different cluster. So between two arbitrary nodes of a cluster, we expect a large number of high-length paths (and thus, we expect the random walk to usually end within the cluster). One of the main advantages of Markov Clustering is its speed. Since we will be using our method on all protein pairs within an organism (for example, for Saccharomyces cerevisiae there are 18 million protein pairs) speed is an important consideration for us.

Markov Clustering takes a column stochastic matrix as its input. A column stochastic matrix is a non-negative matrix with the property that each of its columns sums to 1. The value at row i, column j corresponds to the distance between nodes i and j in original graph normalized so that the total distance of i to all its neighbors equal 1. This value is the probability of traversing from node i to node j.

Markov Clustering works in two main steps called expansion and inflation. In the expansion step, the algorithm simulates one step of random walk. This computes the probability of a random walker ending up on a node after one edge traversal. In the inflation step, the Hadamard power of the matrix is taken, followed by a scaling step

so that the matrix is column stochastic again. The inflation step severs the weak links between nodes, thus forming clusters.

The algorithm can be described as follows:

- Set M_1 to be the current matrix
- While *change* is smaller than a predefined value
    1. $M_2 = M_1 * M_1$
    2. $M_1 = Hadamard(M_2)$
    3. $M_1 = scale(M_1)$
    4. $change = M_1 - M_2$
- The components of M_1 are the resulting clusters

The algorithm may never converge (depending upon the matrix). However, in most cases, algorithm tends to converge after 3-10 iterations.

2.9 Gaussian Radial Basis Function

A radial basis function is a real-valued function whose value only depends on the distance from the origin. For example, Euclidean distance is a radial basis function.

Gaussian Radial Basis Function is a decaying radial basis function. It is very commonly used in Support Vector Machines as a kernel function. However, it can also be used as a way of assessing similarity. It is defined as:

$$\varphi(r) = e^{-\beta r^2} \quad ,$$

where β is the decaying factor and is positive and r is the distance.

2.10 Support Vector Machines

Support Vector Machines (SVMs) are not used in our proposed method but we make use of SVMs in some experiments so they are briefly described here.

Support Vector Machines are supervised learning methods used for classification and regression. Input data is represented by pairwise similarity values computed by a kernel function. Given two sets of data points, SVMs solve the problem of finding the best separating hyperplane such that the geometric margin is maximized. An example of a maximum margin hyperplane is given in Figure 2.4.

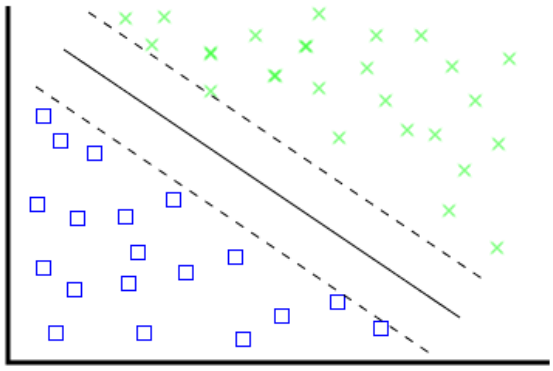

Figure 2.4: Maximum margin hyperplane that separates two data sets

SVMs use a kernel function to compare the data instances. A kernel function can be any function as long as it satisfies two properties: (a) The function must be symmetric (b) The function must be positive semi-definite. If the kernel function is the dot-product of two vectors then SVM is a linear classifier. However, using other kernel functions, such as Guassian Radial Basis Function, result in non-linear classifiers. Using a non-linear kernel function maps the data to a higher dimensional space (in the case of Gaussian RBF, to Hilbert space, i.e., infinite dimensions) such that the mapped data is linearly separable.

2.11 N-Fold Cross Validation

Cross validation is a technique for estimating the performance of a predictive model. In cross validation, data is partitioned into complementary subsets, where one subset is used for training and other subset is used for validation.

In *N*-fold cross-validation, the original sample is partitioned into *N* subsets. One subset is then selected as the validation set and the remaining *N* - 1 subsets are combined to produce the training set. This operation is repeated *N* times (the number of folds), each time a different subset is selected as the validation set. The resulting *N* success rates are then averaged to give the overall estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

CHAPTER 3

MATERIALS AND METHODS

3.1 Data Sets

In this section, we represent the data sets used in this thesis and the tools we use to measure the performance of our method. The protein-protein interaction network of baker's yeast (*Saccharomyces Cerevisiae*) is predicted in this thesis.

3.1.1 Phylogenetic Profiles

The phylogenetic profile of an organism is constructed via the BLAST tool described in Chapter 2. A BLAST search is performed for every protein (approximately 6000) in *Saccharomyces Cerevisiae*, where BLAST aligns the protein to its orthologs in 450 different organisms. BLAST E-values are in the range $[0, \infty]$. We use the negative log of BLAST E-values as a measure of similarity, i.e, smaller E-values are normalized to larger numbers. E-values smaller than 10^{-200} are converted to 200, i.e., 200 is the upper bound of similarity.

In many studies, two sequences with BLAST E-value of 10^{-3} are considered homologs. However, in the conversion method described above, two E-values 10^{-3} and 10^{-150} would be different by a factor of 50 even though they are close values in biological terms. Because of this, we perform an extra quantization step:

1. If the normalized E-value is less than or equal to 3, it is not quantized, i.e., taken as it is.

2. If the normalized E-value is between 3 (exclusive) and 50 (inclusive), it is taken to be 3.

3. If the normalized E-value is between 50 (exclusive) and 100 (inclusive), it is taken to be 4.

4. If the normalized E-value is between 100 (exclusive) and 150 (inclusive), it is taken to be 5.

5. If the normalized E-value is larger than 150, it is taken to be 6.

An example of quantization, for protein YBR160W, is given in Table 3.1.

Table 3.1: Quantization of a portion of a phylogenetic profile

Prot./Org.	hsa	ptr	mmu	rno	cfa	bta	ssc	gga	xla	xtrl
YBR160W (norm.)	106	101	105	105	103	106	48.5	105	102	102
YBR160W (quant.)	5	5	5	5	5	5	3	5	5	5

The organisms used to generate the phylogenetic profiles can be found in Appendix A.1.1.

3.1.2 Phylogenetic Tree

The phylogenetic tree is constructed from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. The file used can be found here[1]. The file format is simple, a line starting with '#' indicates a group and the number of '#'s indicate the level of the group. Every organism is written in a single line in the PIR-PSD[2] (International Protein Sequence Database) format.

However, the tree generated from KEGG database include only 438 of the 450 organisms from BLAST. So, the 12 organisms not in our tree is removed from the generated phylogenetic profile. The phylogenetic tree has a maximum depth of 5 and leaf nodes, i.e., organisms, are in levels 3 and 4. Overall, there are 107 groups, 49 in level 4 and 50 in level 3.

[1] ftp://ftp.genome.jp/pub/kegg/genes/taxonomy
[2] http://pir.georgetown.edu/pirwww/dbinfo/pir psd.shtml

Organisms used in the phylogenetic tree can be seen on Appendix A.1.2. Part of the phylogenetic tree can be seen on Appendix A.2.

3.1.3 Positive and Negative Interaction Sets

We use the datasets used by Guo et al. [7]. They define approximately 6000 protein-protein interaction pairs for positive and negative interactions as their 'gold standard' set. The positive gold standard set come from DIP core interaction dataset[3]. Since our method is unsupervised, we do not use this data in training (as Guo et al. do) but use it to test the performance of our similarity measures. Even without the clustering step, one can expect that the similarity measure should give high scores to protein-protein interaction pairs in the positive set, while giving low scores to the pairs in the negative set.

3.1.4 Gene Ontologies

The Gene Ontology[4] (GO) is a bioinformatics initiative that aims to annotate all genes and gene products across all species and provide tools to facilitate access too all aspects of the provided data.

We downloaded the most recent yeast GO annotations and the most recent ontology definition file as of date. We got a list of significant biological process terms from supplementary material of Myers et al. [13] and extracted the process terms. There were 295 significant terms in the supplementary data, 186 of which is annotated to at least one protein. Every protein in each of these 186 cluster has the same GO annotation, thus can be assumed to be in the same functional path.

GO annotations have many uses, for example, they can be used to predict protein-protein interactions. However, since GO annotations are not likely to be available for newly-sequenced organisms, we do not use GO annotations for PPI prediction in our thesis. Still, our test organism (baker's yeast) is well-studied and well-annotated.

[3] http://dip.doe-mbi.ucla.edu/dip/Main.cgi
[4] http://www.geneontology.org/

Because proteins in the same functional path are likely to evolve together, they are also likely to be phylogenetically similar to one another. So after the clustering step, the resulting clusters may be expected to map to a GO annotated cluster.

Even for well-studied organisms such as baker's yeast, a significant portion of the PPI network is not yet validated by wet-lab experiments. Because of this, for a given GO cluster, it is possible that there are other proteins that are in the same functional path but not yet GO-annotated. This tells us that, after the clustering step, the resulting clusters may contain apparent false-positives that should actually be in that GO cluster.

3.2 Overview

This section describes the overview of our method while the following sections detail the steps in our algorithm.

1. Construct phylogenetic profiles for all proteins in the test organism using BLAST.
2. Normalize, then quantize the BLAST *E*-values
3. Construct the phylogenetic tree for organisms in the phylogenetic profiles.
4. Using the phylogenetic tree, group the phylogenetic profiles into taxonomy units.
5. For all protein pairs, calculate the similarity value for all groups.
6. Average the similarity vector using evolutionary distance information from the phylogenetic tree and reduce the similarity vector to a single value.
7. Cluster the resulting network

3.3 Phylogenetic Profile Grouping

Based on the phylogenetic tree, the generated phylogenetic profiles are divided into smaller groups. To achieve this, we take all the phylogenetic groups one level above

the organisms and group the profile accordingly. For example, for the group Fishes (Eukaryotes → Vertebrates → Animals → Fishes), the profile values for organisms *Danio rerio* (zebrafish), *Fugu rubripes* (Japanese puffer fish), *Tetraodon nigroviridis* (green spotted puffer) and *Oryzias latipes* (Japanese medaka) are put in a group. A portion of a sample profile is given in Figure 3.1.

	Mammals							Birds	Amphibians	
	hsa	ptr	mmu	rno	cfa	bta	ssc	gga	xla	xtr
YBR160W	106.0	101.0	105.0	105.0	103.0	106.0	48.52287875	105.0	102.0	102.0

Figure 3.1: A portion of the grouped phylogenetic profile for protein YBR040W.

3.4 Assessing the Similarity Between Profiles

After dividing the phylogenetic profiles into smaller groups, we then need a similarity measure that can score how biologically related the aforementioned groups are. There are a couple of methods we can use.

3.4.1 Pearson's Correlation

A common method for measuring correlation between vectors is Pearson's Correlation. There are some problems with using Pearson's Correlation in comparing phylogenetic profiles. One problem is that Pearson's Correlation measures the level of dependence between two vectors. This score will not always necessarily be correlated with how similar the profiles are from a biological perspective. Another problem is that if Pearson's Correlation is run on the original 450 dimensional vector, many protein pairs (both actually positive and negative) will get similarity scores due to the high number of dimensions. We aim to mitigiate this for Pearson's Correlation (and other similarity measures) by dividing phylogenetic profile to smaller groups.

Pearson's Correlation also does not work well with a grouped phylogenetic profile. Some of our phylogenetic groups (for example, Eukaryotes → Fungi → Ascomyetes → Fission Yeasts) only include one organism (for Fission Yeasts, *Schizosaccharomyces pombe*), and Pearson's Correlation is undefined for a vector with a single dimension. In our tests, we ignored single-organism groups which means important

evolutionary data is lost.

Also consider these two profile portions:

Table 3.2: Profile portions for YOR275C and YPR173C

	Budding Yeasts				
Protein	dsmi	dsba	ago	dkwa	cal
YOR275C	200.0	200.0	119.0	145.0	71.2
YPR173C	67.3	200.0	200.0	200.0	175.0

Both proteins have highly similar orthologs for all organisms in Budding Yeasts and a good similarity measure should give this pair a large score. However, calculating Pearson's correlation for these two profile portions gives:

$$Pearson's\ Correlation(\begin{bmatrix} 200.0 \\ 200.0 \\ 119.0 \\ 145.0 \\ 71.2 \end{bmatrix}, \begin{bmatrix} 67.3 \\ 200.0 \\ 200.0 \\ 200.0 \\ 175.0 \end{bmatrix}) = -0.40 \ ,$$

implying that the profile portions are negatively-correlated. Because Pearson's Correlation works by first centering the vectors on their respective means (means are 147.0 and 168.4 respectively). After the vectors are centered on their means, first vector has 2 negative, 3 positive values and the second vector has 1 negative, 4 positive values. Since Pearson's Correlation then multiplies the dimensions on mean-centered vectors we end up with a negative value.

Another problem is especially apparent after the quantization step in phylogenetic profiles. Among positive interaction pairs, there are many high-similarity values for many orthologs, but after quantization these values are scaled to a smaller range. After the quantization step, many groups in the phylogenetic profile of proteins consist of a single value. But since Pearson's Correlation measures the correlation of change between two vectors, single repeating value profile pairs can not be calculated. An example is given below in Table 3.3.

Pearson's Correlation of these two profile portions will be undefined even though

Table 3.3: An example of quantization for profiles

	Mammals						
Protein/Organism	hsa	ptr	mmu	rno	cfa	bta	ssc
YHL004W	16.5	16.6	17.0	18.0	17.6	18.3	7.6
YDR036C	44.5	11.3	45.0	39.0	43.3	40.1	10.2
YHL004W (quant.)	3	3	3	3	3	3	3
YDR036C (quant.)	3	3	3	3	3	3	3

from a biological perspective it looks like these two profile portions should be highly correlated. We may try to extend the Pearson's Correlation to assign such profiles a value of 1 but that approach would be problematic, as correlation of two profiles, one made of 3's, the other made of 0's will also be undefined and it is not immediately clear how such profiles should be scored.

3.5 Normalized Euclidean Distance

Because of the problems with Pearson's Correlation, we may consider other similarity measures. One of the possible measures is the Euclidean distance. Euclidean distance has some significant advantages over Pearson's Correlation:

1. It works for single element profile portions.

2. It works for profile portions where standard deviation is zero, i.e., profile portion is made of the same repeating number.

3. It is much more resistant to small changes in vectors.

However, Euclidean distance also has its share of problems. Distances between high-dimensional vectors tend to be higher than low-dimensional vectors. So if we were to use Euclidean distance directly, phylogenetic profile groups with many members would dominate the profile and dominate the overall results.

So, we propose a simple change to Euclidean distance:

$$\sqrt{\frac{\sum_{i=1}^{n}(p_i - q_i)^2}{n}}$$

Another problem is that Euclidean Distance is a measure of distance and not of similarity, i.e., if two profiles are 'closer' to one another, normalized Euclidean Distance value will be small, while for unrelated profiles it will be large. We can get a similarity measure by simply dividing 1 to the normalized Euclidean distance. Alternatively, we use the Gaussian Radial Basis Function:

$$\frac{1}{e^{\beta \frac{\sum_{i=1}^{n}(p_i - q_i)^2}{n}}}$$

where β is the decaying factor and larger than 0.

A Gaussian Radial Basis function is stronger than inverting the result as Gaussian RBF decays to zero faster (depending on β). Thus, only very similar profiles get a high score.

3.6 Averaging Grouped Phylogenetic Profile Similarity Vectors

After the initial BLAST query, we have 450-dimensional profile vectors. We then use the information in the phylogenetic tree to divide the 450-dimensional vector into smaller groups (for example, 47 groups if we take groups to be the taxonomy units that are parents of leaf nodes). We then apply one of the similarity measure described in the previous section to the protein pairs and we get a low-dimensional (continuing the example, 47-dimensional) vector that describes the similarity between two protein phylogenetic profiles within groups. We can use this vector directly to predict protein-protein interactions. But the Markov Clustering algorithm that we used in our work expects a single scalar value as the edge length between nodes. Because of this, we need to reduce the new similarity vector to a single dimension.

One approach is to average the values in the similarity vector. While simple, this approach ignores the evolutionary distance between different organisms. Profile similarities that are coming from organisms similar to our test organism from an evolutionary perspective are more likely to have an effect on overall profile scoring than those organisms that are not related evolutionarily. Because of this, we propose another averaging method where the evolutionary distance is weighted accordingly. We assume the phylogenetic tree to be an undirected graph and calculate the shortest distance between all nodes (both organisms and other taxonomy units). After that, Gaussian Radial Basis Function is applied to every group.

$$\frac{1}{e^{\beta r^2}}$$

where β is the decaying factor and r is the shortest distance between test organism, i.e., yeast, and the organism in the profile.

3.7 Scaling and Clustering

After averaging is complete, we now have a single scalar value that represents the similarity between two phylogenetic profiles. These scalar values are calculated for every possible protein-protein pair in the network, thus forming a complete clique with 6000 nodes. The resulting k-clique is converted to matrix form and scaled so that the sum of values in a column add up to 1 as Markov Clustering expects a column-stochastic matrix. The scaled matrix is then given to Markov Clustering.

Since proteins in a functional path have a high probability of co-evolving, clustering approach enables us to consider all proteins in the functional path together, instead of just focusing on the probability of interaction between two proteins. These proteins in functional paths will likely have a high intra-cluster distance so Markov Clustering is a good approach for clustering them.

CHAPTER 4

RESULTS

In this chapter, we represent the experimental results of our method.

4.1 Quantization and Dimension Reduction

As discussed in Chapter 3, after taking the negative log values of BLAST E-values, we perform an extra quantization where we reduce the range of values. For our experiments, we also define a simple binarization of BLAST E-values (as used by other researches) where BLAST E-values smaller than or equal to 10^{-3} are taken to be 1 and other values are taken to be 0.

Grouping phylogenetic profiles can be considered a way of dimension reduction. We start with a 450 dimensional vector where every dimension represents an *organism* and values represent the similarity between the current protein and protein in the represented organism. We then reduce this phylogenetic profile to a lower dimensional vector by grouping. In the grouped phylogenetic profiles, every dimension represents a *phylogenetic group* and values represent the overall similarity between the current protein and all the proteins of organisms in that phylogenetic group. For the next steps of our algorithm to work, i.e., the averaging and the clustering steps, this dimensionality reduction needs to be meaningful. Even though our method is unsupervised, we use classification to show that dimensionality reduction does not lose important information.

We use the dataset described in Section 3.1.3. We employed Support Vector Machines

(SVM) as the classification algorithm. We used the widely-used libsvm[1] library as the Support Vector Machines implementation. The libsvm tool is used with the default settings. In its default mode, libsvm tool uses Gaussian Radial Basis Function as its kernel method. We used 5-fold cross validation to measure the accuracy of classification.

We also need a 450 dimensional representation of the similarity between two profiles. To that end, we define a simple similarity algorithm that takes two 450 dimensional vectors as its input and returns a 450 dimensional similarity vector as its output:

- Convert both profile vectors to binary vectors using the binarization method described above.
- For dimension i
    1. If the i^{th} dimension of both profiles are 1, then set dimension i to 1.
    2. If the i^{th} dimension of both profiles are both 0, then set dimension i to 0.5.
    3. If the i^{th} dimension of the profiles are different values, then set dimension i to 0.

This method is similar to XOR-ing the profiles except if both dimensions are 0, the resulting value is 0.5 instead of 1.0.

Accuracies for different methods are given in Table 4.1. Grouped Euclidean represents the similarity measure that we propose in this thesis (the similarity measure described in Section 3.5 used with grouped phylogenetic profiles). Grouped Pearson's Correlation represents the similarity measure where Pearson's Correlation is applied to every group for protein pairs. In both grouped methods, we assumed the parents of leaf nodes to be the groups. This gives us 47 groups. For completeness, we also use regular Euclidean distance (again, Gaussian Radial Basis Function is applied to Euclidean distance to convert it from a distance measure to a similarity measure) and Pearson's Correlation between two profiles. In these two methods, Euclidean distance

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

or Pearson's Correlation is calculated between two full profiles without using phylogenetic trees. Finally, Modified XOR represents the similarity measure described above.

For the descriptions in between parantheses, "(normalized)" indicates that negative logs of BLAST E-values are taken. Label "(quantized)" indicates that the quantization step is performed. Label "(binary)" indicates that the simple binarization scheme described above is used.

For consistency, every dimension of every vector (for all methods) given to libsvm is scaled between [0.0, 1.0].

Table 4.1: Classification Results.

Method	Accuracy
Grouped Euclidean (normalized)	63.5555%
Grouped Euclidean (quantized)	64.7532%
Grouped Euclidean (binary)	64.103%
Grouped Pearson's Correlation (normalized)	50.9111%
Grouped Pearson's Correlation (quantized)	51.125%
Grouped Pearson's Correlation (binary)	50.2438%
Regular Euclidean (normalized)	50.1754%
Regular Euclidean (quantized)	54.6326%
Regular Euclidean (binary)	50.1668%
Regular Pearson's Correlation (normalized)	54.7096%
Regular Pearson's Correlation (quantized)	61.1087%
Regular Pearson's Correlation (binary)	63.3929%
Modified XOR	66.3701%

Generally, for all methods, normalized forms perform worse than quantized and binarized forms. Also, except for regular Pearson's Correlation, quantized form outperforms binary form.

Grouped Pearson's Correlation performs no better than a coin toss, thus is not very useful. This is expected because of the problems mentioned in Section 3.4.1. Regular Euclidean also perform no better than a coin toss. The so-called 'curse of dimensionality' applies here where the Euclidean distance between two high dimensional (450 dimensions in this case) for many different vectors are similarly large values and so, the Gaussian Radial Basis function converts all values to very small values.

Grouped Euclidean similarity (in quantized form) performs very close to Modified XOR method. This shows that the lower-dimensionality of a grouped phylogenetic profile does not sacrifice important information if the right similarity measure is used for comparison.

For the above table, the decaying factor β in Gaussian Radial Basis function for grouped Euclidean is taken to be 1.0. However, different values of β have little effect on the accuracy of classification, as can be seen in Figure 4.1. Even for much larger beta values (such as 10 or 20), accuracy remains largely the same.

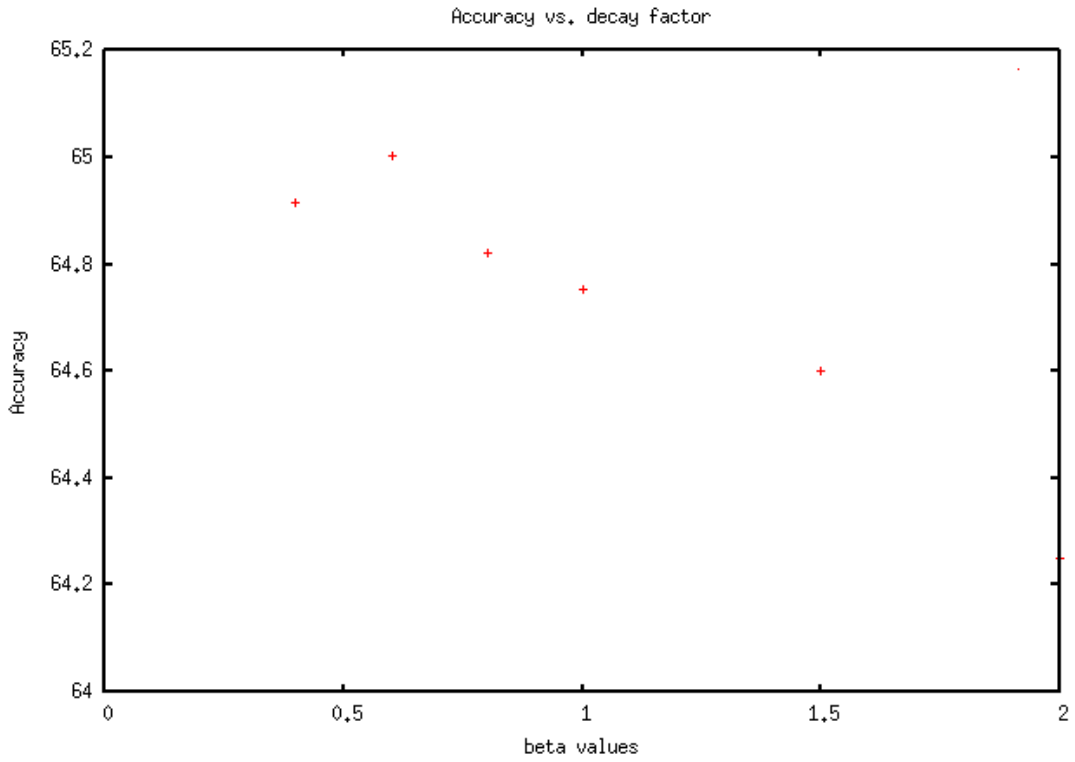

Figure 4.1: Accuracy of classification for various β.

We also expect that the similarity vectors for positive and negative protein-protein interactions to be distant from each other. So, for both the positive and the negative set, the mean of each dimension is calculated. This corresponds to computing the mean of every group. The results are given in Figure 4.2.

The group ids correspond to the group given in Appendix A.2.1. For example, group id 1 correspond to Acidobacteria. For almost all groups, means of positive interaction values are consistently higher than means of negative interaction values.

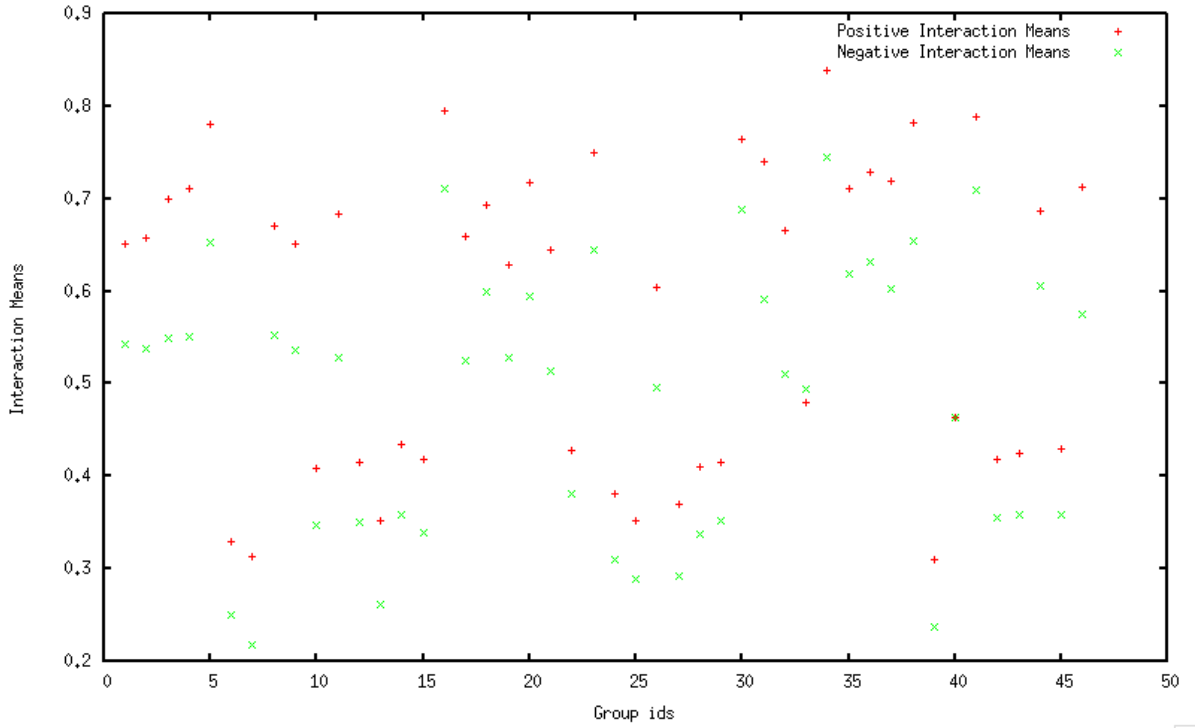

Figure 4.2: Means of Interaction values for different groups.

4.2 Clustering

After the previous step, 450-dimensional phylogenetic profitles were reduced to smaller dimensional vectors. However, for the clustering algorithm, this value needs to be reduced to a scalar. To achieve this, we can simply average the correlations or use a weighted average method using evolutionary distance as described in Section 3.6. We call the first averaging method 'simple' and the second 'tw' (short for tree-weighted).

During our tests, we saw that all methods, save for one, fail to produce meaningful results. For most of the methods, the Markov Clustering algorithm simply returned 1 cluster containing all proteins. We can create more clusters by experimenting with Markov Clustering or method parameters to extreme ranges but these will not lead to good results so are skipped. The one method that generates useful clusters is quantized grouped Euclidean similarity averaged using the tree-weight method.

Markov Clustering algorithm is mainly controlled by the inflation parameter. This parameter indicates the level of aggressiveness in severing links between nodes. The

default value is 2.0. We also use the most aggressive value in the recommended range (4.0) to sever all the weak links that are likely to be false-positives. We also tried the value 8.0 for experiments. The number of resulting clusters for different values of β (decaying factor for Gaussian RBF) and inflation parameters is given below in Table 4.2.

Table 4.2: Number of clusters for different values of *inflation* ($\beta = 1$).

Inflation	Number of Clusters
2.0	8
4.0	16
8.0	26

The number of resulting clusters for different values of β is given in Table 4.3.

Table 4.3: Number of clusters for different values of β (*inflation* = 1).

Decaying Factor	Number of Clusters
0.5	2
1.0	8
10.0	168
20.0	247

As expected, increasing decaying factor or inflation increases the number of clusters.

To test the performance of clustering, we use the DIP dataset which includes approximately 18000 protein protein interaction pairs where all interactions are obtained by high-throughput experiments. This dataset only contains a portion of the possible 40 to 80 thousand interaction pairs, so it is probable that some of the protein pairs in clusters interact in reality but is missed because it is not in the dataset. So we only calculate the ratio of clustered protein pairs to all protein pairs in dataset, i.e., recall. We assume that two proteins are interacting if they are in the same cluster and that they are not interacting if they are in different clusters.

Unfortunately, clustering results are pretty bad and are in fact, *worse* than a coin flip as shown in Table 4.4:

Table 4.4: Recall results.

Inflation/β	Recall
I=4.0,β=20.0	8%
I=8.0,β=10.0	8%
I=8.0,β=10.0	9%
I=8.0,β=1.0	25%
I=4.0,β=1.0	29%
I=2.0,β=1.0	45%
I=2.0,β=0.5	91%

Due to high-number of false-positives, Markov Clustering does not perform well when used with default inputs and generates very few clusters. If we force a larger number of clusters with high inflation and β values, then recall is very low. With default values, while recall is naturally high, the precision is extremely low. However, Markov Clustering can still be a useful first step in identifying groups of proteins that are likely and not likely to interact. For example, with inflation 2.0 and β 0.5, there are two clusters, first one with 5237 and the second with 626 members. If, for the moment, we were to assume that the approximately 18000 interactions were the *only* interactions for yeast (which is, of course, not true, but useful for a thought experiment) and since there are 18 million possible interactions, we can say that, for every 1 real interaction there are 1000 possible interactions. In other words, real to possible interaction ratio is 1/1000. For the second cluster (with 626 members) there are 195938 possible interactions. However, from DIP dataset, only 37 interactions exist within the second cluster. So real to possible interaction ratio is about 1/5000. So, Markov Clustering was successful in dividing the network into two groups, one likely to interact and the other unlikely to interact.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusions

Protein-protein interaction prediction remains an interesting problem in Bioinformatics. In this thesis, we set out to develop novel methods for protein-protein interaction prediction using only the evolutionary information and sequence of proteins in an unsupervised framework. We used the phylogenetic profiles of proteins as our main data source and used the knowledge of co-evolutionary pressure hypothesis to predict interactions based on the similarity of profiles. We showed that preprocessing of phylogenetic profile vector (by normalization, quantization, etc.) can lead to significantly different outcomes. By going beyond a simple binary phylogenetic vector and including the BLAST E-values in a carefully scaled form led to better accuracy.

We also proposed a way to group phylogenetic profiles based on phylogenetic trees. We identified the shortcomings of various other similarity measures and introduced a new robust similarity measure to compute the similarity of group pairs for proteins. We used Support Vector Machines and 5-fold cross validation to show that this grouping step results in a significant dimensional reduction without the loss of important information as SVM performance for high-dimensional and lower-dimensional vectors were mostly identical.

We applied a weighted average method to reduce the lower-dimensional similarity vectors to 1 dimension. This weighted average method made use of phylogenetic trees and the concept of evolutionary distance. Tree-weighted method fared better

compared to simple averaging when they are used in clustering.

We made use of Markov Clustering to cluster likely interaction pairs together. While this approach did not perform very well, it can still be used as an initial step in identifying the protein-protein interaction pairs in a new organism.

5.2 Future work

An important extension would be to extend the Markov Clustering tool so that it also returns the resulting edge probabilies instead of just clusters. Alternatively, another clustering algorithm, such as k-means clustering, self-organizing maps or hierarchical clustering, may be used.

While the co-evolutionary pressure hypothesis suggests that interacting proteins are likely to evolve or disappear together, it does not take into account the random possibility of two non-interacting proteins that happen to evolve together. Two proteins evolving together will always have a high similarity value even though they may not be interacting. Since the ratio of the number of total protein-protein interactions to the number of total possible protein-protein interactions is tiny, there may be a huge number of such false-positive interactios. So, a false-positive elimination algorithm may work before the clustering step, eliminating the easily-noticeable false-positives. This will result in a less dense graph, thus making clustering easier. One such algorithm is subcellular localization prediction. The idea behind subcellular localization is that interacting proteins are likely to be localized in the same part of a living cell. Following this line of thought, proteins in different localizations are not likely to interact. Subcellular localization prediction (SLP) algorithms predict the likely localization of a protein. For example, the BaCelLo algorithm proposed by Pierleoni et al.[18] do SLP by using protein sequence data (but they use a supervised method). One of the downsides is that, there may be interacting proteins even though they are in different subcellular localization so SLP can predict some false-negative interactions. Another option may be Gene Ontology. For a newly-sequenced organism, Gene Ontology data will probably not be available. But, since, Gene Ontology annotates proteins according to their functions (among other things), such annotated proteins are likely to

evolve together. So ontology clusters from a well-studied organism may be mapped to the proteins in the test organism to identify likely interacting protein pairs in the same functional path. The weights of such edges can then be increased to emphasize the importance of these protein homologs.

REFERENCES

[1] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] J. R. Bock and D. A. Gough. Predicting protein protein interactions from primary structure. *Bioinformatics*, 17:455–460, 2001.

[3] P. M. Bowers, M. Pellegrini, M. J. Thompson, J. Fierro, T. O. Yeates, and D. Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, 5:R35, 2004.

[4] S. V. Dongen. A cluster algorithm for graphs. *Information Systems*, 10:1–40, 2000.

[5] R. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *Bioinformatics*, 5:113, 2004.

[6] O. Gonzales and R. Zimmer. Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. *Bioinformatics*, 24:1257–1263, 2008.

[7] Y. Guo, L. Yu, Z. Wen, and M. Li. Using support vector machine combined with auto covariance to predict protein protein interactions from protein sequences. *Nucleic Acids Research*, 36:3025–3030, 2008.

[8] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006.

[9] R. Jothi, T. M. Przytycka, and L. Aravind. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, 8:173, 2007.

[10] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome wide coevolutionary networks. *PNAS*, 105:934–939, 2008.

[11] Y. Kim and S. Subramaniam. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins*, 62:1115–1124, 2006.

[12] S. Martin, D. Roe, and J.-L. Faulon. Predicting protein protein interactions using signature products. *Bioinformatics*, 21:218–226, 2005.

[13] C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7:187, 2006.

[14] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. StÃ$\frac{1}{4}$mpflen, H.-W. Mewes, A. Ruepp, and D. Frishman. Mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[15] B. Pamuk. Coevolution based prediction of protein-protein interactions with reduced training data. Master's thesis, Computer Engineering Department, Middle East Technical University, 2008.

[16] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proceedings of National Academy of Sciences USA*, 85:2444, 1988.

[17] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *National Academy of Sciences of the United States of America*, 96:4285–4288, 1999.

[18] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio. Bacello: a balanced subcellular localization predictor. *Bioinformatics*, 22(14):408–416, 2006.

[19] T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh. Prediction of protein protein interactions based on real valued phylogenetic profiles using partial correlation coefficient. *GIW*, page 122, 2004.

[20] J. D. Thompson, G. D. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4608, 1994.

[21] J.-P. Vert. A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18:276–284, 2002.

[22] J. Wu, S. Kasif, and C. DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19:1524–1530, 2003.

[23] I. Xenarios, L. Salwinski, X. J. Duan, K. Higney, P., S.-M., and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002.

APPENDIX A

APPENDIX A

A.1 ORGANISMS

A.1.1 Organisms Used for the Construction of Phylogenetic Profiles

The short names of the 450 organisms used to generate the phylogenetic profiles is given below: bqu cfa eba sme rxy dpyo tan sma dcbr nme reh cyb cya nma jan tac dme cel ttj tth tte lpp cef lpn dbmo lpl vpa pho mxa xla bps ilo lpf hdu bpn bpm rde pha bpe bpa gox cdi aha rco osa hwa pgi ccr ddha ago hch mca erw dkwa eru cch wbr pfu wbm cca pfo cvi pfl erg mbo afu cbu rbe dtni pfa rba afm ayw sit sil lmo pen bms rty lmf lme yps tpv lma gme ypn mtu ypm ypk hal bmf bme mtp mac bmb ype bma cal aeh hac ypa ctr mth shm tpa cac cab mtc wsu she blo cte sha llc ctc dyli lla cta rsp nha msu bli rso bld hso pcu ade sgl pcr ngo ooe hsa rru csa xft pca gka sfx sfv dge sfu dncr sfr aci ace crp syw ljo sfl xfa uur syn dfru ava syg nfa syf sye syd abo syc bja ser neu sep pub net pau pat neq par lin tma aba lil dsba det dcnb sec pai lic ptr rpr pae pto mpu pac fal pab hpy sdy deh ftu mpn atu cpv cpt rpe rpd cps ftl aae sdn cpr rpc rpb rpa mpe hpj fth cpn ftf mpa cpj ath sde swo eli hpa cpf pst cpe atc psp cpa wol xcv ddi nwi tko bhe dde bha sco psb xcc xcb lga rno dvu gga bga art vfi sbo sus bfs hne bfr fra cne mmy bfl mmu mmr mmp sav mmo sau sat sas sar cmu rme sao san sty nar sam xtr oih sal sak lxx sai stt mmc xac mma sag sto sac dar ppu hma sab stm stl cme saa ppr sth ste stc mlo ppe dcin rle bxe mle dame poy aph ssp sso ehi ape ssn dspd ldb dsy ddpo pol ssc lwe vvy nse vvu aor dkla vch bcz sru mka dpkn bcn dmgr tfu bcl lca bci bch fnu bce bcc gvi dcgr bca lbu lbr cjr dre lbl bbu mja zmo lbj cjk ani dra bbr pmu pmt cje ana gbe pmn ter pmm efa bur pmi tel hit bba spz hin pma spy buc

spt sps bat spr bas bar dsmi spo spn plu plt spm ban mhy lac spk bam ama spj spi sph spg baf dps spd nph tdn spb spa chy bab baa btk chu tws bth tde rha bte cho xoo bta xom hhe twh son tcx gsu bsu tcr mgm noc ecv ecu tvo ecs mge ecp eco ecn cgl mga ecj eci ech ece rfr cgb ecc tbr eca mfl rfe tbd smu daga nmu cfe reu ret

A.1.2 Organisms Used for the Phylogenetic Tree

The short names of the organisms used in the phylogenetic tree are given below: hsa ptr mmu rno cfa bta ssc gga xla xtr dre dfru dtni dme dbmo cel ath osa cme dsmi dsba ago dkwa cal spo ani afm aor cne ecu ddi pfa dpkn tan tpv cpv cho tbr tcr lma ehi eco ecj ece ecs ecc eci ecp ecv sty stt spt sec stm ype ypk ypm ypa ypn yps sfl sfx sfv ssn sbo sdy eca plu buc bas bab bcc wbr sgl bfl bpn hin hit hdu hso pmu msu xfa xft xcc xcb xcv xac xoo xom vch vvu vvy vpa vfi ppr pae pau ppu pst psb psp pfl pfo pen par pcr aci son sdn sfr she shm ilo cps pha pat sde cbu lpn lpf lpp mca ftu ftf ftl fth tcx noc aeh hch csa abo aha bci crp nme nma ngo cvi rso reu reh rme bma bxe bur bcn bch bam bps bpm bte bpe bpa bbr rfr pol neu net nmu eba dar tbd hpy hpj hpa hhe hac wsu tdn cje cjr gsu gme pca dvu dde bba dps ade mxa sat sfu rpr rty rco rfe rbe wol wbm ama aph eru erw erg ecn ech nse pub mlo sme atu atc ret rle bme bmf bmb bms bja rpa rpb rpc rpd rpe nwi nha bhe bqu ccr sil sit rsp jan rde mmr hne zmo nar sal eli gox gbe rru mgm aba sus bsu bha ban bar baa bat bce bca bcz btk bli bld bcl oih gka sau sav sam sar sas sac sab saa sao sep ser sha ssp lmo lmf lin lwe lla llc spy spz spm spg sps sph spi spj spk spa spb spn spr spd sag san sak smu stc stl ste lpl ljo lac ldb lbu lbr lca lga ppe efa ooe lme cac cpe cpf cpr ctc sth swo chy dsy tte mge mpn mpu mpe mga mmy mmo mhy uur poy ayw mfl mtu mtc mbo mle mpa mmc cgl cgb cef cdi cjk nfa rha sco sma twh tws lxx art pac tfu fra fal ace blo rxy fnu rba ctr cta cmu cpn cpa cpj cpt cca cab cfe pcu bbu bga baf tpa tde lil lic lbj lbl syn syw syc syf syd sye syg cya cyb tel gvi ana ava pma pmm pmt pmn pmi ter bth bfr bfs pgi sru chu cte cch plt det deh dra dge tth ttj aae tma mja mmp mac mma mtp mth mka afu hal hma hwa nph tac tvo pto pho pab pfu tko ape sso sto sai pai neq

A.2 PHYLOGENETIC TREE

The full phylogenetic tree is too large to reasonably fit in this page, so a portion of the phylogenetic tree are given below:

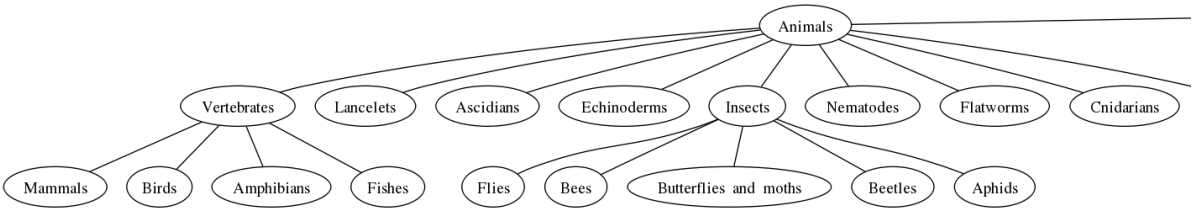

Figure A.1: A portion of the phylogenetic tree showing the animal kingdom

A.2.1 Phylogenetic Tree Groups

The groups given below are used in our tests:

Acidobacteria, Actinobacteria, Alpha/others, Alpha/rhizobacteria, Alpha/rickettsias, Amphibians, Apicomplexa, Bacillales, Bacteroides, Basidiomycetes, Beta, Birds, Budding yeasts, Butterflies and moths, Cellular slime molds, Chlamydia, Clostridia, Crenarchaeota, Cyanobacteria, Deinococcus-Thermus, Delta, Entamoeba, Epsilon, Euglenozoa, Eurotiales, Euryarchaeota, Fishes, Fission yeasts, Flies, Fusobacteria, Gamma/enterobacteria, Gamma/others, Grass family, Green nonsulfur bacteria, Green sulfur bacteria, Hyperthermophilic bacteria, Lactobacillales, Magnetococcus, Mammals, Microsporidians, Mollicutes, Mustard family, Nematodes, Red algae, Spirochete