Prediction of Regulatory Networks for Non-Model Organisms

by

Rachita Sharma

Bachelor of Engineering in Computer Technology,
Nagpur University, 2000
Masters in Computer and Information Science,
State University of New York, 2003

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctorate of Philosophy in Computer Science

In the Graduate Academic Unit of Computer Science

Supervisor(s):	Patricia A. Evans, Ph.D., Faculty of Computer Science
	Virendrakumar C. Bhavsar, Ph.D., Faculty of Computer Science
Examining Board:	Michael Fleming, Ph.D., Faculty of Computer Science
	Weichang Du, Ph.D., Faculty of Computer Science
	Denise Clark, Ph.D., Department of Biology, Chairperson
External Examiner:	Rajendra R. Joshi, Ph.D.
	Associate Director and HOD
	Bioinformatics Resources and Applications Facility (BRAF)
	Centre for Development of Advanced Computing (C-DAC)
	University of Pune

This dissertation is accepted by the

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

May, 2013

Abstract

Identification of gene regulatory networks is useful in understanding gene regulation in any organism. Some regulatory network information has already been determined experimentally and using statistical methods for model organisms, but much less has been identified for non-model organisms. The limited amount of data available for non-model organisms makes inference of regulatory networks difficult using the commonly used statistical methods. This thesis proposes a method to determine the regulatory links that can be mapped from a distant model to a non-model organism. Experiments are performed to map the regulatory network data of *S. cerevisiae* to *A. thaliana* and analyze the results. Mapping a regulatory network involves mapping the transcription factors and target genes from one genome to another. In the proposed method, different techniques for predicting transcription factors and target genes for the non-model organism using the available data for the model organism are compared and analyzed. The techniques that obtain the best results overall should be the ones chosen for these predictions. These predicted transcription factors and target genes are then integrated into predicted regulatory links for the non-model organism. A set of rules is then defined on the gene expression experiments to filter these predicted regulatory links that are well supported. Very limited available gene expression data of the non-model organism is used to filter the predicted regulatory links based on these rules to get rid of the high number of false positives. Finally, the filtered regulatory links are tested against a large dataset of gene expression

experiments to illustrate that correctly predicted regulatory links are obtained. The links thrown out by filtration are also tested against the same gene expression dataset to illustrate the significance of this step to refine the results.

Dedication

I dedicate this dissertation to my parents and my husband. A special feeling of gratitude towards my loving parents who always encouraged me and taught me persistence. My husband believed in me and stood by me throughout the long hours of research and writing. It would not have been possible to reach this goal without his support.

Acknowledgements

I would like to express my deepest appreciation to my supervisors, Dr. Patricia Evans and Dr. Virendra Bhavsar, for their tremendous support throughout my doctoral studies. Dr. Evans and Dr. Bhavsar shared their exceptional knowledge and experience in the field of Computational Biology and Parallel/ Distributed Computing, respectively to help address the challenging problems in our work. Without their guidance and persistent help this dissertation would not have been possible.

Lu Yang and Joey Bernard, both from the ACEnet group at the University of New Brunswick, provided timely technical support for the development of my proposed workflow.

I would also like to acknowledge Jodi O'Neill and Linda Sales of the Department of Computer Science, who were always prompt in taking care of any required paperwork or arranging resources at any time. A special thanks goes to my ex-group members, Aijazuddin Syed, En Zhang, Marc Cooper and Zheng Wang, who engaged in discussions ranging from technical to personal topics. I would also like to thank my Examiners for their engaging participation during the defense.

The members of my entire family greatly supported my study in Canada. I appreciate them all.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Determination of gene regulatory networks has now become an extremely important process in the field of computational biology. Experimental techniques such as gene-knockout experiments are extremely time-consuming and inadequate to find regulatory networks at the genome level. The amount of genetic information available for newly sequenced genomes is increasing exponentially and, therefore, it is essential to evolve mechanisms to bridge the gap and infer regulatory networks for newly sequenced genomes that are of interest to biologists. These regulatory relationships in a network can help answer current biological questions, such as the identification of genes and proteins related to various diseases, and are useful in novel drug design and development. These regulatory relationships can also be useful in understanding the differences in gene regulation between different organisms. Since it is extremely important to determine gene regulatory networks for understanding gene regulation by identifying these genes and their relationships, it is critical to study how genes are involved in regulation, or the way they are themselves regulated by other genes.

A lot of methods have been developed for identifying regulatory networks for model organisms. These methods work well with large amounts of gene expression data that are readily available for model organisms. Model organisms are the ones that are investigated thoroughly by the biologist by virtue of being simpler and easy to manipulate and having short life cycles. The information from these model organisms can be mapped to newly sequenced organisms about which less is known, called non-model organisms. These non-model organisms have very limited data available for using the previously mentioned methods. Therefore, there is need for a method to determine some gene regulation information about these newly sequenced genomes for the researchers.

1.2 Objective

The proposed method determines how gene regulatory information can be mapped from a model organism (source genome) to a non-model organism (target genome). Identification of regulatory information for a non-model organism will help the biologists investigate any new organism at the genome level and provide information about the common regulatory relationships between the two genomes. The method in this thesis involves mapping the transcription factors (TFs), target genes (TGs) and their regulatory relationships from one genome to another. Gene expression data available for the target genome is used in this thesis to refine and evaluate the predicted regulatory network for the target genome. A set of rules is established to use the expression information in the gene expression experiments for analyzing the predicted regulatory relationships for the target genome. In the proposed method, any model organism with available experimentally confirmed regulatory network and transcription factor binding site (TFBS) motif information can be used as the source genome. However, a model organism closer to the target genome is preferable, since

evolutionarily closer organisms will tend to have more similar regulatory relationships. The target genome can be any non-model organism with available nucleotide sequence data. For experimentation and testing purposes *Saccharomyces cerevisiae* and *Arabidopsis thaliana* have been used in this thesis as source and target genomes, respectively. *Saccharomyces cerevisiae* is a model organism with all the required information required for a source genome. *Arabidopsis thaliana*, not being a non-model organism, is used as the target genome for experimentation purposes only, so that the gene expression data available for *Arabidopsis thaliana* can be used for analysis and verification of the mapped regulatory relationships determined using the proposed method. *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are used as source and target genomes, respectively, since there is no suitable pair of organisms that are evolutionarily close that have both sufficient regulatory network and gene expression information available for verification. Therefore, the results in this thesis identify the proportion of regulatory relationships that can be mapped from an evolutionarily distant model organism.

1.3 Organization

Chapter 2 gives background information about genes, gene regulation and transcription regulatory networks. This chapter also concisely describes the computational models that have been used for determining regulatory networks for model organisms. Additionally, different biological data types and bioinformatics tools used in this thesis work are introduced in this chapter.

The first step in predicting a non-model organism's regulatory links is mapping transcription factors from one genome to another, which involves finding similar protein sequences in the target genome performing similar functions. The method, experiments and results for mapping transcription factors are described in Chap-

ter 3 for bacterial genomes and in Chapter 4 for eukaryotic genomes. These are investigated separately because yeast and plants are complex eukaryotic genomes that contain a nucleus and bacteria are a large domain of prokaryotic microorganisms that do not contain a nucleus. The mapping of transcription factors between genomes based on evolutionary distance has been investigated in Chapter 3 using 14 bacterial genomes. Transcription factors from model bacterial genomes, *Bacillus subtilis* and *Escherichia coli* are mapped to other sequenced bacterial genomes. For each of the bacterial genomes, we determined the best e-value threshold for BLAST and the best model organism to use between *Bacillus subtilis* and *Escherichia coli*. In Chapter 4, the method with best results was determined from among the three methods used for mapping transcription factors from *Saccharomyces cerevisiae* to *Arabidopsis thaliana*. For each of the three methods, an experiment was conducted to identify a suitable BLAST parameter threshold using a ROC curve. Another experiment was performed to determine if a predicted target genome transcription factor, mapped using sequence similarity, was of the same type or shared the same function as the source genome transcription factor used for mapping.

Chapter 5 presents the different methods and their results for predicting target genes of a non-model organism, the next step of regulatory link prediction. Experiments are conducted by predicting target genes for *Arabidopsis thaliana* using *Saccharomyces cerevisiae* as the source genome. The predicted target gene result sets using the three methods are analyzed. The last step of integrating the regulatory element's information along with its refinement using a small set of gene expression data is explained in Chapter 6. Additionally, the predicted regulatory links for *Arabidopsis thaliana* are analyzed and tested against a very large set of gene expression data. Finally, the conclusion in Chapter 7 provides an overview of the results from the previous chapters and discusses their signficance and that of the overall method.

Chapter 2

Background

This chapter provides the relevant biological background and an overview of the computational tools that have been employed in this thesis.

2.1 DNA

Deoxyribonucleic acid (DNA) [39] is a polymeric chain made up of four nucleotide bases - adenine (A), cytosine (C), guanine (G) and thymine (T), as shown in Figure 2.1. DNA is a double helical structure constituting of two polynucleotide chains, where the bases are paired (Figure 2.2). A DNA sequence is depicted by a sequence of letters - A, C, T and G - representing the four types of constituent bases.

2.2 Genes

A gene is the physical and functional unit of heredity, which consists of protein-coding and non-coding sequences in a DNA that corresponds to a specific polynucleotide chain as shown in Figure 2.2. All genes in an organism make up the organism's genome which contains the entire array of commands needed for its sustenance. The protein-coding and non-coding regions of genes are long DNA sequences made up

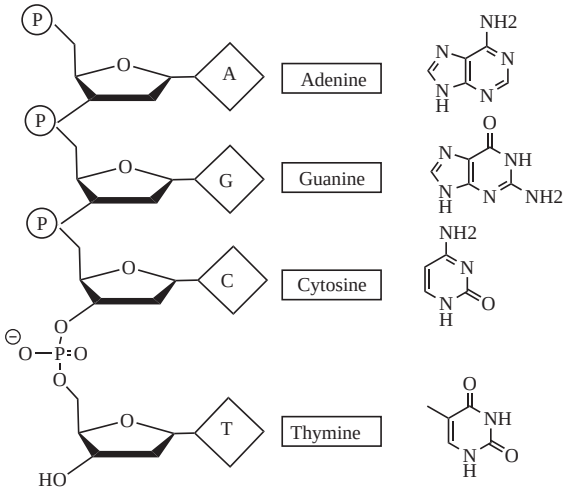


Figure 2.1: DNA chemical structure

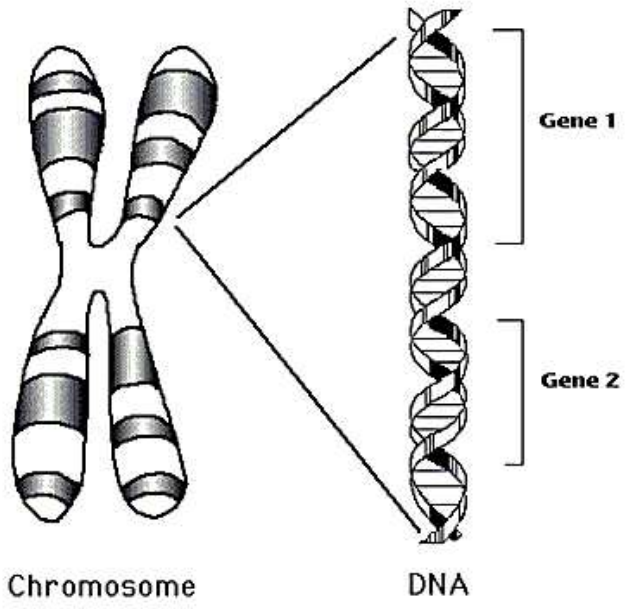


Figure 2.2: Gene in a DNA double helix structure [16]

of promoters, introns, and exons as shown in Figure 2.3. A promoter administers the activity of a gene and is generally located upstream of that gene, while exons mostly are the coding sequences in a gene deciding the final gene product that may be produced. Also, the introns are the non-coding sequences in a gene that determine when the gene is expressed, that is, control the conditions of production of the gene product. The analysis in this thesis is restricted to protein-coding genes and not non-coding RNAs even though there are lots of non-coding RNAs that are part of regulatory networks.

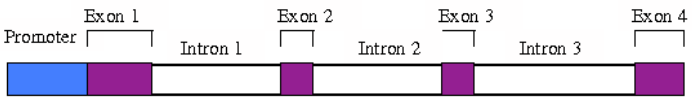


Figure 2.3: Basic gene structure

A gene sequence example in FASTA format, a text-based format for representing either nucleotide sequences or peptide sequences, is shown in Figure 2.4. The first line contains the nucleotide sequence identifier followed by the nucleotide sequence description. The second line consists of the complete nucleotide sequence for the given nucleotide sequence identifier.

```
>sce:YAL068C PAU8; Pau8p
atggtcaaattaacttcaatcgccgctggtgtcgctgccatcgctgctactgcttctgca
accaccactctagctcaatctgacgaaagagtcaacttggtggaattgggtgtctacgtc
tctgatatcagagctcacttagcccaatactacatgttccaagccgcccacccaactgaa
acctacccagtcgaagttgctgaagccgttttcaactacggtgacttcaccaccatgttg
accggtattgctccagaccaagtgaccagaatgatcaccggtgttccatggtactccagc
agattaaagccagccatctccagtgctctatccaaggacggtatctacactatcgcaaac
tag
```

Figure 2.4: Example of a gene sequence in FASTA format

2.3 Gene regulation

All the cells in an organism have the same DNA sequence, though this sequence is used to produce different proteins in different cells that take different functional roles. These proteins are linear polymers of amino acids and their sequences are made up using combinations of 20 biologically available amino acids, an example shown in Figure 2.5. The first line contains the protein sequence identifier followed by the protein sequence description, while the second line consists of the complete amino acid sequence for the given protein sequence identifier. Genes do contain the code for protein sequences, but they have to be expressed (turned on) for the corresponding proteins to be produced and gene regulation is involved in controlling the expression of these genes and allows a cell to have control over its structure and function. Therefore, a gene or a group of genes can be involved in regulating another gene or a group of genes. Only specific genes need to be activated or expressed for the production of specific proteins, so it is important to study how genes are involved in regulation and the way they are themselves regulated by other genes.

>FOSB_MOUSE Protein fosB. 338 bp
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL

Figure 2.5: Example of a protein sequence in FASTA format

2.3.1 Gene expression

Gene expression is a multi-step process resulting in the synthesis of proteins by using the coded information from a gene as shown in Figure 2.6. Transcription [39], the first step, involves copying a strand of a DNA molecule into RNA (RiboNucleic Acid),

which is catalyzed by the RNA polymerase enzyme. Regulatory elements in the DNA sequence, including its promoters and its enhancers, enable gene transcription. In the next step, some parts of RNA, its introns, are removed during RNA processing, and exons, sequences that encode proteins, are converted into mRNA. Finally, mRNA is translated into a protein sequence that can take a structural or functional role and can act as a regulatory protein for regulating another gene.

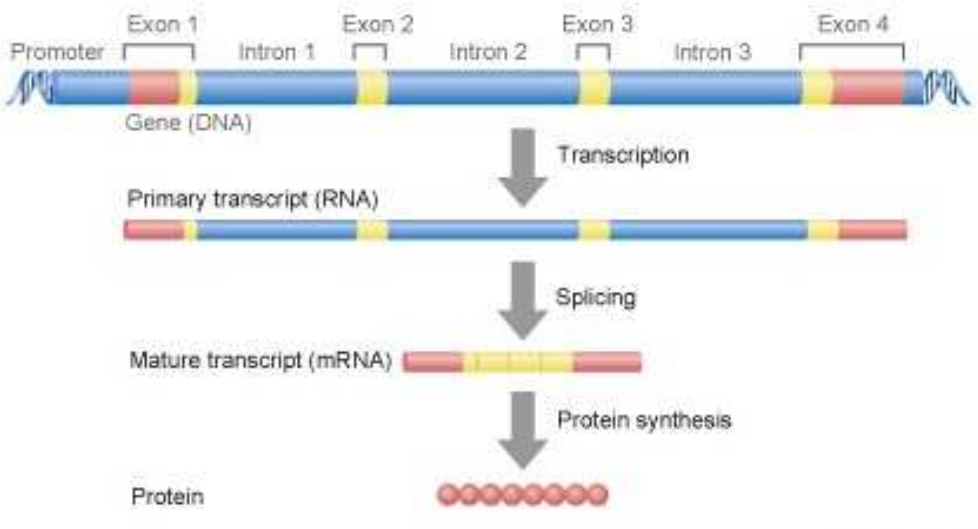


Figure 2.6: Gene Expression [69]

Gene regulation in a cell controls the amount and timing of the appearance of a gene's functional product. Gene regulation can occur at various stages as shown in Figure 2.7. However, most of the regulatory events for controlling expression of protein-coding genes occur at the level of gene transcription [22], the first step in gene expression. Therefore, regulation of gene expression is mostly done by changing the amount and rate of transcription of the gene.

2.3.2 Transcription

The process of transfer of genetic information from DNA to RNA is called transcription. It involves production of a single strand of RNA that is similar in sequence with

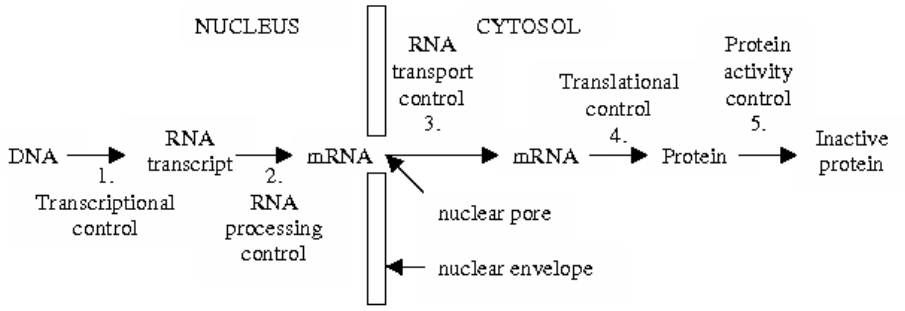


Figure 2.7: Stages of control of gene expression [3]

one of the DNA strands. Transcription [39] of a gene is initiated with the binding of RNA polymerase enzyme with the promoter, a special region at the beginning of the gene. The sequences before and after the start point are called upstream and downstream relative to the start point. The group of key elements such as promoter, operator, structural genes and regulator genes is called an operon in prokaryotes, as shown in Figure 2.8.

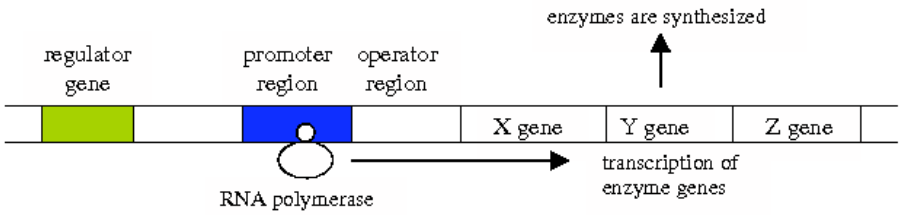


Figure 2.8: Operon structure in prokaryotes

2.3.3 Transcription regulation

Transcription regulation of gene expression [71] can be defined as the binding of transcription factors to binding sites. The groups of proteins that together initiate transcription are termed transcription factors (TFs), and are either transcriptional activators or transcriptional repressors. The binding site is a short DNA sequence

and is generally located upstream of the target gene near the promoter region. Another example of a binding site would be an operator in prokaryotes which is located close to the promoter and binds to the repressor protein to prevent transcription initiation. Binding sites occurring within a hundred or thousand base pairs of DNA from the start site also affect the transcriptional activity of the genes. These binding sites are called enhancers for activation, or silencers for repressing transcription. Regulatory regions of eukaryotes (organisms whose cells contain a nucleus) generally have more sites stretched over a long length of DNA.

In negative gene regulation, transcription is repressed when a silencer (repressor protein) is bound to the promoter (in eukaryotes) or operator (in bacteria), hence inhibiting the binding of the promoter with the RNA polymerase, as shown in Figure 2.9. An activator can combine with this repressed protein to form a complex and dissociate the complex from the promoter. This initiates transcription by allowing RNA polymerase to bind the promoter. In positive gene regulation, the activator combines with an inactive transcription factor to bind to the promoter forming the transcription factor-activator complex. RNA polymerase then binds to the promoter initiating transcription [39], as shown in Figure 2.9.

The DNA binding sites can occur in regulatory regions in different arrangements representing different ways of regulation. The four different arrangements of the binding sites binding to their regulators are illustrated in Figure 2.10. A single regulator has only one binding site for a single transcriptional regulator, while repetitive motifs have multiple binding motifs for a given regulator and multiple regulators have many binding sites corresponding to several regulators. These regulators are involved in the gene regulation collectively. Some pairs of regulators are co-occurring regulators, and these can bind to sites in the same regulatory region and might interact physically or be functionally related [26].

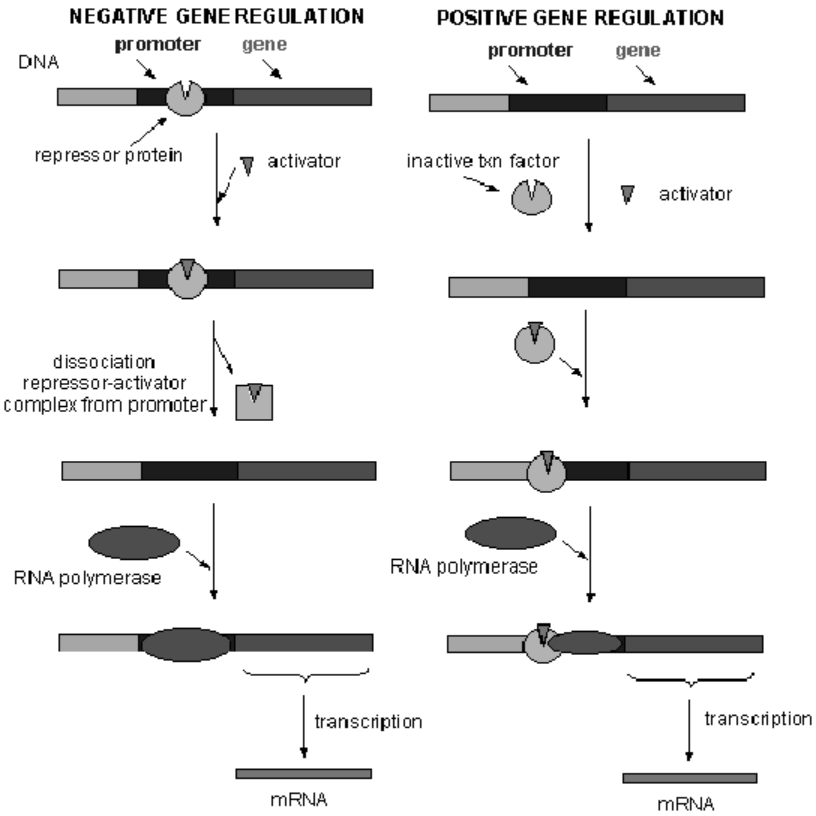


Figure 2.9: Negative and positive gene regulation [31]

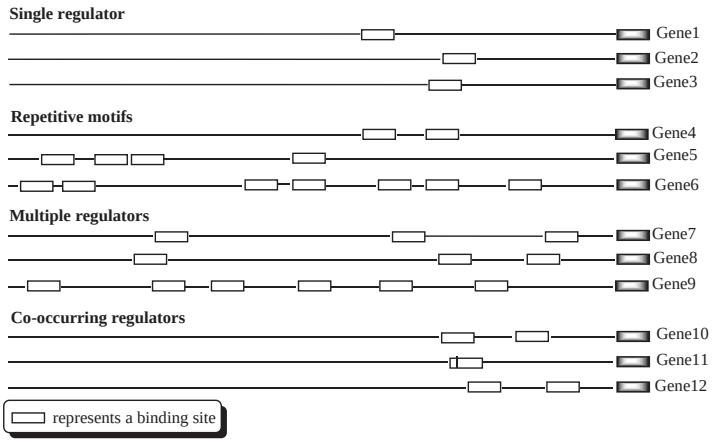


Figure 2.10: Promoter architecture [26]

2.3.4 Transcription regulatory networks

A gene, a part of a DNA sequence, is converted into protein through transcription and translation. Some proteins are transcription factors acting as either activators or repressors, and either act alone or in combination. Transcription factors and the genes they regulate combine in six different possible common classifications of combinations, as shown in Figure 2.11 [38]: autoregulation, multi-component loop, feedforward loop, single input motif, multi-input motif and regulator chain. These different combinations of regulatory elements are called network motifs, and together constitute a gene regulatory network.

Each of these different network motifs characterizes a different type of regulatory combination, as follows. In an autoregulation network motif, a protein binds to the promoter region of its own protein-coding gene activating transcription and translation, leading to more production of the same protein. A multi-component loop consists of two proteins that each bind to the promoter region of the other one. In a feedforward loop, one protein binds to the promoter region of the protein-coding gene of another protein, and then these two proteins both bind to the promoter region of the same third gene. A single input motif consists of a protein that binds to the promoter regions of a group of genes. In a multi-input motif, a group of proteins bind to the promoter regions of another group of genes. Finally, a regulatory chain is a chain of three or more proteins where one protein binds to the promoter region of another protein-coding gene producing the second protein, which then binds to another protein-coding gene producing the third protein, and so on.

A gene regulatory network, as shown in Figure 2.12, describes a collection of genes and gene products interacting with each other and with other chemicals in the cell, regulating the rate of transcription of genes in the network into mRNA. It can be represented as a complex directed graph consisting of genes as nodes and regulatory relationships between genes and gene products as directed edges. These graphs can

also represent a group of different network motifs.

2.4 Computational models for Regulatory Network Inference

Significant time and resources are required for the experimental determination of these gene regulatory networks. Experimental techniques such as gene-knockout experiments [23] are extremely time-consuming and in many cases inadequate to identify a regulatory network for an organism at the genome level. The amount of genetic information available for newly sequenced genomes is increasing exponentially, and therefore it is essential to develop methods to bridge the gap and infer regulatory networks for these new genomes.

Several computational models have been used to represent gene regulatory pathways

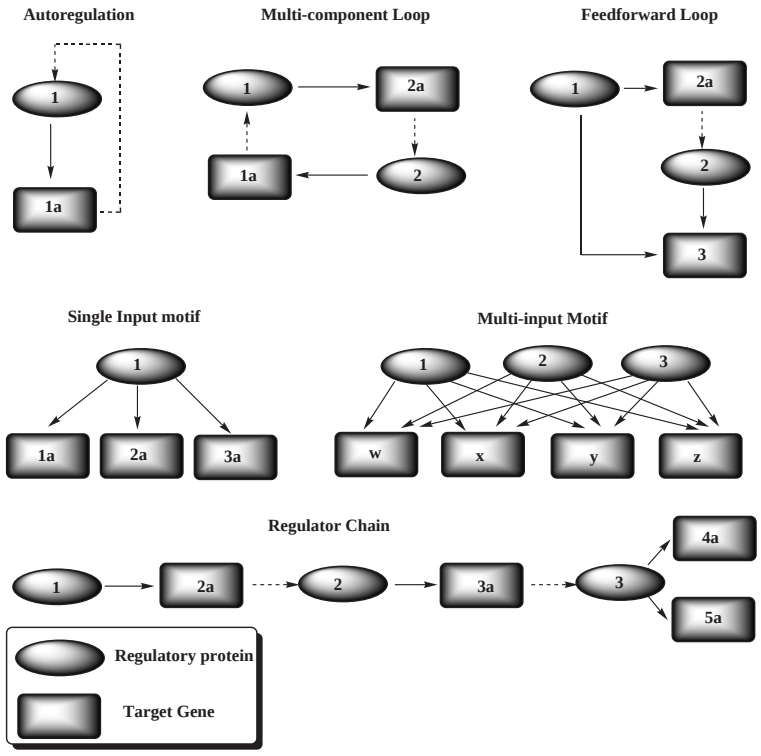


Figure 2.11: Types of network motifs [38]

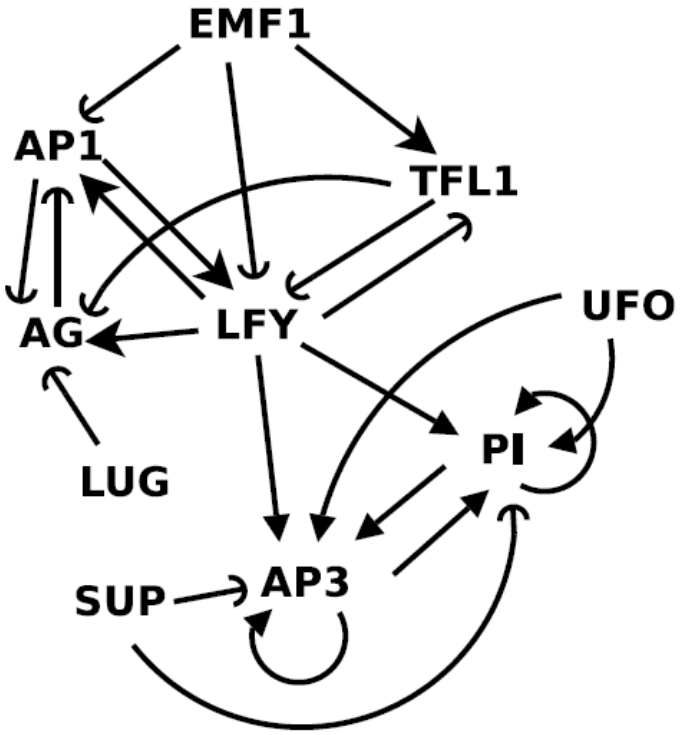


Figure 2.12: Gene regulatory network of *A. thaliana* flower morphogenesis [21]: activation relationship is represented by pointed arrows, whereas a repression relationship is represented by blunt arrows between genes. Relationship example: EMF1 activates TFL1, but it is also a repressor of LFY.

for model organisms using gene expression data, and some models incorporate additional available biological knowledge. These models include Boolean Networks [40, 2], Bayesian Networks [29, 4, 42], Differential equation models [15, 20] and Hybrid Petri Nets [43]. Recently, many techniques have started incorporating additional data, such as protein-protein interaction data, protein-DNA interaction data, and binding site data, along with the gene expression data to get more accurate gene regulatory networks. Additionally, many methods are now focusing on using time-course behavior of gene expression along with available biological knowledge. Some of these techniques have been compared in recent reviews [27, 76] of reconstruction of gene regulatory networks.

The above mentioned techniques cannot be used for non-model organisms due to the sparseness of the data available for non-model organisms. There have been so far two methods developed, P-MAP [49] and another one used in KEGG [34], that use a regulatory network of a model organism to infer a regulatory network for a non-model organism. However, these methods are limited in their usability for mapping regulatory pathways (see Section 2.5 for details).

2.4.1 State Transition Tables and Boolean Network Models

Boolean Networks are used to represent regulatory networks based on directed graphs, and can be inferred using time series of gene expression data. A gene is implemented as a binary variable and is characterized as ON or OFF. The state of each gene is related to other genes using logical rules. The network consists of a set of nodes as genes and a set of boolean functions as logical rules.

REVerse Engineering Algorithm (REVEAL) [40] was developed based on state transition tables using Information Theoretic Principles of Mutual Information [59]. Using this method [40], the set of genes regulating every gene in a regulatory network could be identified using a quantitative information measure between input and out-

put states. Mutual information is calculated between a gene and its regulatory genes to infer gene regulations and boolean functions. The results of computational experiments for this method showed that it performed well for low values of indegree (number of inputs per gene).

Another algorithm [2] was developed to be simpler compared to REVEAL but with worse efficiency of time and memory space. The results of computational experiments for this method showed that for maximum indegree of each node as 2 and for about 100 input/output pairs, a boolean network of size 100,000 can be inferred.

The advantage of using Boolean Networks is that they are rule-based and computationally simple. But the limitation is that extremely high computing times are required to construct reliable network structures and the gene expression cannot be represented appropriately by only two states.

2.4.2 Bayesian Networks

Bayesian Networks are used to infer regulatory networks using a graph-theoretic approach. They use a joint probability distribution of the estimated network and decompose it into the product of conditional probability distributions. Mostly, gene expression data has been used with Bayesian networks to construct a regulatory network. The methods using Bayesian Networks may work well for small size networks but not for actual large regulatory networks, since they generally have many parameters and require huge amounts of data for reliable network inference.

2.4.3 Differential Equations

Differential equation models are used to represent regulatory networks, and these range from traditional linear ODEs (Ordinary Differential Equations) to power law (non-linear) ODEs using continuous variables. These techniques show theoretically how linear algebra and Fourier transforms can be used to estimate parameters for

the models. Differential equations can represent frequent non-linear behaviors based on reaction kinetics, binding thermodynamics or molecular diffusion.

One of the methods to represent regulatory networks [15] uses a linear transcription model using differential equations for gene expression. The limitation is that the reliable identification of large scale gene regulatory networks using gene expression data cannot be done adequately by simplicity of linearization. On the other hand, a large amount of experimental data is required for the reliable determination of non-linear interactions because non-linear models [70, 72] consist of many parameters. Hence, the practical application of non-linear models is limited by data insufficiency.

2.4.4 Other Methods

Aside from the recently used models and techniques described above, there are some old techniques used infrequently to infer regulatory networks. A method based on statistical techniques [73] has been used to identify the binding site clusters in a genome to construct a regulatory network. This method is based on the fact that a lot of transcription factors work together in transcription activation, leading to closely spaced transcription factor binding sites; this type of arrangement occurs rarely on its own. To identify the genes, which might be regulated by a transcription factor, the method used information about the known biochemical functions of the transcription factor, the structure of its binding site and the function of the genes near the cluster. The results of this method showed a useful feature that genes regulated by similar transcription factors in different organisms can be functionally related.

Unlike the discrete variable models described above, a continuous variable model, Hybrid Petri Net [43], is also used to represent gene regulatory networks. Petri nets are bipartite graphs consisting of places, transitions and arcs connecting them and are used as graphical and mathematical modeling tools. Hybrid Petri Net is

an extension of Petri nets and can represent the relationship between discrete and continuous values explicitly. It can handle continuous factors (mRNA or protein concentration), which cannot be handled by ordinary Petri net models.

2.4.5 Remarks

Although all the above methods have been used to infer gene regulatory networks using databases of a model organism, there are various issues related to using the above methods for regulatory network inference. A large amount of data is required for accurate network prediction using these methods, and they may only be able to find individual relationships and may not find the direction of the regulatory relationships. They use gene expression data, which makes it difficult to determine the actual regulatory relationships because the predicted relationships will also contain indirect relationships through other genes. The above methods also do not work well for large actual regulatory networks of model organisms, even with the substantial information that is available for model organisms. Non-model organisms, on the other hand, have very limited data available to use in these network inference methods.

2.5 Mapping Regulatory Pathways

Due to the limitation mentioned in the previous section, pathway mapping is used to infer regulatory networks for non-model organisms. The following methods describe how a biological pathway from one organism's genome can be used to predict a biological pathway for another organism's genome using limited data.

2.5.1 P-MAP

P-MAP [49] is used to map regulatory pathways from one genome to another. The mapping, taken as a constrained minimum spanning tree problem, is done based on keeping overall sequence similarity between mapped gene pairs and preserving the co-regulated gene structures in the predicted pathway. The advantage of this method is that it can be used for closely related and remotely related microbial genomes (such as, *B. subtilis* to *E. coli*). However, the limitation of this method is that it does not predict the direction and type of regulatory relationships between the genes. Also, it does not consider other possible data that might be available for the target genome, such as gene expression and proteomics data.

2.5.2 KEGG

Another technique to map regulatory pathways is used in KEGG [34]. The KEGG database is a publicly available collection of pathway maps which consists of 38,187 pathways generated from 296 reference pathways. These reference pathways are from bacterial or archaeal genomes and contain about one-fourth to one-third of genes similar to the genes in other pathways. Therefore, organism-specific pathways are constructed based on ortholog identifier assignments in the GENES database. Orthologs are evolutionarily related genes, having common ancestors, in different species. KEGG relies on previously determined information about homologous genes, and it does not consider any other data such as gene expression data or protein-DNA interaction data. Therefore, the use of only identified orthologs limits the pathways that can be mapped in this way.

2.5.3 Remarks

Determination of regulatory networks is an important part of the process of investigating non-model organisms. There is a lot of experimental data available for a few model organisms, which is being used for construction of their gene regulatory networks, quite often confirmed by gene knockout experiments. These experiments are very expensive, time consuming and cannot be performed for every organism. Only limited data are available for non-model organisms. Therefore, the previously described computational models cannot be used to predict regulatory networks for a non-model organism due to insufficient and sparse data.

There are also issues using the two methods discussed above for mapping pathways. The regulatory relationships predicted may not have the correct direction predicted and these methods do not use additional data such as gene expression data, binding site data, transcription factor data, network motif data, and protein-protein interaction data.

2.6 Tools used in Regulatory Network Prediction

Various tools commonly used for different tasks have been integrated in this thesis work. These tools are selected based on the requirements of the proposed system. To minimize the execution time and dependability of the system, all the tools selected and integrated in this system are publicly available, free of cost and can be easily downloaded, installed and executed on local machines. These tools are BLAST [5], a sequence alignment tool; InterProScan [78], a protein domains identifier tool; a nucleosomes position prediction tool [58] and GALF-P [14], a motif search tool.

2.6.1 BLAST

Sequence alignment tools are useful in identifying similarity among DNA sequences of genes or amino acid sequences of proteins that could have similar structure and function. These tools assist in identifying a group of similar sequences and investigating their unique characteristics that make these sequences similar. There are various sequence alignment tools available that are used for identifying similar sequences based on the type of sequence alignment method used, including the widely-used BLAST (Basic Local Alignment Search Tool) [5]. In this thesis, BLAST has been used for mapping regulatory sequences from a model organism's genome to a non-model organism's genome. It compares a query nucleotide or protein sequence to all the sequences in a sequence database to identify local similarity regions and generates gapped alignments of these regions. Local sequence alignment is used in BLAST to find similar regions in sequences as opposed to global sequence alignment that aligns every piece of the sequence. Each aligned pair is associated with an alignment score and e-value (expectation value) parameter representing sequence similarity. The score S corresponds to the degree of sequence similarity and is computed as the sum of aligned and gap scores. The e-value E is the number of times one would expect a hit by chance with scores more than or equal to S, with lower e-values corresponding to higher sequence similarity.

2.6.2 InterProScan

In addition to sequence alignment, more knowledge can be obtained about proteins in relation to their structure and function. This information about protein domains, families and functional sites has been collected in different protein signature databases. InterPro [44] is a non-redundant database that integrates these commonly used protein signature databases into one resource. The corresponding scanning or protein signature recognition methods for these integrated protein signa-

ture databases are combined into the InterProScan tool [78]. InterProScan tool can be downloaded and executed on a stand-alone machine and integrated into another tool for sequence analysis, and produces output that indicates both the protein signature results and the specific methods used.

In this thesis, the PANTHER (Protein ANalysis THrough Evolutionary Relationships) database [67] and its corresponding scanning tools, BLAST and hmmsearch, are used within InterProScan. This is because the PANTHER database is used in the thesis to determine genes with similar functionality and this database has protein sequences classified into families and subfamilies with similar function based on published experimental evidence and evolutionary relationships.

2.6.3 Nucleosomes Position Prediction Tool

Eukaryote genomes consist of an array of nucleosomes packaged tightly into a chromatin structure consisting of DNA and proteins. The function of this chromatin structure and its involvement in controlling transcription has become valuable in understanding gene regulation. A nucleosome is a basic unit of chromatin structure, consisting of 147 base pairs of DNA and four pairs of proteins named histones, repeated over the genome DNA. The DNA base pairs are closely rolled up around the histones as shown in Figure 2.13.

About 75 to 90 percent of the DNA in a genome is bundled up in nucleosomes. The unwrapped DNA between two neighboring nucleosomes is about 10-50 base pairs of linker DNA. The DNA wrapped around the histones in a nucleosome is not available to be involved in any kind of interaction with proteins to form different complexes. Therefore, gene expression is controlled by the presence of nucleosomes on DNA. Hence, it is necessary to determine the location of nucleosomes on the genome sequences to aid in locating active binding sites and understanding gene regulation for that genome. The Nucleosome Position Prediction Tool [58] has been

developed to find nucleosome occupancy locations within a given sequence. The output of this tool provides the probability of a nucleosome being present and the start of nucleosome occupancy at every position in the sequence.

2.6.4 GALF-P Tool

Transcription factor binding site (TFBS) motifs are a small part of a nucleotide sequence that binds to the transcription factors and thus controls gene expression. The interaction of a transcription factor with a TFBS for a gene affects transcription of that gene. These TFBS motifs are located in the cis-regulatory region or promoter region and are generally found upstream of the genes they regulate. Computational methods have been developed to predict these TFBS motifs, which are critical elements for gene regulation.

Current genetic algorithms for this purpose use position-led and consensus-led rep-

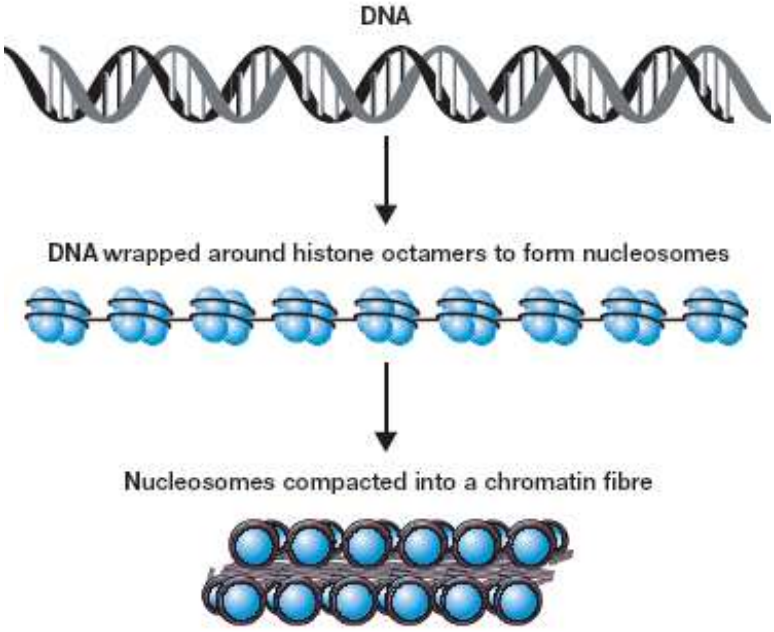


Figure 2.13: DNA bundled around nucleosomes [47]

resentations independently, whereas the GALF-P (Genetic Algorithm with Local Filtering and adaptive post-processing techniques) tool [14] combines both representations and uses local filtering to decrease false positives and improve efficiency. The length of the TFBS motif to be searched can be varied based on requirements. The results based on synthetic and real datasets using the genetic algorithms, GAME [75], BioOptimizer [32], MEME [7], and BioProspector [41] compared to GALF-P demonstrated better performance of GALF-P. GALF-P is used in this work to find known TFBS motifs in nucleotide sequences.

2.7 Data used in Regulatory Network Prediction

Different biological data, such as nucleotide sequence data, protein sequence data, transcription factor data, transcription factor binding sites motif data, network motif data, gene expression data and regulatory network data, have been used in this thesis.

A nucleotide sequence database for a genome consists of all the DNA sequences for that genome that have been sequenced. Nucleotide sequence data for different genomes are publicly available in different databases such as EMBL Nucleotide Sequence database [17] and NCBI GenBank database [10].

A protein sequence database for a genome consists of all the amino acid sequences for that genome. Protein sequence data for different genomes is publicly available in different databases such as NCBI GenBank database [10], PIR (The Protein Information Resource) database [9], SWISSPROT [11] and TrEMBL database [11].

A transcription factor is a regulatory protein in a genome that controls the expression of a gene (activate or repress) by binding to the regulatory regions located generally upstream of that gene. Transcription factor data consist of protein sequences in FASTA format as a transcription factor is a specific type of protein. Transcription

factor data for different experimentally investigated genomes is publicly available in different databases such as PlantTFDB (Database of Plant Transcription Factors) [25], and Yeastract Database [66].

Groups of transcription factors are characterized into families that tend to share a common structure among their binding sites (in the DNA sequence they bind to) though there is a considerable diversity in the binding sites for any transcription factor. These common structures are represented as transcription factor binding site motifs. Transcription factor binding site data is represented in a regular expression or position-weight matrix format. Transcription factor binding site data is publicly available in databases such as TRANSFAC [28] and JASPAR [57].

A promoter sequence is a small specific nucleotide sequence called a regulatory sequence present upstream of a gene sequence. It contains specific DNA sequences (binding site motifs) that are recognized by the transcription factors. A promoter sequence is represented in FASTA format in the same way as a gene sequence. Promoter sequence data for various experimentally investigated genomes is publicly available in different databases such as EPD (The Eukaryotic Promoter Database) [52], and TRED (Transcriptional Regulatory Element Database) [33].

Gene expression data are comprised of expression levels of genes in a genome obtained experimentally under specific conditions. There are different formats of gene expression data depending upon the database resource used. Gene expression data are publicly available in databases such as Stanford MicroArray database [64] and EBI ArrayExpress Microarray database [12].

A gene regulatory network is a collection of positive and negative regulatory interactions between genes and proteins in a genome. There is no predefined format for these regulatory interactions, and they are expressed in different formats depending on the resource. They provide information about the regulatory protein, target gene and the type of gene regulation, which is either positive or negative. There are

some databases for regulatory networks of various genomes publicly available such as RegulonDB [13] and AGRIS (Arabidopsis Gene Regulatory Information Server) [19, 50].

Chapter 3

Transcription Factor mapping between Bacterial Genomes

3.1 Overview

The first step of mapping a regulatory link from a source genome to a target genome is mapping the transcription factor. This chapter [61, 62] [1] [2] focuses on finding transcription factors for bacterial species as researchers need to investigate more bacterial species below the genus level. There are more than 600 bacterial species living in the human saliva, some of which cause specific diseases such as strep throat, meningitis and bacterial pneumonia [46]. Biologists are interested in examining these genomes to research the cure for various diseases.

In this chapter, the genomes of two common model bacteria, *Escherichia coli* and *Bacillus subtilis*, are used to compare two tools, BLASTP [5] and PSI-BLAST [6], and determine the most suitable E-value threshold to use to map transcription fac-

[1] Sharma, R.; Evans, P.; Bhavsar, V. C., Transcription Factor mapping between Bacteria Genomes, International Journal of Functional Informatics and Personalised Medicine 2009, Vol. 2, 4, 424-441

[2] Sharma, R.; Evans, P.; Bhavsar, V. C., Mapping Transcription Factors from a Model to a non-Model Organism, Proceedings of the International Joint Conferences on Bioinformatics, Systems Biology and Intelligent Computing 2009, 160-166

tors between bacterial genomes. These tools are two different types of a commonly used sequence alignment tool, BLAST. The BLASTP tool is used to find similar protein sequences for a variety of applications. The PSI-BLAST tool is designed primarily to find distant evolutionary relationships among sequences by performing iterative search over protein databases. Higher E-values for sequences identified in later iterations do not necessarily validate these sequences with homology in PSI-BLAST results [30]. BLASTP uses pre-defined scoring matrices such as BLOSUM62 compared to the position-specific scoring matrices (PSSMs) used by PSI-BLAST. Different E-value thresholds are used for many different experiments based on the best performance for any particular application. A BLAST E-value threshold range of $10^{-50} < E < 10^{-10}$ has been suggested for finding a closely related sequence and another range, $10^{-5} < E < 1$, for finding a true homologue [54]. For one BLAST application [55], a suitable E-value threshold of 10^{-3} for protein sequence comparisons and 10^{-1} for nucleotide sequence comparisons has been suggested. These E-value thresholds, however, are application-dependent and cannot be generalized. Therefore, in this chapter we determine the most suitable E-value threshold for effectively mapping transcription factors between bacterial genomes, and compare BLASTP and PSI-BLAST results. This chapter also investigates the effect of using a closer or a more distant model organism as a source genome for transcription factor mapping in bacteria.

3.2 Methodology

A transcription factor (regulatory protein) of a model organism (source genome) is mapped to a non-model organism (target genome) by finding similar proteins in the target genome based on sequence similarity. The transcription factors of any source genome will not necessarily all be mapped to another target genome, as some

transcription factors of any genome are not highly similar to any factors in other genomes. The three different possible mappings of transcription factors between any two genomes are illustrated in Figure 3.1, showing that there can be one, more than one, or no protein in the target genome similar to a transcription factor of the source genome.

The Basic Local Alignment Search Tool (BLAST) [5], as discussed in Section 2.6.1,

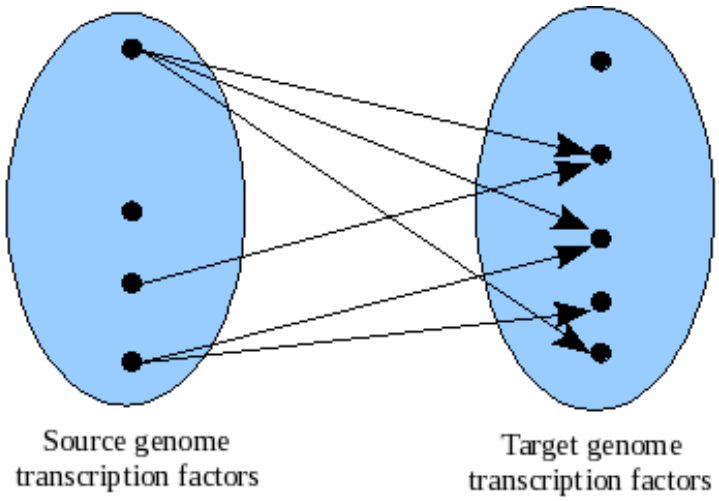


Figure 3.1: Transcription Factor mapping types from model to a non-model organism: one to many (A to A1, B1 and C1), one to one (C to C1), and none (zero from B) mapping

is a widely used sequence alignment tool used to find similarities among DNA sequences of genes, or amino acid sequences of proteins, that, as regions preserved by evolution, can indicate similar structure or function. BLASTP is a type of BLAST tool that is used for finding similarity among protein sequences. PSI-BLAST [6] is another type of BLAST tool that uses position-specific scoring matrices built using results from the BLASTP tool. This tool is recommended for use when the user gets very few homologous sequences or when the initial results are not annotated, such as if they are labelled as a hypothetical protein or as similar to a protein.

In this transcription factor mapping method, protein sequences of the target genome

that are similar to transcription factor protein sequences of the source genome are identified. These similar sequences are found by running either the BLASTP or PSI-BLAST tool (version 2.2.16) on the protein sequence database of the target genome, with transcription factor protein sequences of the source genome as input query sequences, as shown in Figure 3.2. The identified sequences in the target genome are the mapped sequences considered to be possible transcription factors. The method used in this experiment classifies the identified sequences in the target genome as positives for being a transcription factor, and negative otherwise. Various exper-

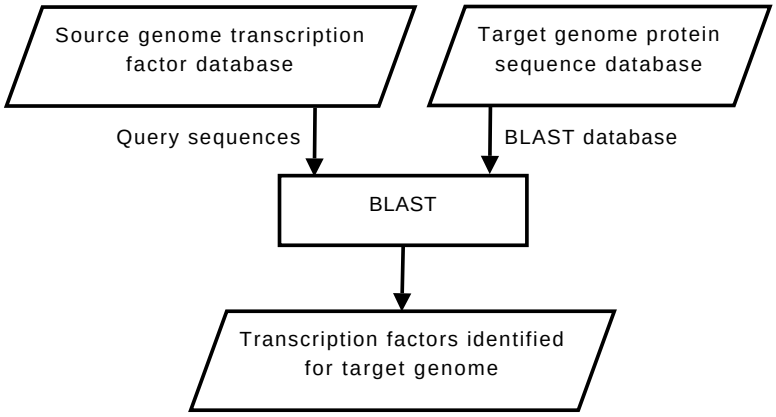


Figure 3.2: Mapping Transcription Factors from a source to a target genome

iments were performed on bacteria genomes to test this mapping method with a BLAST E-value threshold parameter ranging from 10^{-5} to 10. The results from this method are evaluated on the basis of how likely a sequence predicted as a transcription factor is actually a transcription factor of the target genome. For any pair of genomes, all the transcription factors from one genome cannot be mapped to another genome even if they are from the same species as there are bound to be some differences based on the evolutionary distance between these genomes. Therefore, the number of confirmed mappings between any two genomes is unknown as it depends on the definition of a confirmed mapping used in the experiment. It can be based on the number of proteins correctly predicted as transcription factors or the number

of transcription factors correctly mapped with the same function or similar function or the number of transcription factors that should be similar based on evolutionary distance between the genomes. In this thesis, the results of the transcription factor mapping are analysed based on proteins being correctly predicted as transcription factors by comparing the available transcription factor data of the target genome with the predicted sets of transcription factors.

The objectives of these experiments are to determine if there is a trend in transcription factor mapping between bacterial genomes with varying evolutionary distance, to determine a suitable BLAST E-value threshold for transcription factor mapping between bacterial genomes, and to compare the results using BLASTP and PSI-BLAST. The phylogenetic tree [48], representing evolutionary distance between the bacterial genomes, is used to decide which bacterial genomes to use in this experiment. There are two known model organisms in Bacteria, *Escherichia coli* (a gram negative bacterium) and *Bacillus subtilis* (a gram positive bacterium), which are used as source genomes in these experiments. Based on the phylogenetic tree [48], *Escherichia coli* and *Bacillus subtilis* are a part of the proteo-bacteria and bacillus/clostridium genus groups, respectively. Seven bacterial genomes (*H. influenzae*, *V. cholerae*, *P. aeruginosa*, *X. fastidiosa*, *N. meningitidis*, *R. prowazekii*, *H. pylori*) from the proteo-bacteria genus group are used as target genomes in this experiment, referred to as the E.coli group. Also, there are seven bacterial genomes (*S. pneumoniae*, *S. pyogenes*, *E. faecalis*, *S. aureus*, *M. pneumoniae*, *M. genitalum*, *U. urealyticum*) of the bacillus/clostridium genus group that are used as target genomes in this experiment, referred to as the Bacillus group. The accession numbers of these bacterial genomes from GenBank [10] are shown in Table 3.1. The experiment is conducted by mapping the transcription factors from the two model bacterial genomes using the proposed method to the two groups of bacterial genomes, E.coli and Bacillus groups, using a BLAST E-value threshold parameter ranging from 10^{-5} to 10.

Table 3.1: GenBank Accession Numbers for Bacterial Genomes used

Bacterial Genomes	**Accession Numbers**
H. influenzae	NC_000907
V. cholerae	NC_002505
P. aeruginosa	NC_002516
X. fastidiosa	NC_010513
N. meningitidis	NC_003112
R. prowazekii	NC_000963
H. pylori	NC_000921
S. aureus	NC_002951
E. faecalis	NC_004668
S. pneumoniae	NC_008533
S. pyogenes	NC_002737
U. urealyticum	NC_002162
M. pneumoniae	NC_000912
M. genitalum	NC_000908

The transcription factors of the two model bacterial genomes are mapped to the bacteria genomes of their own respective group first and then to the other group. Hence, the first set of mappings is between evolutionarily closer bacterial genomes and the second set of mappings is between bacterial genomes that are evolutionarily farther apart.

A binary classifier [45] is used here to analyse the predicted results and classify them into four groups. True Positives (TP) are calculated as the number of sequences, out of the total sequences predicted as transcription factors, that are experimentally confirmed transcription factors present in the available transcription factor database of the target genome. False Positives (FP) are calculated as the number of sequences, out of the total sequences predicted as transcription factors, that are not experimentally confirmed as transcription factors since they are not present in the available transcription factor database. True Negatives (TN) are the number of sequences that are discarded by the method as not a transcription factor and are not experimentally confirmed transcription factors due to not being present in the available target tran-

scription factor database. False Negatives (FN) are the number of sequences that are discarded by the method as not a transcription factor but are experimentally confirmed transcription factors present in the available target transcription factor database [77]. To further analyse the results and to evaluate the performance of the binary classifier [45], the values of Sensitivity (Sens.), Specificity (Spec.) and Positive Predictive Value (PPV), the most widely used statistics, are calculated using equations 3.1, 3.2, and 3.3 given below. Sensitivity of a test is its measure to identify positive results, i.e., all actual positives are identified for a highly sensitive test. On the other hand, specificity refers to a test's ability to identify negative results, i.e., for a highly specific test, a positive result would be a true positive as a false positive result would be very unlikely. Therefore, Sensitivity is the proportion of positives that are true and Specificity is the proportion of negatives that are true. Positive Predictive Value (PPV) is the ratio of true positives to positives. It is a key measure of how a diagnostic method performs because it refers to the probability of positive results actually being true positive. The association between sensitivity and specificity is determined by plotting a ROC (Receiver Operating Characteristic) curve with sensitivity vs. 1 - specificity.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{3.1}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{3.2}$$

$$PPV = \frac{TP}{(TP + FP)} \tag{3.3}$$

The results and analyses of the experiments are presented in the following section.

3.3 Experimental Results and Analyses

The results for all the *Escherichia coli* and *Bacillus subtilis* mappings to the E.coli and Bacillus groups are determined based on the experiments described in the previous section. There are 169 transcription factors available for *Escherichia coli* and 317 transcription factors available for *Bacillus subtilis*, that are used for transcription factor mapping in this thesis. The sensitivity, specificity and PPV values are calculated for all these results and plotted on the graphs. This section illustrates the analyses of these results, based on sensitivity, PPV and ROC curve.

3.3.1 Analysis based on Sensitivity

The Sensitivity for *Escherichia coli* and *Bacillus subtilis* mappings using BLASTP is plotted in Figures 3.3-3.4, and using PSI-BLAST (maximum 5 iterations) in Figures 3.5-3.6, with E-value thresholds ranging from 10^{-5} to 10.

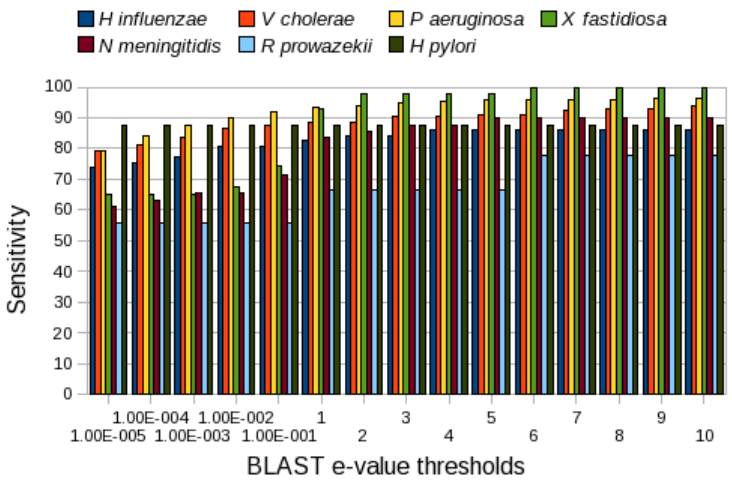
The results in Figure 3.3 show the comparison between the *Escherichia coli* and *Bacillus subtilis* mappings to the bacteria genomes from the E.coli group. The percentages for all the genomes increase in both graphs with the increase in the BLAST E-value threshold, since a higher E-value threshold leads to a higher number of BLAST hits or similar sequences in the target genome. Figure 3.3 illustrates that changing the source genome from *Escherichia coli* to *Bacillus subtilis* for finding transcription factors in the E.coli group does not produce much change in the sensitivity percentage. Therefore, the transcription factor mapping to the E.coli group is not significantly affected by the evolutionary distance between the chosen model organism and the bacteria genomes in the E.coli group.

The comparison between the *Bacillus subtilis* and *Escherichia coli* mappings to the bacteria genomes from the Bacillus group is presented in Figure 3.4. As in the previous mappings, the percentages increase in both graphs for all the genomes with

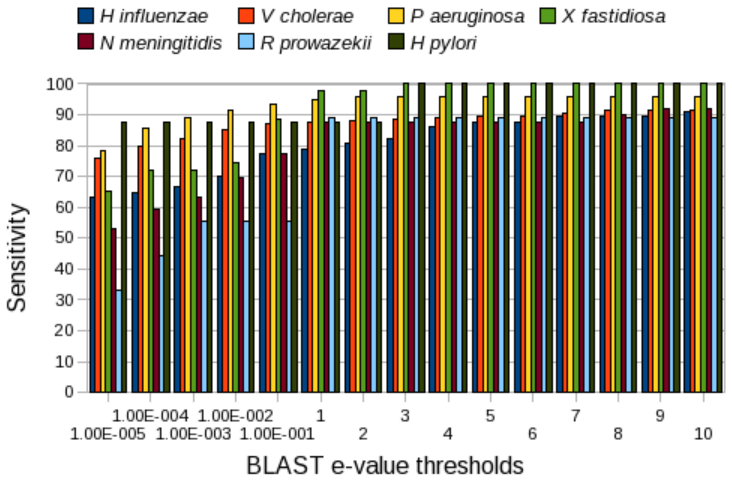
the increase in BLAST E-value threshold except that there is no mapping for low E-value thresholds shown between *Escherichia coli* and three Bacillus group genomes in Figure 3.4(b). Figure 3.4 illustrates that changing the source genome from *Bacillus subtilis* to *Escherichia coli* for finding transcription factors in the Bacillus group does show a large change in sensitivity, as opposed to the little change shown when mapping to the E.coli group. The percentage is much higher when *Bacillus subtilis* is used to map to the Bacillus group. This indicates that more transcription factors are predicted for the genomes of the Bacillus group when *Bacillus subtilis*, an evolutionarily closer model genome, is used for transcription factor mapping.

Sensitivity for PSI-BLAST is calculated and plotted in Figures 3.5-3.6 with E-value thresholds ranging from 10^{-5} to 10. As for using BLASTP, the sensitivity for all the genomes increases using PSI-BLAST as the number of hits increases with increase in the E-value threshold. The results in Figure 3.5 show that there is less than 10 percent change in the sensitivity when changing the source genome from *Escherichia coli* to *Bacillus subtilis* for mapping to the E.coli group. This implies that choosing an evolutionarily closer source genome does not change the mapping to the E.coli group. *Bacillus subtilis* and *Escherichia coli* mappings to the Bacillus group shown in Figure 3.6 illustrate that the sensitivity is higher for *Bacillus subtilis* mapping and there is no mapping between *Escherichia coli* and three Bacillus group genomes for low E-value thresholds. The PSI-BLAST results thus show the same characteristics as the BLASTP results.

Comparing Figure 3.3 and Figure 3.5 shows that the results using BLASTP and PSI-BLAST for mappings to the E.coli group do not demonstrate more than 10 percent difference in sensitivity with the change in tool type. However, there is more than 10 percent increase in sensitivity using PSI-BLAST over BLASTP for the mappings to the Bacillus group, as shown by the differences between Figure 3.4 and Figure 3.6. Therefore, the choice of using one of the BLAST tools does not make a difference for

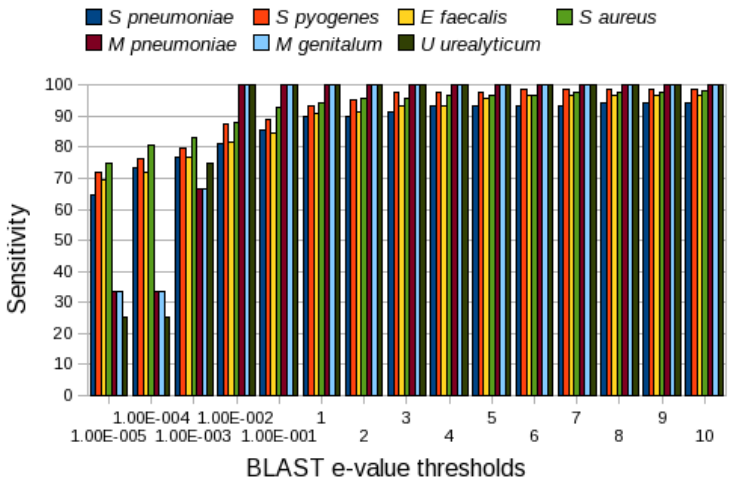


(a)

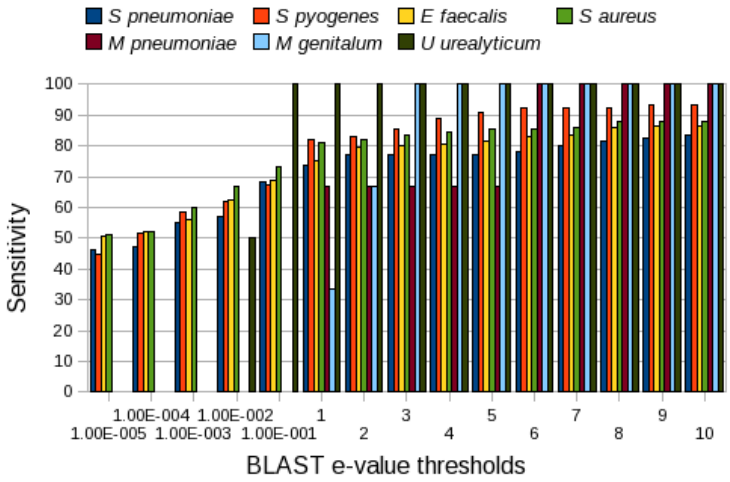


(b)

Figure 3.3: Sensitivity percentage for mapping to the E. coli group using BLASTP: (a) mapping from *Escherichia coli* and (b) mapping from *Bacillus subtilis*

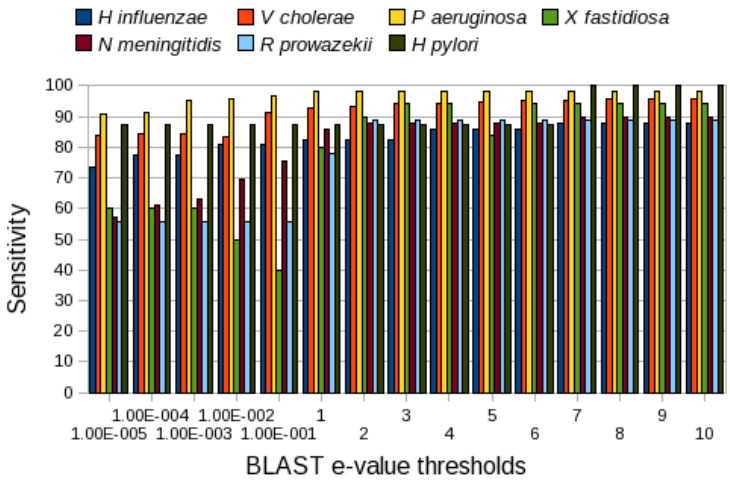


(a)

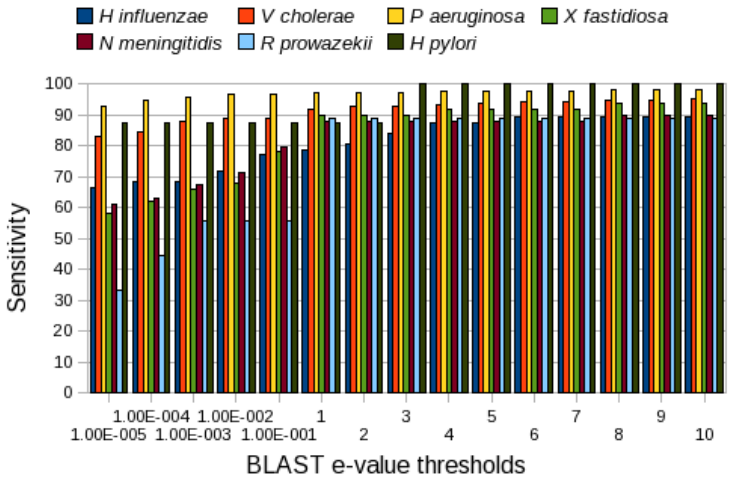


(b)

Figure 3.4: Sensitivity percentage for mapping to the Bacillus group using BLASTP: (a) mapping from *Bacillus subtilis* and (b) mapping from *Escherichia coli*

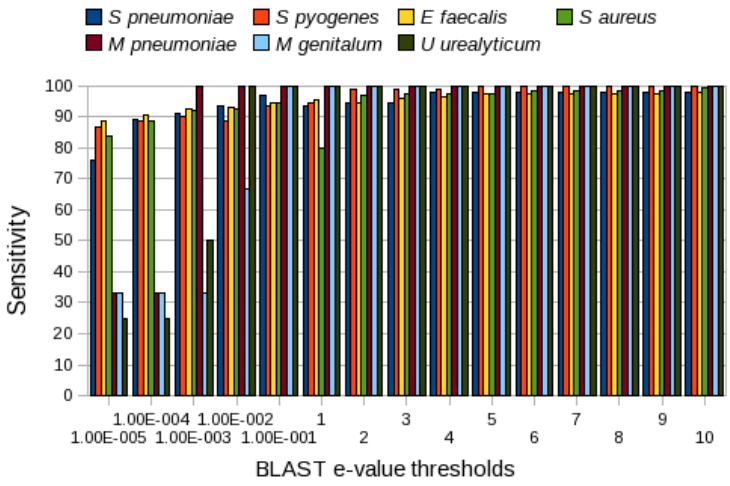


(a)

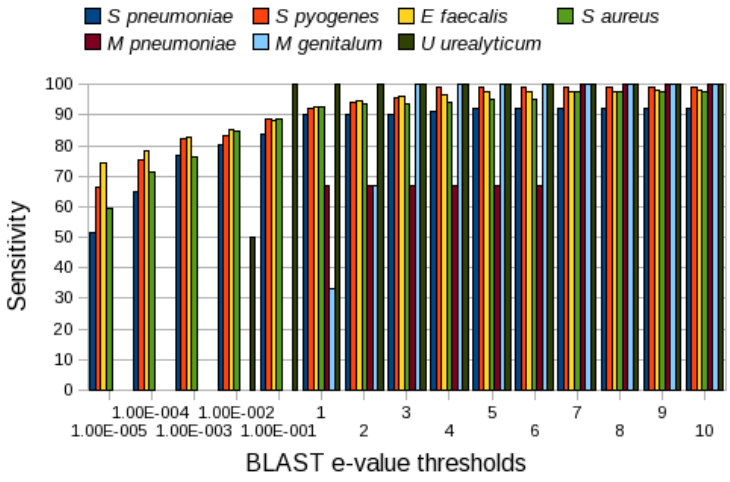


(b)

Figure 3.5: Sensitivity percentage for mapping to the E. coli group using PSI-BLAST: (a) mapping from *Escherichia coli* and (b) mapping from *Bacillus subtilis*



(a)



(b)

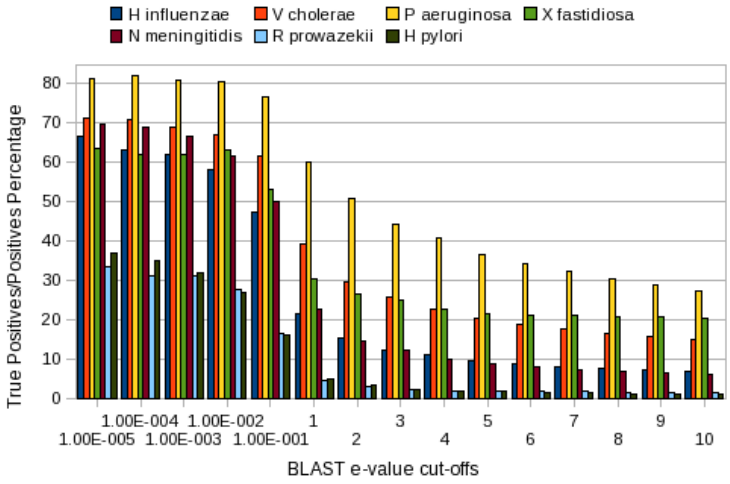
Figure 3.6: Sensitivity percentage for mapping to the Bacillus group using PSI-BLAST: (a) mapping from *Bacillus subtilis* and (b) mapping from *Escherichia coli*

mapping to the E.coli group, though it does matter for transcription factor mapping to the Bacillus group for which PSI-BLAST predicts more transcription factors.

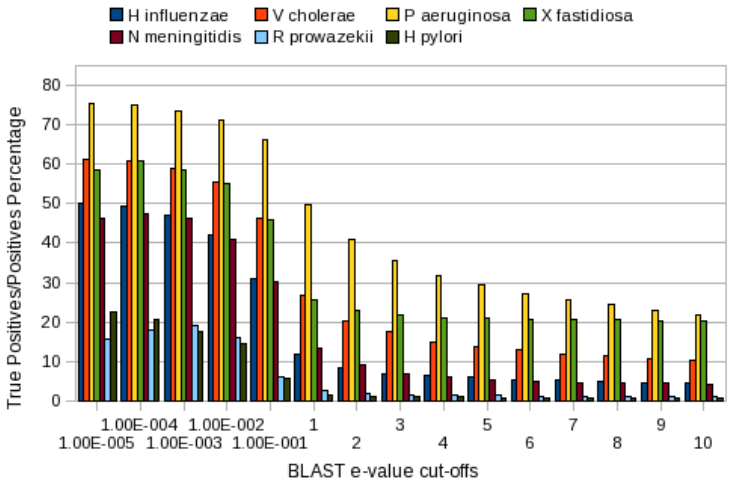
3.3.2 Analysis based on Positive Predictive Value (PPV)

The Positive Predictive Value (PPV) percentage is calculated and plotted in Figures 3.7-3.8 for BLASTP and in Figures 3.9-3.10 for PSI-BLAST using the tested range of E-value thresholds. This ratio represents the likely accuracy of the predicted factors. The results in Figure 3.7 illustrate the PPV percentages for the *Escherichia coli* and *Bacillus subtilis* mappings to the E.coli group using BLASTP. The percentage is a little higher for the *Escherichia coli* mapping than for the *Bacillus subtilis* mapping to the E.coli group, showing that not many false positives were found in the E.coli group when using *Escherichia coli* as a source genome, as compared to the higher number of false positives found using *Bacillus subtillis*. Therefore, for the E.coli group, using the closer model genome is able to ignore more of the similar sequences that are not transcription factors while still having high sensitivity.

The comparison between the PPV percentages of *Bacillus subtilis* and *Escherichia coli* mappings to the Bacillus group for BLASTP is illustrated in Figure 3.8. The percentage is higher for a few Bacillus group genomes when *Escherichia coli* is used, and there is very little or no mapping for the other genomes. When mapping to the other group of bacterial genomes, the false positives can be fewer, leading to a better PPV percentage. When using *Escherichia coli* and a low E-value threshold, some bacterial genomes have no true positives and a few false positives. The method finds more true transcription factors along with more false positives when using *Bacillis subtilis* as a source genome than by using *Escherichia coli*, potentially because *Bacillus subtilis* may have more closely related sequences in the Bacillus group that are not transcription factors. These sequences increase the false positives, decreasing the PPV percentage despite increasing the number of true positives found.

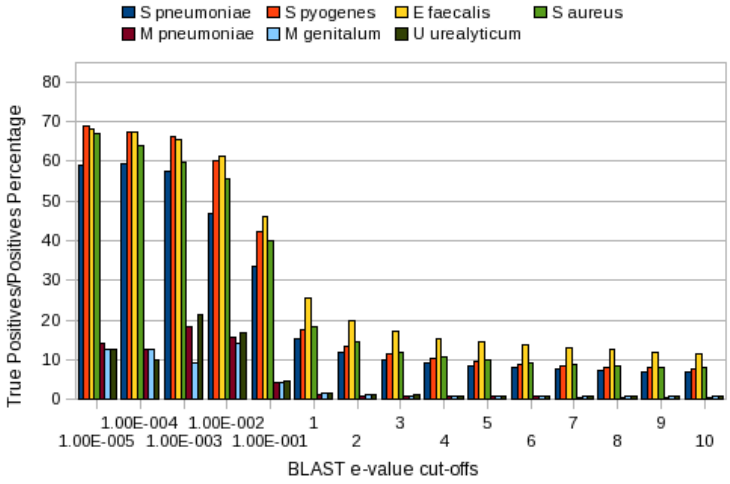


(a)

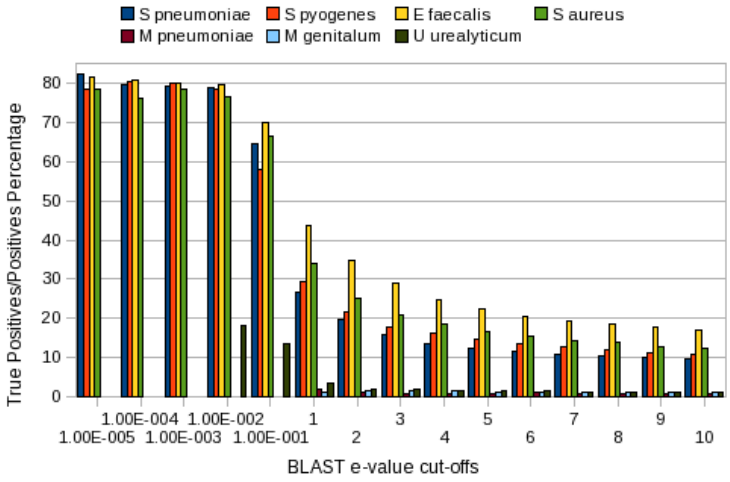


(b)

Figure 3.7: PPV percentage for mapping to the E. coli group using BLASTP: (a) mapping from *Escherichia coli* and (b) mapping from *Bacillus subtilis*



(a)



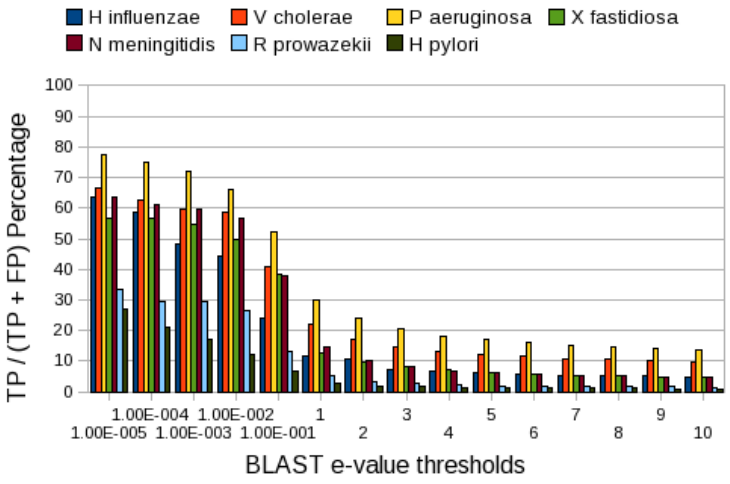
(b)

Figure 3.8: PPV percentage for mapping to the Bacillus group using BLASTP: (a) mapping from *Bacillus subtilis* and (b) mapping from *Escherichia coli*

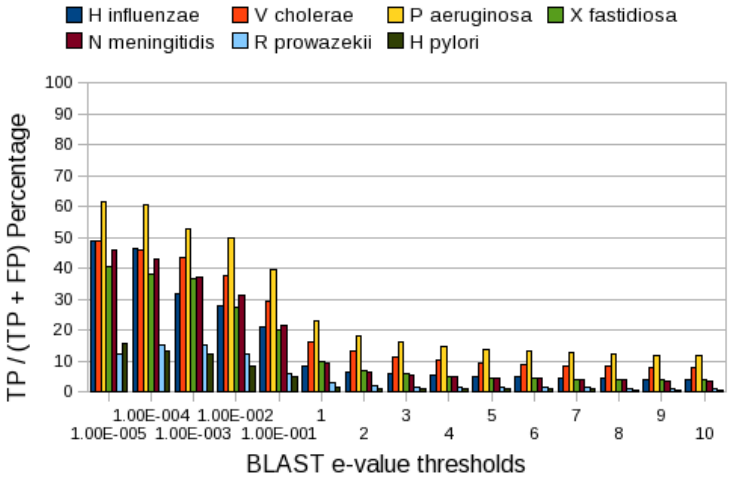
Figure 3.4(a) and 3.8(a) (*Bacillus subtilis* mapping to the Bacillus group) illustrate that for E-value thresholds ranging from 10^{-5} to 10^{-3}, the percentage of transcription factors that are found is lower, and the percentage of predicted factors that are true factors is higher. Both percentages increase and decrease, respectively, for higher E-value thresholds. Therefore, at lower E-value thresholds some of the true transcription factors are lost but many similar sequences that are not transcription factors are ignored. This type of trade-off is an expected effect of the variation in E-value threshold, and holds for all these experiments.

Similarly, the sensitivity is lower and the percentage of predicted positives that are true transcription factors is higher for E-value thresholds from 10^{-5} to 10^{-1}, as shown in Figure 3.4(b) and 3.8(b) (*Escherichia coli* mapping to the Bacillus group). For higher E-value thresholds, the sensitivity increases and the PPV decreases. Also, there are no correctly predicted transcription factors for three Bacillus group genomes for the E-value threshold range of 10^{-5} to 10^{-1}. These three bacterial genomes (*M. pneumoniae*, *M. genitalum*, *U. urealyticum*) are in a separate subgroup of the Bacillus group, evolutionarily farther away from *Bacillus subtilis* than the others in the Bacillus group, and their results are quite poor for the lower E-value thresholds when *Bacillus subtilis* is used. For higher E-value thresholds, the sensitivity is very high while the percentage of positives that are true is very low, reflecting the prediction of many distantly related sequences. While they should still be the same evolutionary distance from *Escherichia coli*, the differences between their subgroup and the rest of the Bacillus group greatly affects the transcription factors that are related to those in *Escherichia coli*.

The results of *Escherichia coli* and *Bacillus subtilis* mapping to the E.coli group (Figure 3.9), and to the Bacillus group (Figure 3.10) using PSI-BLAST show a decrease in the PPV percentage with the increase in the E-value thresholds. The percentage is higher for *Escherichia coli* mapping to the E.coli group, similar to

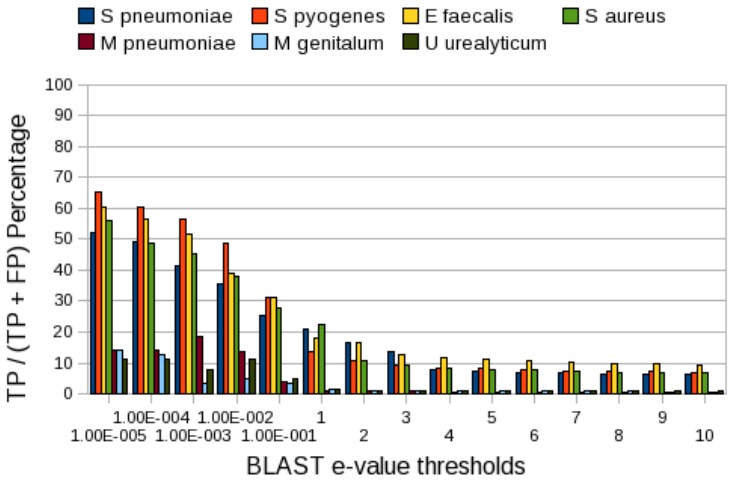


(a)

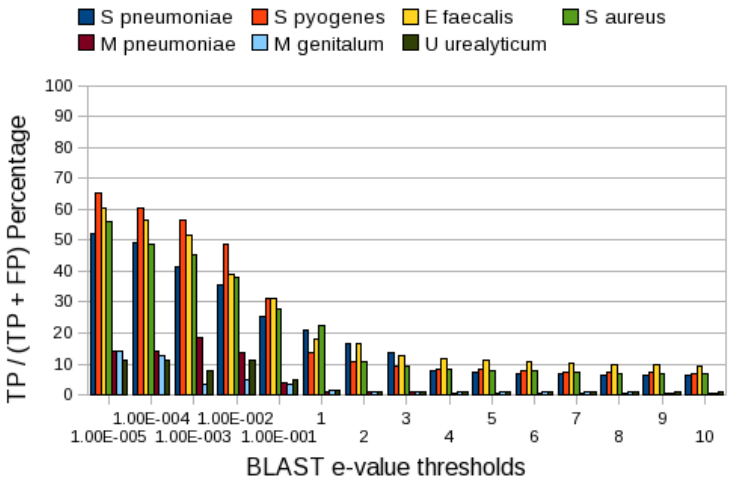


(b)

Figure 3.9: PPV percentage for mapping to the E. coli group using PSI-BLAST: (a) mapping from *Escherichia coli* and (b) mapping from *Bacillus subtilis*



(a)



(b)

Figure 3.10: PPV percentage for mapping to the Bacillus group using PSI-BLAST: (a) mapping from *Bacillus subtilis* and (b) mapping from *Escherichia coli*

the results from using BLASTP (Figure 3.7). More false positives are found using *Bacillus subtilis* than using *Escherichia coli* (the closer genome) as a source genome. For mapping to the Bacillus group, the *Escherichia coli* mapping is better for lower E-value thresholds but no mapping is found between three genomes of Bacillus group and *Escherichia coli*, similar to the results from BLASTP (Figure 3.8). There is a small increase in PPV percentage (accuracy) when using PSI-BLAST (Figure 3.9-3.10) compared to BLASTP (Figure 3.7-3.8).

3.3.3 Analysis based on ROC Curve

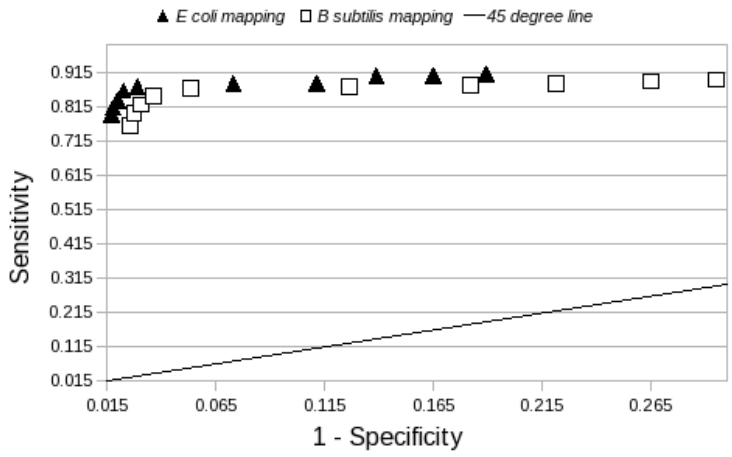
The best E-value threshold, based on the optimal trade-off between sensitivity and specificity, is determined by plotting a ROC (Receiver Operating Characteristic) curve, with Sensitivity versus 1 - Specificity plotted for each E-value threshold. An ideal point on the classification graph would be the upper left corner (0, 1), with no false negatives or false positives. The line $x = y$ at an angle of 45 degrees classifies the points on the graph into good (above the line) and bad (below the line). The more a point is above the line, the better the classification is, since the classification is less random [37]. The Point-Line (PLine) distance of the points (x, y) on the graph from the line $x = y$ is calculated using Equation 3.4, and the most suitable E-value threshold for a genome is selected based on the maximum PLine distance.

$$PLine(x, y) = \frac{(y - x)}{\sqrt{2}} \tag{3.4}$$

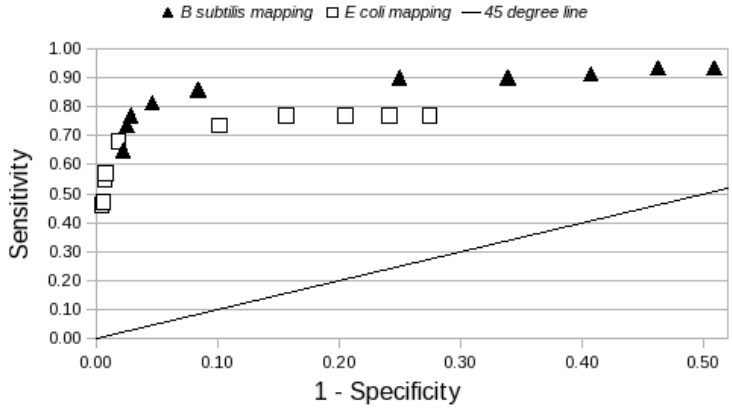
Sensitivity versus 1 - Specificity is plotted for the *Escherichia coli* and *Bacillus subtilis* mappings to *V. cholerae* (Figure 3.11(a)) and *S. pneumoniae* (Figure 3.11(b)), using BLASTP with E-value thresholds ranging from 10^{-5} to 5 (appearing from left to right). The corresponding values of the two graphs are shown in Tables 3.2 and 3.3 with the maximum PLine distance highlighted for both the mappings. The Sensitivity and 1 - Specificity values increase as the BLAST E-value threshold increases.

In Table 3.2, the PLine distance is higher for the *Escherichia coli* mapping to *V. cholerae* than for mapping from *Bacillus subtilis* when an E-value threshold of 0.1, the most suitable threshold based on the highest PLine distance, is used. However, the *Bacillus subtilis* mapping with an E-value threshold of 0.1 is better than the *Escherichia coli* mapping with lower and higher E-value thresholds, despite using the more distant model organism *Bacillus subtilis*. This illustrates that using an evolutionarily closer model and the most suitable E-value threshold are both important for predicting transcription factors in the E.coli group; it especially shows the importance of using the most suitable E-value threshold, since using a less suitable threshold with a closer model gives poorer results than using the most suitable threshold with a more distant model. Many of these poorer thresholds are still well within the range of $10^{-5} < E < 1$ suggested for finding homologues [54]. In Table 3.3, the PLine distance is higher for the *Bacillus subtilis* mapping to *S. pneumoniae* than for mapping from *Escherichia coli* with the most suitable E-value threshold of 0.1. Using a threshold of 10^{-5}, or of 1 or greater, for the *Bacillus subtilis* mapping produces a PLine distance less than that of the Escherichia coli mapping with an E-value threshold of 0.1. This situation shows the importance of choosing both an evolutionarily close model organism and a suitable E-value threshold to predict transcription factors in the Bacillus group.

Figure 3.12 shows Sensitivity versus 1 - Specificity plotted for the *Escherichia coli* and *Bacillus subtilis* mappings to *V. cholerae* and *S. pneumoniae* using PSI-BLAST with E-value thresholds ranging from 10^{-5} to 5 (appearing from left to right). Tables 3.4 and 3.5 show the corresponding values of the two graphs with the maximum PLine distance highlighted. With the increase in BLAST E-value threshold, the Sensitivity and 1 - Specificity values increases. The results from PSI-BLAST in Tables 3.4 and 3.5 are analogous to the BLASTP results in Tables 3.2 and 3.3. The PLine distance with the most suitable E-value thresholds is higher for the *Escherichia coli* mapping



(a)



(b)

Figure 3.11: Sensitivity vs. (1 - Specificity) for the *Escherichia coli* and *Bacillus subtilis* mapping using BLASTP: (a) mapping to *V. cholerae* and (b) mapping to *S. pneumoniae*

Table 3.2: Sensitivity and 1 - Specificity for *Escherichia coli* and *Bacillus subtilis* mapping to *V cholerae* using BLASTP

E-value threshold	***Escherichia coli* mapping**			***Bacillus subtilis* mapping**		
	1 - Spec	**Sens**	**PLine**	**1 - Spec**	**Sens**	**PLine**
0.00001	0.0170	0.7927	0.5485	0.0255	0.7617	0.5205
0.0001	0.0178	0.8135	0.5626	0.0275	0.7979	0.5448
0.001	0.0200	0.8342	0.5757	0.0305	0.8238	0.5610
0.01	0.0225	0.8653	0.5959	0.0362	0.8497	0.5752
0.1	0.0288	0.8756	0.5988	0.0533	0.8705	0.5778
1	0.0730	i0.8860	0.5749	0.1263	0.8756	0.5299
2	0.1112	0.8860	0.5479	0.1818	0.8808	0.4943
3	0.1387	0.9067	0.5431	0.2210	0.8860	0.4702
4	0.1647	0.9067	0.5247	0.2647	0.8912	0.4430
5	0.1889	0.9119	0.5112	0.2946	0.8964	0.4255

Table 3.3: Sensitivity and 1 - Specificity for *Bacillus subtilis* and *Escherichia coli* mapping to *S pneumoniae* using BLASTP

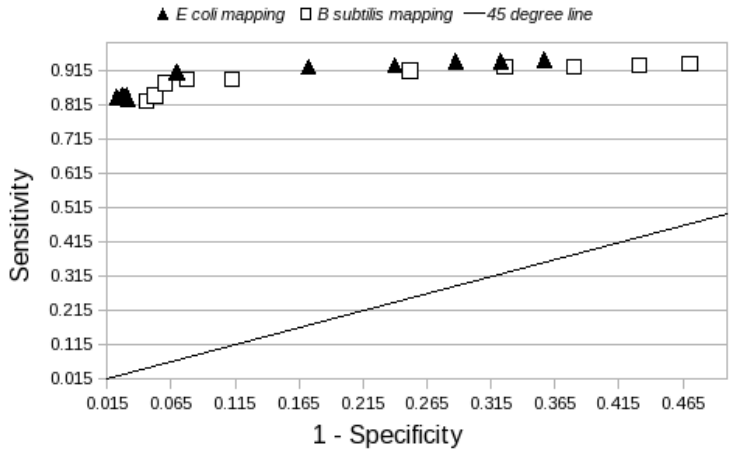
E-value threshold	***Bacillus subtilis* mapping**			***Escherichia coli* mapping**		
	1 - Spec	**Sens**	**PLine**	**1 - Spec**	**Sens**	**PLine**
0.00001	0.0225	0.6484	0.4426	0.0049	0.4615	0.3229
0.0001	0.0252	0.7363	0.5028	0.0060	0.4725	0.3299
0.001	0.0285	0.7692	0.5238	0.0071	0.5495	0.3835
0.01	0.0461	0.8132	0.5424	0.0077	0.5714	0.3986
0.1	0.0845	0.8571	0.5464	0.0187	0.6813	0.4686
1	0.2501	0.9011	0.4603	0.1015	0.7363	0.4489
2	0.3390	0.9011	0.3975	0.1569	0.7692	0.4330
3	0.4076	0.9121	0.3567	0.2052	0.7692	0.3989
4	0.4630	0.9341	0.3331	0.2419	0.7692	0.3729
5	0.5096	0.9341	0.3001	0.2748	0.7692	0.3496

Table 3.4: Sensitivity and 1 - Specificity for *Escherichia coli* and *Bacillus subtilis* mapping to *V. cholerae* using PSI-BLAST

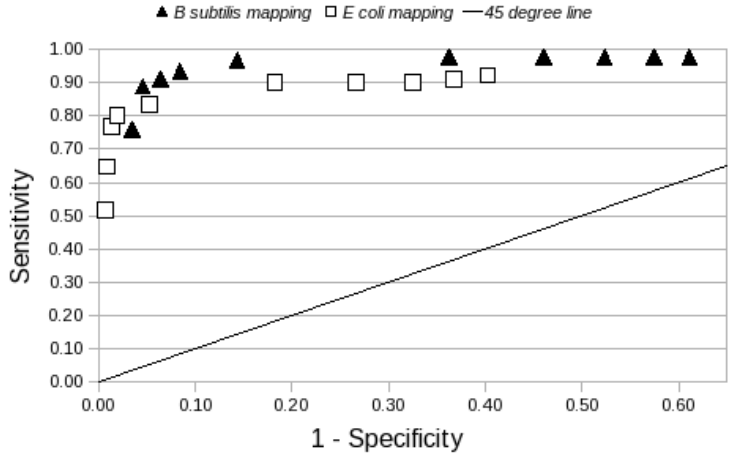
E-value threshold	***Escherichia coli* mapping**			***Bacillus subtilis* mapping**		
	1 - Spec	**Sens**	**PLine**	**1 - Spec**	**Sens**	**PLine**
0.00001	0.0225	0.8394	0.5776	0.0456	0.8290	0.5540
0.0001	0.0269	0.8446	0.5782	0.0524	0.8446	0.5601
0.001	0.0302	0.8446	0.5758	0.0604	0.8808	0.5801
0.01	0.0310	0.8342	0.5679	0.0774	0.8912	0.5754
0.1	0.0697	0.9119	0.5955	0.1129	0.8912	0.5504
1	0.1722	0.9275	0.5341	0.2518	0.9171	0.4704
2	0.2400	0.9326	0.4898	0.3259	0.9275	0.4254
3	0.2872	0.9430	0.4637	0.3800	0.9275	0.3871
4	0.3226	0.9430	0.4387	0.4308	0.9326	0.3549
5	0.3567	0.9482	0.4183	0.4703	0.9378	0.3306

to *V. cholerae* than for mapping from *Bacillus subtilis*, and is also higher for the *Bacillus subtilis* mapping to *S. pneumoniae* than for mapping from *Escherichia coli*. This illustrates that no matter which BLAST tool version is used, the results show the importance of choosing both source organism and a suitable E-value threshold to predict transcription factors.

Sensitivity versus 1 - Specificity is plotted for the *Escherichia coli* mapping to *V. cholerae* (Figure 3.13(a)) and *Bacillus subtilis* mapping to *S. pneumoniae* (Figure 3.13(b)) using BLASTP and PSI-BLAST, with E-value thresholds ranging from 10^{-5} to 5 (appearing from left to right). The PLine distance is higher using BLASTP than PSI-BLAST for *Escherichia coli* mapping to *V. cholerae* (E.coli group) but lower using BLASTP for *Bacillus subtilis* mapping to *S. pneumoniae* at the most suitable E-value threshold. These results illustrate that BLASTP is a better tool than PSI-BLAST for transcription factor mapping to the E.coli group whereas PSI-BLAST works better for mapping to the Bacillus group.

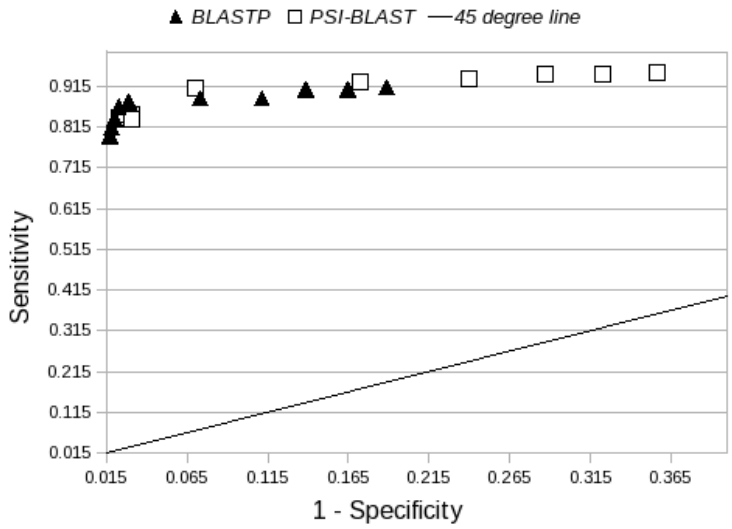


(a)

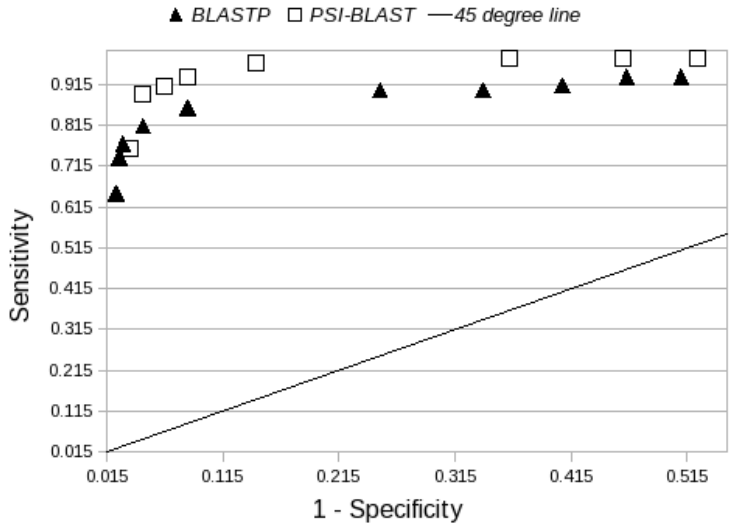


(b)

Figure 3.12: Sensitivity vs. (1 - Specificity) for the *Escherichia coli* and *Bacillus subtilis* mapping using PSI-BLAST: (a) mapping to *V. cholerae* and (b) mapping to *S. pneumoniae*



(a)



(b)

Figure 3.13: Sensitivity vs. (1 - Specificity) for mapping using BLASTP and PSI-BLAST: (a) *Escherichia coli*mapping to *V. cholerae* and (b) *Bacillus subtilis* mapping to *S. pneumoniae*

Table 3.5: Sensitivity and 1 - Specificity for *Bacillus subtilis* and *Escherichia coli* mapping to *S. pneumoniae* using PSI-BLAST

E-value threshold	***Bacillus subtilis* mapping**			***Escherichia coli* mapping**		
	1 - Spec	**Sens**	**PLine**	**1 - Spec**	**Sens**	**PLine**
0.00001	0.0346	0.7582	0.5117	0.0066	0.5165	0.3606
0.0001	0.0455	0.8901	0.5972	0.0082	0.6484	0.4526
0.001	0.0642	0.9121	0.5996	0.0137	0.7692	0.5342
0.01	0.0839	0.9341	0.6011	0.0192	0.8022	0.5537
0.1	0.1432	0.9670	0.5826	0.0527	0.8352	0.5533
1	0.1783	0.9341	0.5344	0.1821	0.9011	0.5084
2	0.2414	0.9451	0.4976	0.2666	0.9011	0.4487
3	0.2946	0.9451	0.4600	0.3247	0.9011	0.4075
4	0.5749	0.9780	0.2851	0.3670	0.9121	0.3855
5	0.6111	0.9780	0.2595	0.4021	0.9231	0.3684

3.3.4 Overall Analysis

The best E-value threshold, based on the highest PLine distance, is determined for all the mappings from *Escherichia coli* and *Bacillus subtilis* to the E.coli and Bacillus groups using BLASTP and PSI-BLAST as shown in Tables 3.6 and 3.7. From these BLAST E-value thresholds, 0.001, 0.01 and 0.1 are the most common best thresholds for mapping to the E.coli group. In addition, 0.1 and 1 are the most common best thresholds for mapping to the Bacillus group. Above and below these thresholds the PLine distance decreases. The important elements for transcription factor mapping, which are the best source genome, E-value threshold and BLAST version to use, where best is chosen based on the PLine distance, are shown in Table 3.8 for all the bacterial genomes of the E.coli and Bacillus groups. According to these results, *Escherichia coli* is the most common source genome and BLASTP is the better version of BLAST for transcription factor mapping to the E.coli group. For transcription factor mapping to the Bacillus group, *Bacillus subtilis* is the most common source genome with PSI-BLAST as the better version of BLAST. The

Table 3.6: Best Suitable E-value Threshold for mapping to E.coli Group

Bacteria genomes	**Best BLAST E-value threshold**			
	***Escherichia coli* mapping**		***Bacillus subtilis* mapping**	
	BLASTP	**PSI-BLAST**	**BLASTP**	**PSI-BLAST**
H. influenzae	0.01	0.01	0.1	0.1
V. cholerae	0.1	0.1	0.1	0.001
P. aeruginosa	0.1	0.001	0.1	0.0001
X. fastidiosa	2	0.1	3	0.1
N. meningitidis	1	1	1	0.1
R. prowazekii	0.00001	1	1	1
H. pylori	0.00001	0.00001	0.00001	0.00001

results based on one or a combination of two elements do not give the best result with highest PLine distance, as the results vary based on the elements not taken into consideration. For example, using BLASTP is better than PSI-BLAST when *Escherichia coli* is used as a source genome and the most suitable E-value threshold is used (Figure 3.13(a)), but PSI-BLAST gives better results with the most suitable E-value threshold compared to the results using BLASTP with higher or lower E-value thresholds. Also, changing the source genomes changes the PLine distance, so the better one needs to be chosen.

For mapping to *V. cholerae* (E.coli group), *Escherichia coli*, the most suitable source genome, 0.1, the most suitable E-value threshold, and BLASTP, the better version of BLAST, is the best mix that corresponds to the highest PLine distance, leading to the most efficient mapping with high true positives and low false positives. The combination of *Bacillus subtilis* as the source genome, 0.01 as the E-value threshold and PSI-BLAST as the BLAST version gives the highest PLine distance for mapping to *S. pneumoniae* (Bacillus group). Therefore, choosing the better source genome, better BLAST version and the best E-value threshold is critical in transcription

Table 3.7: Best Suitable E-value Threshold for mapping to Bacillus Group

Bacteria genome	**Best BLAST E-value threshold**			
	***Escherichia coli* mapping**		***Bacillus subtilis* mapping**	
	BLASTP	**PSI-BLAST**	**BLASTP**	**PSI-BLAST**
S. aureus	0.1	0.001	1	0.1
E. faecalis	0.01	0.001	2	0.01
S. pneumoniae	0.1	0.01	0.1	0.01
S. pyogenes	0.01	0.001	1	0.1
U. urealyticum	0.01	0.01	0.1	0.1
M. pneumoniae	0.01 (1.0E-02)	0.001 (1.0E-03)	6	1
M. genitalum	0.01	0.1	3	3

Table 3.8: Best suitable source genome, E-value threshold and BLAST version for E. coli and Bacillus group genomes

		Source Genome	**E-value threshold**	**BLAST version**
E. coli group genomes	*H. influenzae*	*E. coli*	0.01	BLASTP
	V. cholerae	*E. coli*	0.1	BLASTP
	P. aeruginosa	*E. coli*	0.001	PSI-BLAST
	X. fastidiosa	*B. subtilis*	3	BLASTP
	N. meningitidis	*E. coli*	1	BLASTP
	R. prowazekii	*E. coli*	1	PSI-BLAST
	H. pylori	*E. coli*	0.00001	BLASTP
Bacillus group genomes	*S. aureus*	*B. subtilis*	0.001	PSI-BLAST
	E. faecalis	*B. subtilis*	0.001	PSI-BLAST
	S. pneumoniae	*B. subtilis*	0.01	PSI-BLAST
	S. pyogenes	*B. subtilis*	0.001	PSI-BLAST
	U. urealyticum	*E. coli*	0.01	BLASTP
	M. pneumoniae	*B. subtilis*	0.001	PSI-BLAST
	M. genitalum	*E. coli*	0.01	BLASTP

factor mapping between bacterial genomes.

For suitable E-value thresholds, the PPV ratio (accuracy) is high as shown in Figures 3.7-3.10, along with a high sensitivity as shown in Figures 3.3-3.6. Lower and higher E-value thresholds lose either accuracy or sensitivity. There is a sudden decrease in accuracy at higher E-value thresholds, as more proteins that are not transcription factors are predicted, while there is a sudden decrease in sensitivity for lower E-value thresholds, losing many true transcription factors. Many transcription factor sequences will be lost irrespective of the chosen E-value threshold because not all sequences in one genome are related to sequences in the other genome, even for closely related organisms. Also, there are sequences in the target genomes that are similar to a particular transcription factor but are not confirmed to be a transcription factor. These sequences probably contain the same protein domains, but are not confirmed factors and thus add to the false positive count.

3.4 Conclusion

Based on the analysis of the results of transcription factor mapping, BLAST versions BLASTP and PSI-BLAST are able to map most transcription factors from one bacterial genome to another. This success indicates that transcription factor sequence motifs are preserved well among bacterial genomes, and transcription factors can be found for newly sequenced bacterial genomes using a related model bacterial organism as a source genome. The analysis also determines the appropriate version of BLAST and the most suitable BLAST E-value parameters for mapping transcription factors between bacterial genomes, with accuracy decreasing significantly for higher thresholds and sensitivity decreasing significantly for lower thresholds. Using an evolutionarily close model organism as a source genome and the more suitable version of BLAST improves transcription factor mapping, while a proper choice of E-value

threshold also significantly improves the transcription factor prediction quality, often more than the selection of the closer model organism and the better BLAST version. The right combination of source genome, E-value threshold and BLAST version gives the best result with high sensitivity and PPV ratio. These results also illustrate that BLASTP is the better version of BLAST for transcription factor mapping to the E.coli group whereas PSI-BLAST is a better version of BLAST for mapping transcription factors to the Bacillus group. According to the phylogenetic tree [48] of these species, *Escherichia coli* is closer to the E.coli group, being part of the same branch as all the other genomes except *H. pylori*, while *Bacillus subtilis* is farther away from the Bacillus group, this result is not surprising. The differences in the tool choice between the groups is likely due to the different characteristics of the tools, since PSI-BLAST is better at detecting more distant evolutionary relationships such as those found in the Bacillus group.

Chapter 4

Transcription Factor Mapping to Plant Genomes

4.1 Overview

Chapter 3 described the transcription factor mapping methods and analyses for bacterial genomes. Bacteria are a large domain of prokaryotic microorganisms that do not contain a nucleus whereas plants are complex eukaryotic genomes that do contain a nucleus. Therefore, this chapter [60, 63] [1] [2] presents the transcription factor mapping methods used for mapping to plant genomes. As explained earlier, the transcription factors available for the source genome are used to identify the transcription factors of the target genome. Mapping transcription factors from one genome to another involves finding similar protein sequences in the target genome performing similar functions. This is the first step towards mapping a regulatory link from a source genome to a target genome. This would be followed by target gene mapping in Chapter 5 and integration of the predicted transcription factors

[1] Sharma, R.; Evans, P.A.; Bhavsar, V. C., Regulatory link mapping between organisms, BMC Systems Biology 2011, 5(Suppl 1): S4.

[2] Sharma, R.; Evans, P.A.; Bhavsar, V. C., Mapping a Regulatory Network Between Organisms, Proceedings of the Fourth International Conference on Computational Systems Biology 2010, 204-218.

and target genes in Chapter 6 to complete the process of mapping a regulatory link from one genome to another.

Three methods are used in this chapter for identification of transcription factors based on sequence similarity, same protein family and subfamily classification using BLAST [5] and InterProScan [78]. Various experiments have been performed to test and compare these methods. The results and analysis of these experiments are presented here. Yeast and *Arabididopsis thaliana* are used as the source and target genomes, respectively.

4.2 Transcription Factor Mapping to *Arabidopsis thaliana*

This section explains the three methods for the identification of transcription factors and the experiments conducted for the analysis of their results.

4.2.1 Methods and their Comparison

The sequence alignment tool BLAST [5] and the functional similarity tool InterProScan [78] are used for transcription factor mapping, which is the first step in regulatory link mapping. Similar regions in the gene sequences tend to indicate similar structure or function preserved by evolution. Additionally, specific conserved motifs in the protein sequences, called protein signatures, define the structure and function of the proteins. Therefore, similar protein sequences with common protein signatures generally perform similar functions and belong to the same functional group. As described in Section 2.6, InterPro [44] is a non-redundant database that integrates the commonly used protein signature databases. The PANTHER (Protein ANalysis THrough Evolutionary Relationships) database [67] and its corresponding scanning tools, BLAST and hmmsearch, are used within InterProScan. The PAN-

THER database consists of protein sequences classified into families and subfamilies with similar function based on published experimental evidence and evolutionary relationships.

The experiment involves comparing the three methods named as TF_Seq, TF_Fam and TF_SubFam, respectively to recognize three sets of transcription factors for the *Arabididopsis thaliana* genome using Yeast (*Saccharomyces cerevisiae*) as the source genome. The three methods for mapping transcription factors are explained below.

TF_Seq uses the BLAST tool, with its e-value cut-off set to 0.1 to get sequences in the target genome with high sequence similarity to the transcription factors of the source genome.

TF_Fam uses the BLAST tool as in TF_Seq followed by the InterProScan tool to refine the BLAST results based on similar family protein signature motifs.

TF_SubFam uses the BLAST tool and InterProScan tool as in TF_Fam but refines the results even further based on similar sub-family protein signature motifs.

Yeast is a small unicellular fungus, while *Arabidopsis thaliana* is a plant species. Yeast also has a much smaller genome of about 12,495,682 bases, compared to *Arabidopsis thaliana* with about 115,409,949 base pairs [65]. This makes it necessary to perform one to many mapping of transcription factors between them as one transcription factor in the Yeast genome can be similar to more than one transcription factor in the *Arabidopsis thaliana* genome. Additionally, there can be many to one mapping of transcription factors as more than one transcription factor of the Yeast genome can be similar to one particular transcription factor of the *Arabidopsis thaliana* genome. All the transcription factors of Yeast will not be mapped to *Arabidopsis thaliana* as all pairs of species have some significant differences between their transcription factors. The numbers of mappings of transcription factors from

Yeast to *Arabidopsis thaliana*, using the three methods given above, are shown in Table 4.1. The number of mappings decreases from 1344 for TF_Seq, to 903 for TF_Fam, and further decreases to 212 for TF_SubFam.

Table 4.1: Number of mappings of transcription factors (TFs) from Yeast genome to *Arabidopsis thaliana* genome using the three methods

Method	Number of mappings
TF_Seq	1344
TF_Fam	903
TF_SubFam	212

The numbers of distinct (non-repetitive) transcription factors identified by each of the three methods are shown in Table 4.2. The three sets of predicted transcription factors are: TFbl, TFf and TFsf. The number of distinct transcription factors identified decreases from 767 to 118 from set TFbl to set TFsf.

Table 4.2: Number of sequences identified as transcription factors for *Arabidopsis thaliana* using the three methods

Method	Number of sequences identified as transcription factors
TF_Seq: Result set TFbl	767
TF_Fam: Result set TFf	434
TF_SubFam: Result set TFsf	118

The results obtained in this experiment in Table 4.2 are analyzed to test and compare the results of the three sets of transcription factors identified by using the three methods. Not all the transcription factors can be mapped from one genome to another genome, even when these genomes are evolutionarily very close. Moreover, there is no knowledge of the number of transcription factors that should be mapped from one genome to another. Hence, the number of confirmed transcription factor

mappings is based on the definition of correct mapping used. Each of the results from the method are analyzed to find out if a sequence predicted as a transcription factor is an experimentally verified transcription factor of the target genome or not. Therefore, the results are analyzed by comparing the 1922 available transcription factors of *Arabidopsis thaliana* [24] with the three predicted sets. There are 35351 sequences in the nucleotide database of *Arabidopsis thaliana* genome used as the BLAST database in this experiment to search for transcription factors.

The three methods run in this experiment classify the identified result sequences in the target genome as positives for being a transcription factor, and the rest of the sequences as negative for not being a transcription factor. Binary classifier [45], as explained earlier in Chapter 3, involves comparing the predicted results to the actual results and classifying the predicted results into four groups as described below: true positives, false positives, true negatives and false negatives.

True Positives (TP) are calculated as the number of sequences out of the total sequences predicted as transcription factors that are present in the available transcription factor database of *Arabidopsis thaliana* [24].

False Positives (FP) are calculated as the number of sequences out of the total sequences predicted as transcription factors that are not present in the available transcription factor database.

True Negatives (TN) are the number of sequences that are discarded by a method as not a transcription factor and are not present in the available transcription factor database.

False Negatives (FN) are the number of sequences that are discarded by a method as not a transcription factor but are present in the available transcription factor database.

To analyze the predicted results of the experiment, true positives, false positives, true negatives and false negatives for the three sets of identified transcription factors are calculated as illustrated in Table 4.3 and Figure 4.1. There is a large decrease of 261 in false positives from 367 to 106 comparing set TFbl to set TFf and a low decrease of 72 in true positives from 400 to 328. Therefore, set TFf has a better result than set TFbl as there is only an 18 percent decrease in true positives from set TFbl to set TFf with a huge decline of 71.1 percent in false positives. There is also a low decline in percentage of available transcription factors correctly identified in result sets TFbl and TFf from 20.8 percent to 17.1 percent. Comparing result sets TFf and TFsf, there is a huge decline of 237 in true positives from 328 to 91 and a decrease of 79 in false positives from 106 to 27. Therefore, with a huge decline of 72.3 percent and 74.5 percent in true positives and false positives respectively and a huge decline in percentage of available transcription factors correctly predicted from 17.1 percent to 4.7, set TFf has a better result over set TFsf by not losing many true positives. True negatives in all the three result sets are very high because most of the sequences that are not transcription factors have very low similarity to the transcription factor sequences.

Table 4.3: True positive, false positive, false negative and true negative values for transcription factors identified in the three result sets

	TFbl	TFf	TFsf
True Positives	400	328	91
False Positives	367	106	27
True Negatives	33062	33323	33402
False Negatives	1522	1594	1831

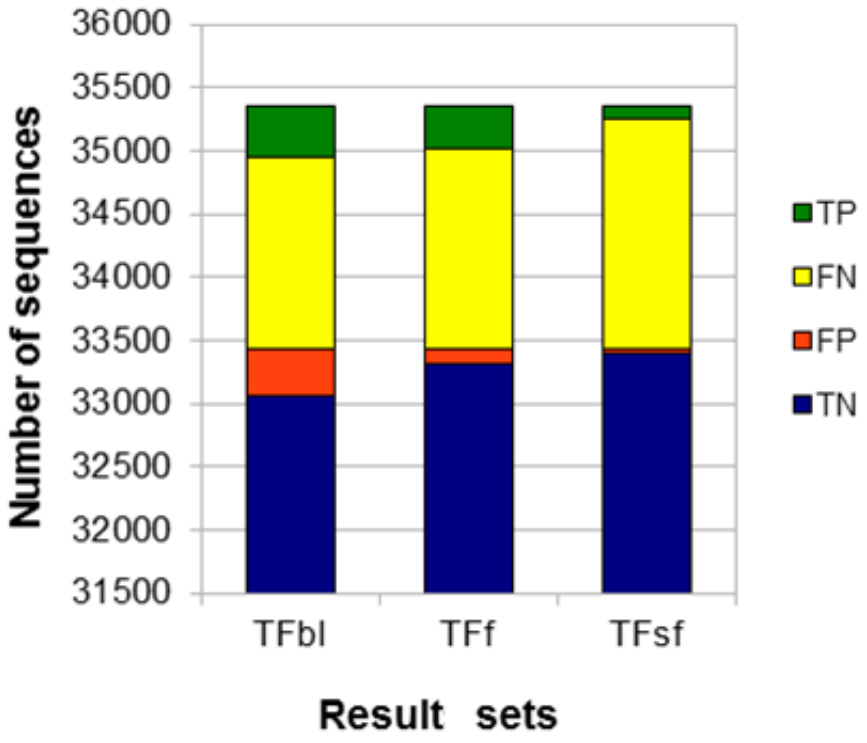


Figure 4.1: True positive, false positive, false negative and true negative values for transcription factors identified in the three result sets

4.2.2 Identification of a suitable BLAST parameter threshold

4.2.2.1 TF_Seq method

The purpose of this experiment is to estimate the most suitable BLAST parameter e-value cut-off for the BLAST tool used in the TF_Seq method for identification of transcription factors for *Arabidopsis thaliana* genome using Yeast as the source genome. This experiment is performed by comparing the available transcription factor data of *Arabidopsis thaliana* genome consisting of 1922 transcription factors to predicted transcription factors for e-value cut-offs ranging from 0.01 to 10. There are 35351 sequences in the nucleotide database of *Arabidopsis thaliana* used as BLAST database in this experiment. The number of mapped links of transcription factors and number of distinct transcription factors identified in this experiment are shown in Table 4.4. There is almost a constant increase in the number of mapped links and transcription factors from BLAST e-value cut-offs of 0.01 to 10, except for a steep increase of 647 and 378, respectively from e-value cut-off of 0.1 to 1.

To analyze the above results, binary classification is used and the comparison consisting of true positives, false positives, true negatives and false negatives is shown in Table 4.5 and Figure 4.2. There is a large increase of 71 and 122 in the beginning in true positives from e-value cut-off of 0.01 to 0.1 and 0.1 to 1, respectively. From e-value cut-off of 1 to 10, there is a gradual increase in true positives from 522 to 701. For false positives, there is almost a steady increase throughout, except a large increase of 256 and 178 from e-value cut-off of 0.1 to 1 and 1 to 2, respectively.

The value of the ratio of false positives to false negatives as e-value cut-off increases from 0.01 to 10 is shown in Figure 4.2. There is a sudden increase in the false positives to false negatives ratio from e-value cut-off of 0.1 to 1. To further analyze, and to evaluate the performance of a binary classifier, values of sensitivity and specificity

Table 4.4: Number of mapped links and number of sequences identified as transcription factors using TF_Seq for BLAST e-value ranging from 0.01 to 10

BLAST e-value cut-off	Number of mapped links	Number of sequences identified as transcription factors
0.01	1030	588
0.1	1344	767
1	1991	1145
2	2300	1369
3	2552	1531
4	2744	1658
5	2934	1778
6	3130	1922
7	3280	2016
8	3438	2135
9	3568	2241
10	3740	2365

Table 4.5: True positive, false positive, false negative and true negative values for transcription factors identified in TF_Seq for different e-value cut-offs

BLAST e-value cut-off	True positives	False positives	True negatives	False negatives
0.01	329	259	33170	1593
0.1	400	367	33062	1522
1	522	623	32806	1400
2	568	801	32628	1354
3	598	933	32496	1324
4	621	1037	32392	1301
5	634	1144	32285	1288
6	650	1272	32157	1272
7	655	1361	32068	1267
8	672	1463	31966	1250
9	681	1560	31869	1241
10	701	1664	31765	1221

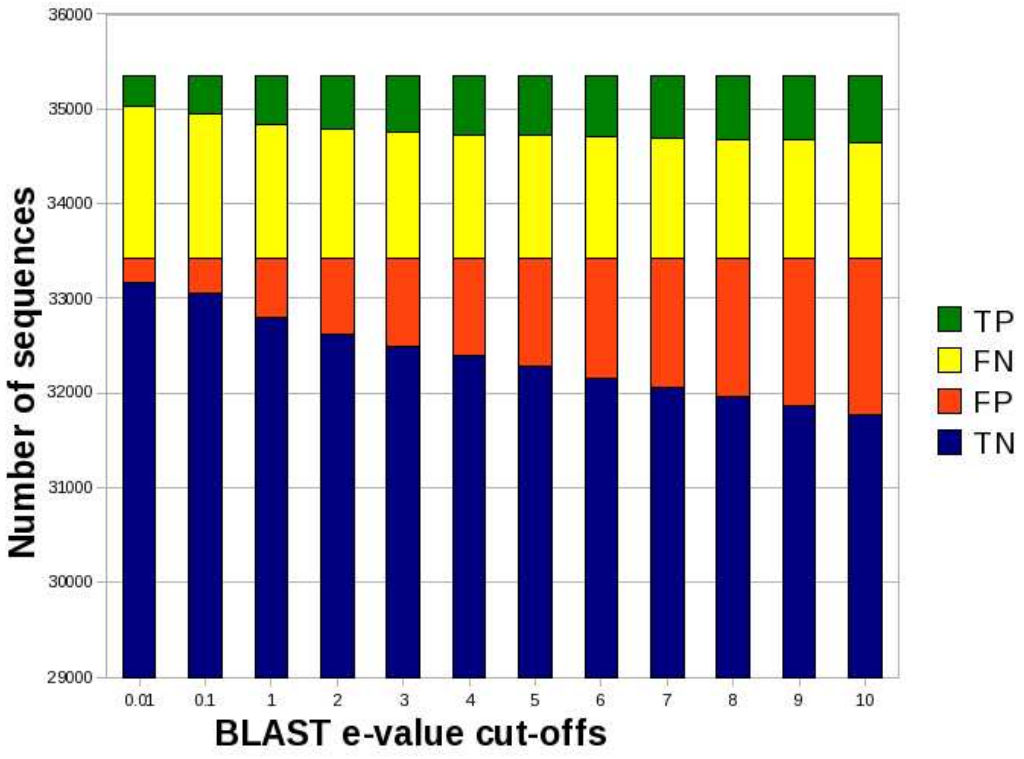


Figure 4.2: True positive, false positive, false negative and true negative values for transcription factors identified in TFmethod TF_Seq for different e-value cut-offs

are calculated using Equation 4.1 and 4.2. Sensitivity is measured as the proportion of positives that are true and specificity is measured as the proportion of negatives that are true.

$$Sensitivity = TP/(TP + FN) \tag{4.1}$$

$$Specificity = TN/(TN + FP) \tag{4.2}$$

The association between sensitivity and specificity is determined by plotting a ROC (Receiver Operating Characteristic) curve with sensitivity vs. 1 - specificity as shown in Figure 4.3 with corresponding BLAST e-value cut-offs marked on the points on the graph. Sensitivity and 1 - specificity values are shown in Table 4.6. An ideal point on the graph of perfect classification would be the upper left corner with a coordinate of (0,1) with no false negatives and false positives. A diagonal with an angle of 45 degrees classifies the points on the graph into good and bad classification based on whether the points are above or below the diagonal respectively. The larger the distance of a point (above the diagonal) from the diagonal, the better the classification (less randomness). The distance of the points on the graph in Figure 4.3 from the point (0,1) and from the diagonal are calculated as Point-Point and Point-Line using equation 4.3 and 4.4, respectively and are shown in Table 4.7. In Figure 4.3, the point corresponding to BLAST e-value cut-off of 10 has the largest distance from the diagonal and the shortest distance from the point (0,1).

$$Point - Point = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4.3}$$

$$Point - Line = |am + bn + c|/\sqrt{a^2 + b^2} \tag{4.4}$$

Table 4.6: Sensitivity and 1 - specificity values based on TF_Seq results with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	1 - Specificity	Sensitivity
0.01	0.0077	0.1712
0.1	0.0110	0.2081
1	0.0186	0.2716
2	0.0240	0.2955
3	0.0279	0.3111
4	0.0310	0.3231
5	0.0342	0.3299
6	0.0381	0.3382
7	0.0407	0.3408
8	0.0438	0.3496
9	0.0467	0.3543
10	0.0498	0.3647

Table 4.7: Distance of points on the ROC curve from the point (0,1) and from the diagonal based on TF_Seq results with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	Point-Point	Point-Line
0.01	0.8289	0.1156
0.1	0.7920	0.1394
1	0.7286	0.1789
2	0.7049	0.1920
3	0.6894	0.2003
4	0.6776	0.2065
5	0.6710	0.2091
6	0.6629	0.2122
7	0.6605	0.2122
8	0.6518	0.2163
9	0.6474	0.2175
10	0.6372	0.2227

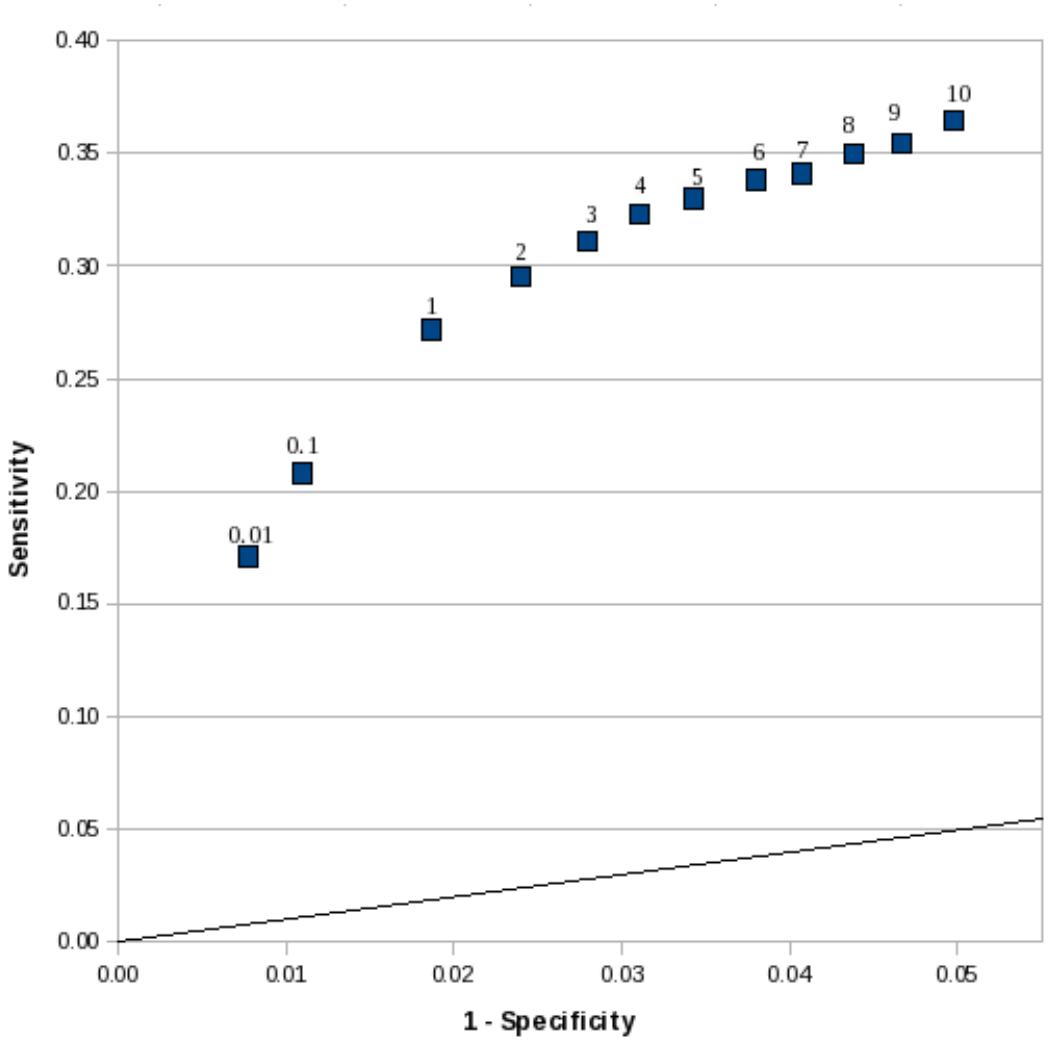


Figure 4.3: Sensitivity versus (1 - specificity) for different BLAST e-value cut-offs ranging from 0.01 to 10 using TF_Seq

4.2.2.2 TF_Fam and TF_SubFam methods

The purpose of this experiment is to estimate a suitable e-value cut-off for the BLAST tool used in the TF_Fam and TF_SubFam methods to achieve lower false negatives and higher true positives. The experiment is conducted to determine if the InterProScan tool can be used to refine the BLAST results with a higher e-value cut-off value. It is done by running BLAST with e-value cut-off ranging from 0.01 to 10 followed by the InterProScan tool as in TF_Fam and TF_SubFam with Yeast and *Arabidopsis thaliana* as source and target genomes repectively. The results are shown in Table 4.8 and Table 4.9. The number of mapped links increase gradually from e-value cut-off of 1 to 10 for TF_Fam and TF_SubFam, though there is a large increase of 20.2 percent and 17.1 percent from e-value cut-off of 0.01 to 0.1, and a large increase of 22 percent and 19.8 percent from e-value cut-off of 0.1 to 1 for TF_Fam and TF_SubFam respectively. The number of distinct transcription factors identified also increase gradually from e-value cut-off of 1 to 10 for TF_Fam and TF_SubFam, though there is a large increase of 15.7 percent and 5.3 percent from e-value cut-off of 0.01 to 0.1, and a large increase of 21.0 percent and 5.1 percent from e-value cut-off of 0.1 to 1 for TF_Fam and TF_SubFam respectively.

The results obtained in this experiment are analyzed by comparing the available transcription factor data, consisting of 1922 transcription factors, of *Arabidopsis thaliana* to the predicted transcription factors. There are 35351 sequences in the nucleotide database of *Arabidopsis thaliana* used as BLAST database in this experiment. For comparing results, binary classification is used and true positives, false positives, true negatives and false negatives are calculated for TF_Fam and TF_SubFam are shown in Table 4.10 and Table 4.13, respectively. For TF_SubFam, as shown in Figure 4.4, there is a large increase of 15.1 percent and 21.0 percent in true positives from e-value cut-off of 0.01 to 0.1 and 0.1 to 1 respectively, though there is a steady small increase from 397 to 454 for e-value cut-off ranging from 1 to 10. The false

Table 4.8: Number of mapped links of transcription factors(TFs) for *Arabidopsis thaliana* using TF_Fam and TF_SubFam with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	TF_Fam	TF_SubFam
0.01	751	181
0.1	903	212
1	1102	254
2	1162	271
3	1194	283
4	1224	299
5	1242	306
6	1272	313
7	1281	318
8	1297	322
9	1301	324
10	1322	330

Table 4.9: Number of sequences identified as transcription factors (TFs) for *Arabidopsis thaliana* using TF_Fam and TF_SubFam with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	TF_Fam	TF_SubFam
0.01	375	112
0.1	434	118
1	525	124
2	550	129
3	563	130
4	573	133
5	579	135
6	591	135
7	592	135
8	596	135
9	596	135
10	603	135

positives increase by 23.4 percent and 22.6 percent from e-value cut-off of 0.01 to 0.1 and 0.1 to 1, though there is a gradual increase from 705 to 868 for e-value cut-off ranging from 1 to 10.

Table 4.10: True positive, false positive, false negative and true negative values for transcription factors identified in TF_Fam for different e-value cut-offs

BLAST e-value cut-off	True positives	False positives	True negatives	False negatives
0.01	285	466	32963	1637
0.1	328	575	32854	1594
1	397	705	32724	1525
2	417	745	32684	1505
3	425	769	32660	1497
4	434	790	32639	1488
5	438	804	32625	1484
6	443	829	32600	1479
7	444	837	32592	1478
8	448	849	32580	1474
9	448	853	32576	1474
10	454	868	32561	1468

For further analysis of TF_Fam, sensitivity and specificity are calculated for the binary classification and the ratios of Sensitivity to (1 - Specificity) for TF_Fam with corresponding BLAST e-value cut-offs marked on the points on the graph are shown in Figure 4.5. In Figure 4.5, the point corresponding to BLAST e-value cut-off of 10 has the largest distance from the diagonal and the shortest distance from the point (0,1).

For TF_SubFam, as shown in Figure 4.6, there is a constant increase from 85 to 102 in true positives from e-value cut-off of 0.01 to 2 respectively; there is a very small increase from 102 to 107 for e-value cut-off ranging from 2 to 5; and then the true positives remain constant for e-value cut-off ranging from 5 to 10. No increase in true positives with higher e-value cut-offs signifies that it is less likely to find

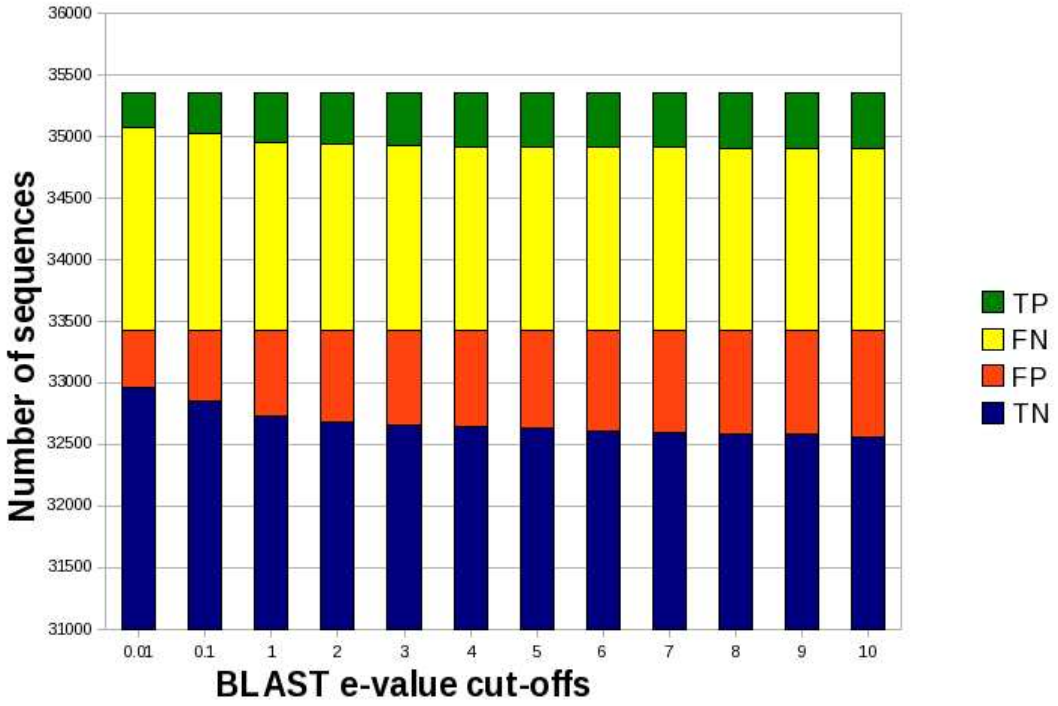


Figure 4.4: True positive, false positive, false negative and true negative values for transcription factors identified in TF_Fam for different e-value cut-offs

Table 4.11: Sensitivity and 1 - specificity values based on TF_Fam results with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	1 - Specificity	Sensitivity
0.01	0.0139	0.1483
0.1	0.0172	0.1707
1	0.0211	0.2066
2	0.0223	0.2170
3	0.0230	0.2211
4	0.0236	0.2258
5	0.0241	0.2279
6	0.0248	0.2305
7	0.0250	0.2310
8	0.0254	0.2331
9	0.0255	0.2331
10	0.0260	0.2362

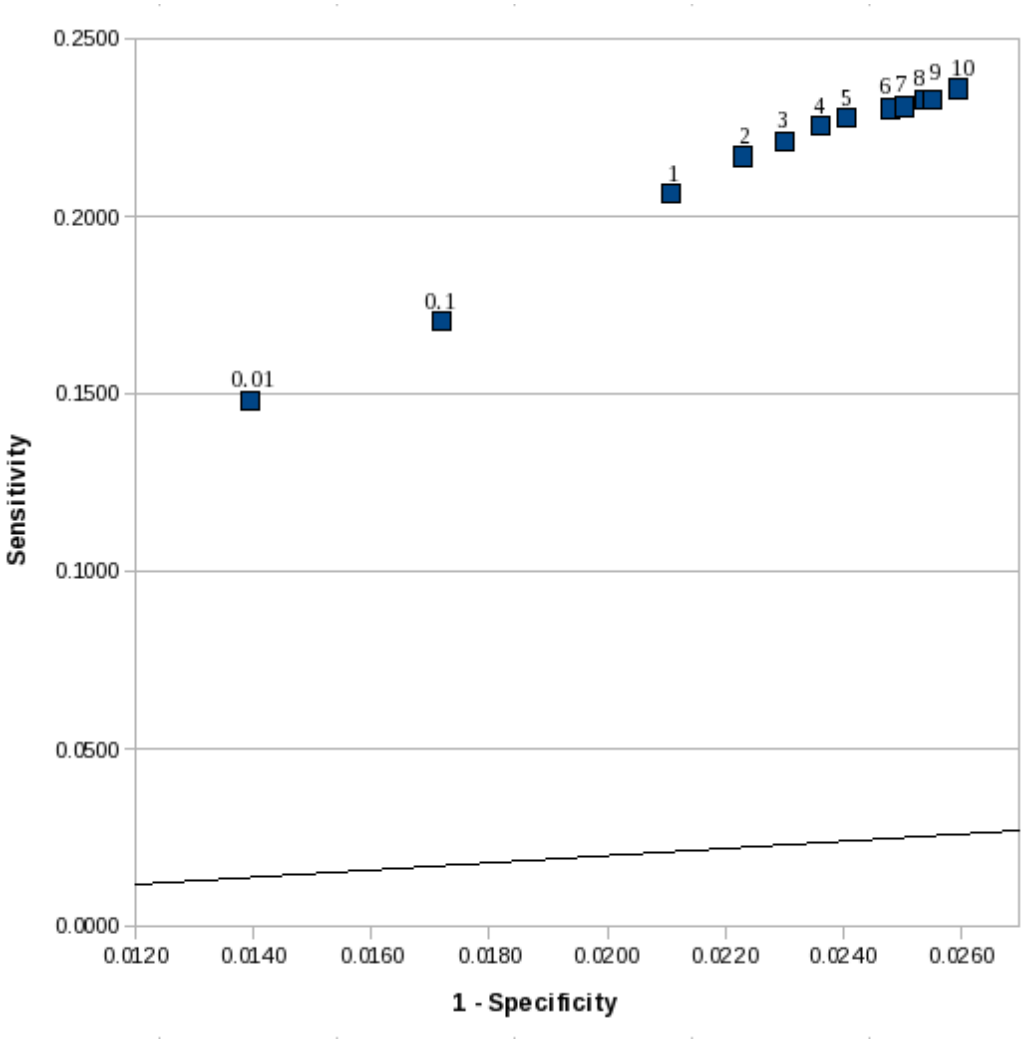


Figure 4.5: Sensitivity versus (1 - specificity) for different BLAST e-value cut-offs ranging from 0.01 to 10 using TF_Fam

transcription factors of the same sub-family in not much similar sequences. There is a large increase in false positives from 96 to 157 for e-value cut-off of 0.01 to 1; there is a gradual increase from 157 to 193 for e-value cut-off ranging from 1 to 4; and then very little increase from 193 to 223 for e-value cut-off ranging from 4 to 10. For further analysis of TF_SubFam, sensitivity and specificity are calculated for the binary classification in Table 4.14 and the ratios of Sensitivity to (1 - Specificity) for TF_SubFam with corresponding BLAST e-value cut-offs marked on the points on the graph is shown in Figure 4.7. In Table 4.15, points corresponding to BLAST e-value cut-off of 8, 9 and 10 have the largest distance from the diagonal and the shortest distance from the point (0,1).

4.2.3 Determination of correct type of transcription factors

In all the above experiments, the results from the three methods have been compared on the basis of how likely a sequence predicted as a transcription factor is actually

Table 4.12: Distance of points on the ROC curve from the point (0,1) and from the diagonal based on TF_Fam results with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	Point-Point	Point-Line
0.01	0.8518	0.0950
0.1	0.8295	0.1085
1	0.7937	0.1311
2	0.7834	0.1377
3	0.7792	0.1401
4	0.7746	0.1430
5	0.7725	0.1441
6	0.7699	0.1454
7	0.7694	0.1456
8	0.7673	0.1469
9	0.7673	0.1468
10	0.7642	0.1487

Table 4.13: True positive, false positive, false negative and true negative values for transcription factors identified using TF_SubFam for different e-value cut-offs

BLAST e-value cut-off	True positives	False positives	True negatives	False negatives
0.01	85	96	33333	1837
0.1	91	121	33308	1831
1	97	157	33272	1825
2	102	169	33260	1820
3	103	180	33249	1819
4	106	193	33236	1816
5	107	199	33230	1815
6	107	206	33223	1815
7	107	211	33218	1815
8	107	215	33214	1815
9	107	217	33212	1815
10	107	223	33206	1815

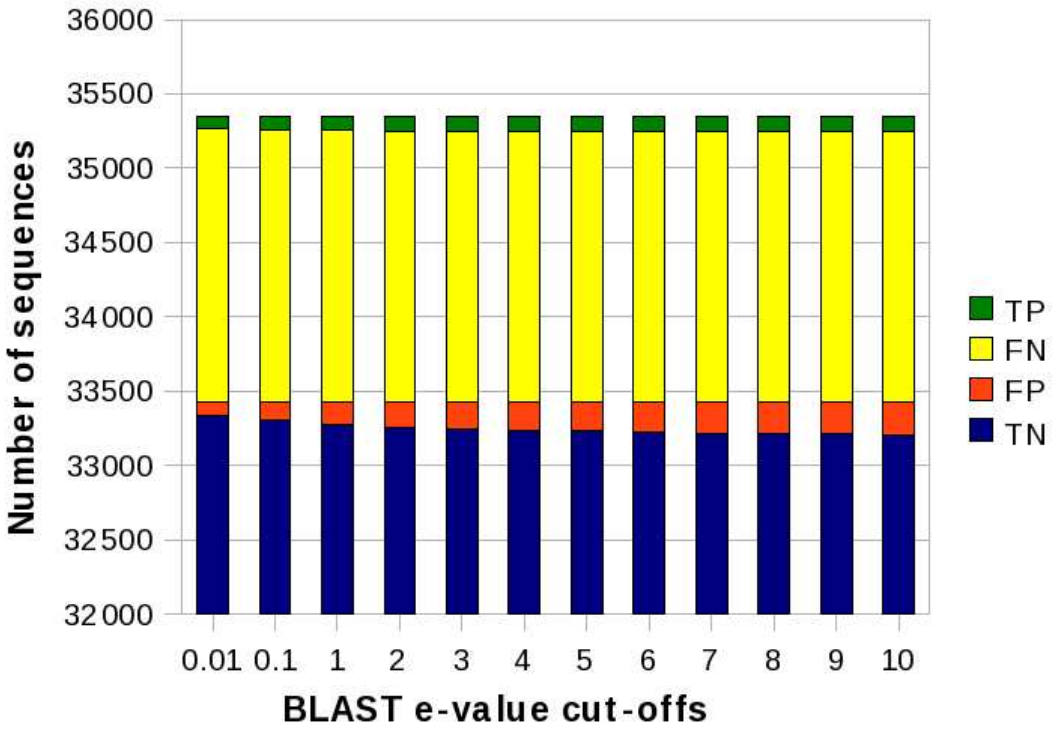


Figure 4.6: True positive, false positive, false negative and true negative values for transcription factors identified in TF_SubFam for different e-value cut-offs

Table 4.14: Sensitivity and 1 - specificity values based on TF_SubFam results with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	1 - Specificity	Sensitivity
0.01	0.0029	0.0442
0.1	0.0036	0.0473
1	0.0047	0.0505
2	0.0051	0.0531
3	0.0054	0.0536
4	0.0058	0.0552
5	0.0060	0.0557
6	0.0062	0.0557
7	0.0063	0.0557
8	0.0064	0.0557
9	0.0065	0.0557
10	0.0067	0.0557

Table 4.15: Distance of points on the ROC curve from the point (0,1) and from the diagonal based on TF_SubFam results with BLAST e-value cut-off ranging from 0.01 to 10

BLAST e-value cut-off	Point-Point	Point-Line
0.01	0.9558	0.0292
0.1	0.9527	0.0309
1	0.9495	0.0324
2	0.9469	0.0340
3	0.9464	0.0341
4	0.9449	0.0349
5	0.9443	0.0352
6	0.9443	0.0350
7	0.9443	0.0349
8	0.9444	0.0348
9	0.9444	0.0348
10	0.9444	0.0346

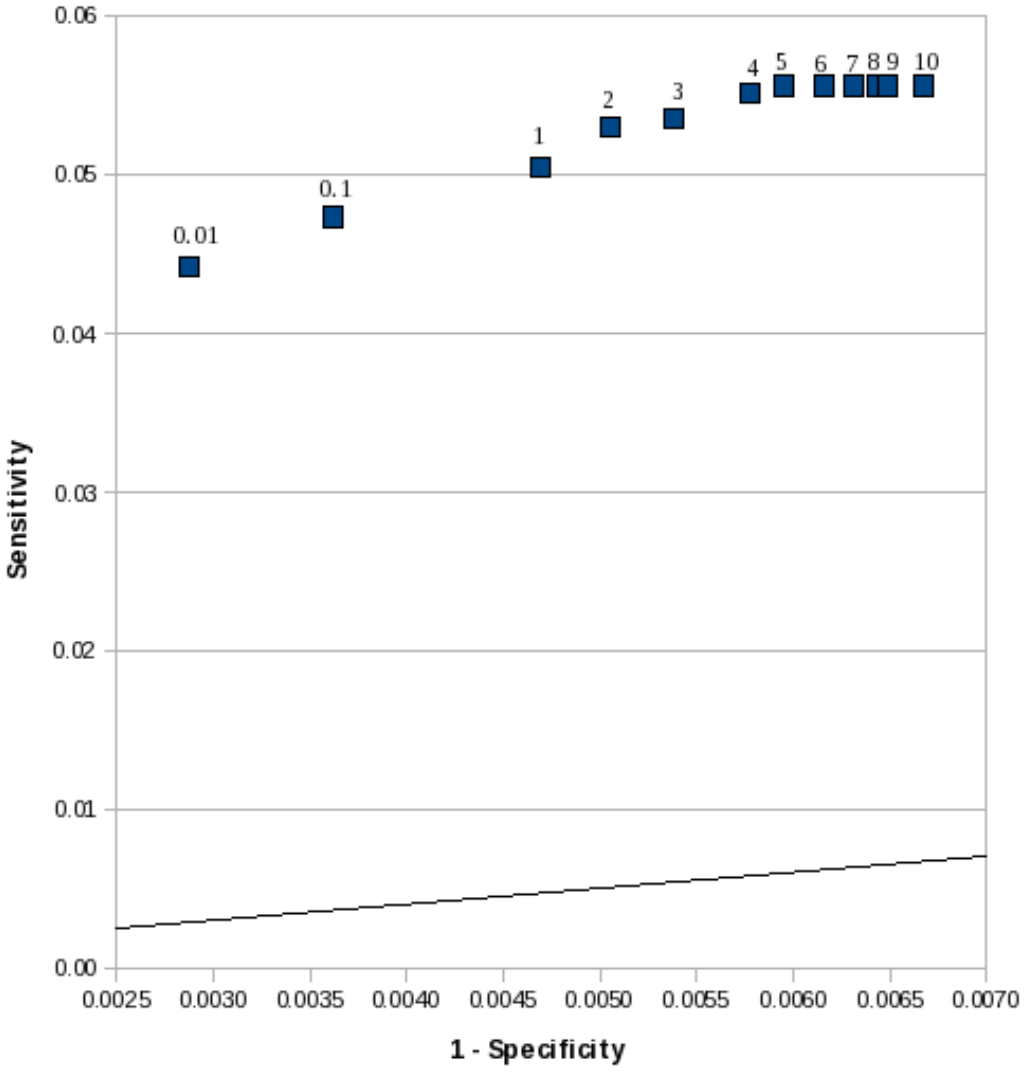


Figure 4.7: Sensitivity versus (1 - specificity) for different BLAST e-value cut-offs ranging from 0.01 to 10 in TF_SubFam

a transcription factor of the target genome. The goal of this experiment is to determine for TF_Seq if the sequences mapped as transcription factors for the target genome are actually of the same type as the transcription factor sequence of the source genome. The experiment is performed by first running BLAST tool on the nucleotide sequence data of *Arabidopsis thaliana* genome as BLAST database, with each transcription factor of Yeast genome representing a query sequence, varying the e-value cut-off value from 0.01 to 10 as shown in Figure 4.8. As a result, a set of BLAST outputs is collected corresponding to each transcription factor of Yeast genome. Next, annotation of each hit in every BLAST output for e-value cut-off of 10 is compared manually to the annotation of query sequence and is categorized into four categories, Confirmed TF, Similar TF and Other TF (transcription factor) and Not TF (not a transcription factor) as shown below.

Confirmed TFs are the hit sequences with the same annotation as the query sequence, that is, the annotations of the query sequence and hit sequences show that they are of the same transcription factor type.

Similar TFs are the hit sequences with annotation described as 'similar' to the query sequence, that is, the annotations of the hit sequences stated being 'similar to' the same transcrition factor type as that of the query sequence.

Other TFs are the hit sequences with annotation of being a transcription factor but not of the same type or similar to the transcription factor type of the query sequence.

Not TFs are the hit sequences that are not marked as Similar TF, Confirmed TF or Other TF.

Table 4.16 and Figure 4.9 shows results of this experiment with the number of hits (mapped links) grouped into four categories. The number of hits marked as Confirmed TF increases by 27.2 percent and Not TF increases by 54.6 percent from

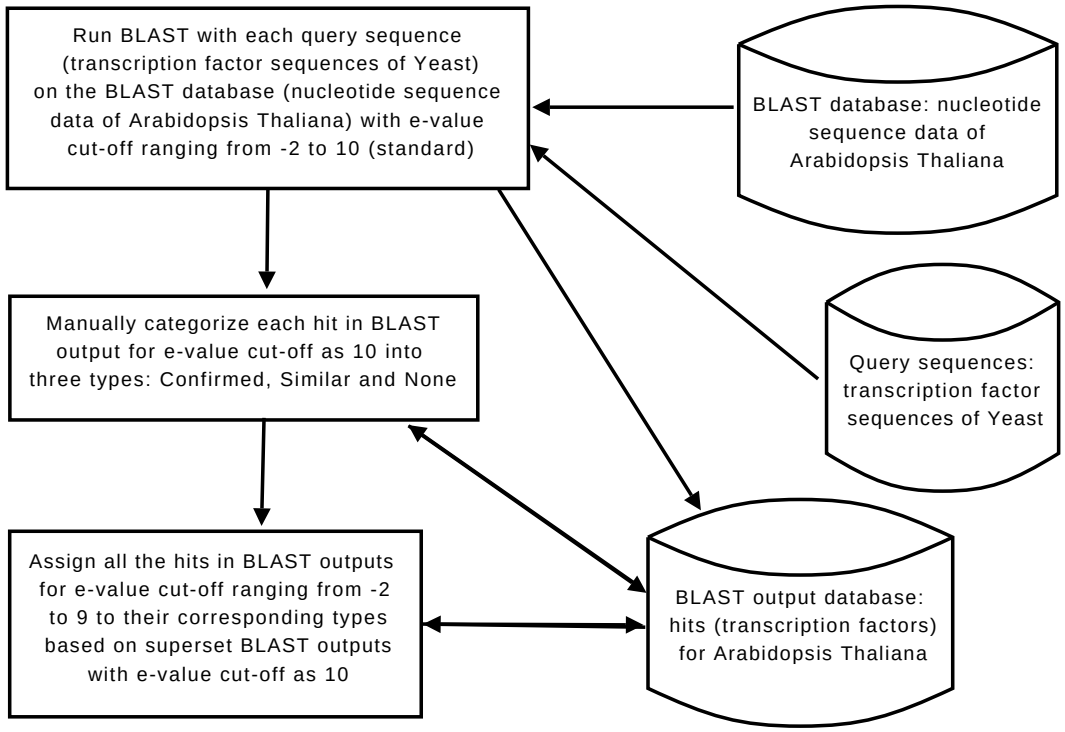


Figure 4.8: Flowchart for analyzing the identified sets of transcriptions factors from Yeast to *Arabidopsis thaliana* using BLAST

e-value cut-off of 0.01 to 0.1 though there is not much increase in number of hits marked as Similar TF and Other TF. From e-value cut-off of 0.1 to 1, there is a very large increase by 250 percent and 129 percent in number of hits categorized as Other TF and Not TF respectively but not much increase in number of hits marked as Confirmed and Similar TF. For all e-value cut-offs ranging from 1 to 10, there is a steep rise from 421 to 1556 in number of hits marked as Not TF, a slow rise from 1378 to 1912 in the number of hits marked as Confirmed TF and very little increase from 185 to 238 and 7 to 34 in the number of hits marked as Similar TF and Other TF respectively.

Table 4.16: Number of Hits divided into four types (Confirmed TF, Similar TF, Other TF and Not TF) for different e-value cut-offs

	Confirmed TF	Similar TF	Other TF	Not TF
0.01	791	119	1	119
0.1	1006	152	2	184
1	1378	185	7	421
2	1502	202	8	588
3	1582	213	12	745
4	1644	223	16	861
5	1699	226	18	991
6	1761	228	20	1121
7	1809	233	21	1217
8	1838	235	24	1341
9	1865	236	30	1437
10	1912	238	34	1556

The total number and percentage of distinct transcription factors identified and marked as Confirmed TF, Similar TF, Other TF and Not TF for e-value cut-off ranging from 0.01 to 10 are shown in Table 4.17, Figure 4.10 and Figure 4.11. The number of transcription factors marked as Confirmed TF and Not TF increases from 451 to 562 and 86 to 144 respectively for e-value cut-off of 0.01 to 0.1 though there is a very small increase from 62 to 77 and 1 to 2 in number of transcription factors

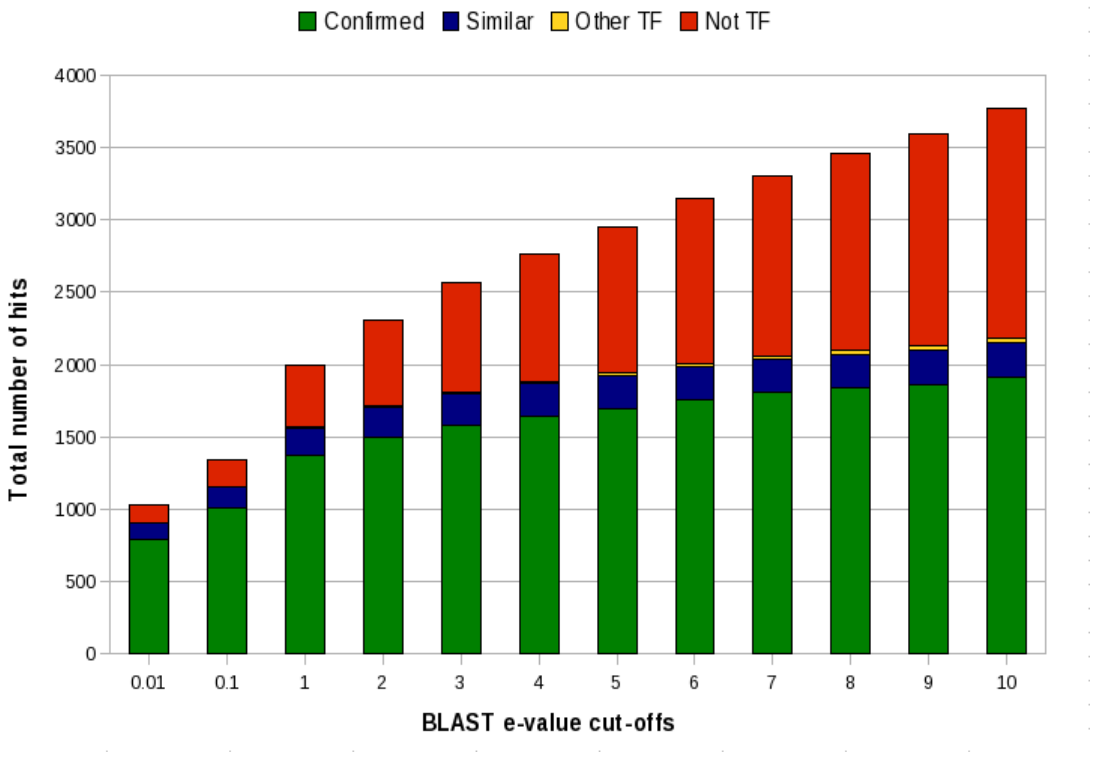


Figure 4.9: Number of Hits divided into four types (Confirmed TF, Similar TF, Other TF and Not TF) for different e-value cut-offs

marked as Similar TF and Other TF respectively. For e-value cut-off of 0.1 to 1, there is also a large increase in number of transcription factors categorized as Not TF and Confirmed TF from 562 to 749 and 144 to 341 respectively, but not much increase in number of transcription factors marked as Similar TF and Other TF from 77 to 83 and 2 to 7 respectively. Although for all e-value cut-offs ranging from 1 to 10, there is a steep rise in number of transcription factors marked as Not TF from 341 to 1345, a slow rise in the number of transcription factors marked as Confirmed TF from 749 to 969 and very little rise in the number of transcription factors marked as Similar and Other TF from 83 to 99 and 1 to 32 respectively. There is a significantly large number of sequences marked as Similar and probably containing the same protein domains as their query sequences but are not annotated as transcription factors containing the same protein domains. The sequences marked as Similar may be part of the false positives in the previous experiments as only the sequences that are annotated as transcription factors are included in true positives.

Table 4.17: Number of sequences identified as transcription factors divided into four types (Confirmed, Similar, Other TF and Not TF) for different e-value cut-offs

	Confirmed	Similar	Other TF	Not TF
0.01	451	62	1	86
0.1	562	77	2	144
1	749	83	7	341
2	809	92	8	498
3	843	93	11	628
4	868	94	15	729
5	885	95	16	838
6	912	97	18	954
7	928	97	19	1036
8	936	97	22	1150
9	952	97	28	1236
10	969	99	32	1345

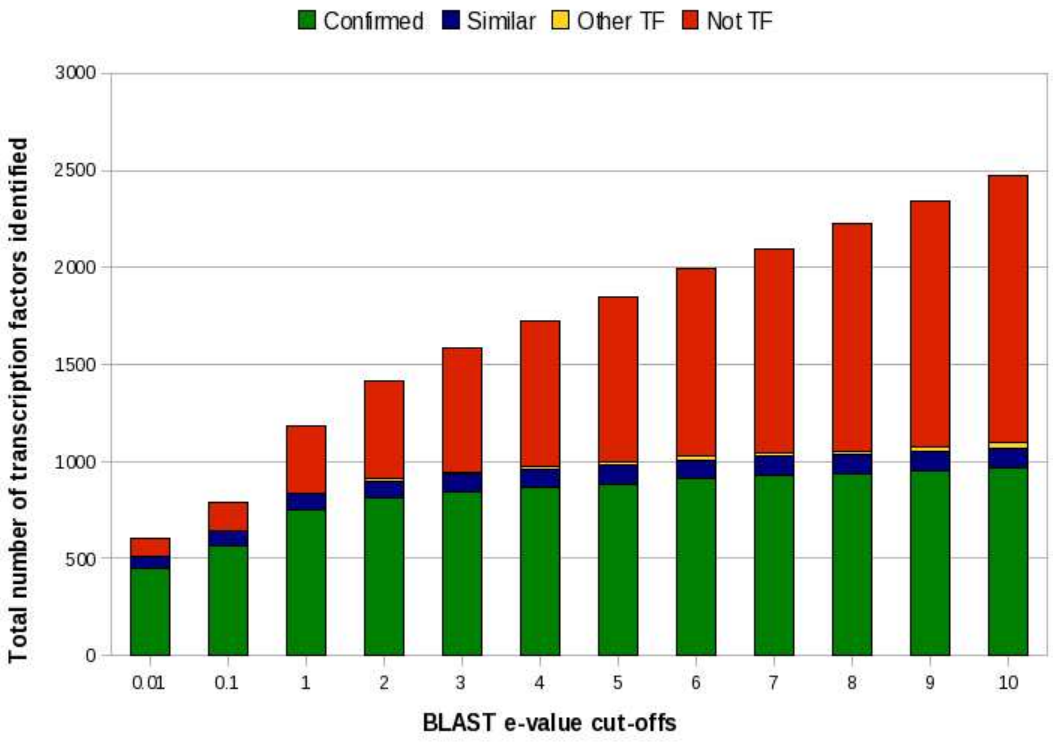


Figure 4.10: Number of transcription factors divided into four types (Confirmed TF, Similar TF, Other TF and Not TF) for different e-value cut-offs

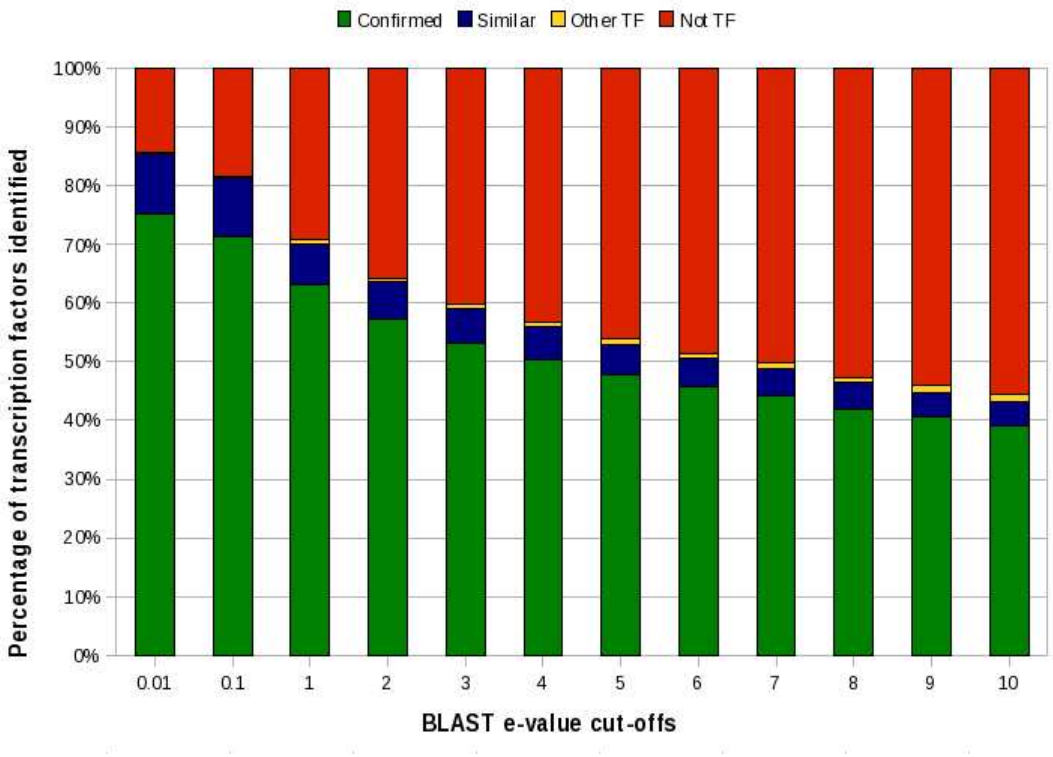


Figure 4.11: Percentage of transcription factors divided into four types (Confirmed TF, Similar TF, Other TF and Not TF) for different e-value cut-offs

4.3 Concluding Remarks

In this chapter, three methods have been used for identification of transcription factors based on sequence similarity, same protein family and subfamily classification. Mapping transcription factors with sequence similarity and protein family classification gave better results than using only sequence similarity or using sequence similarity along with protein subfamily classification. Next, experiments were conducted to identify a suitable BLAST parameter threshold for the three methods using ROC curve with e-value threshold ranging from 0.01 to 10. For the methods TF_Seq and TF_Fam, the most suitable BLAST e-value cut-off identified is 10 and for the method TF_SubFam it is 8, 9 and 10. Additionally, another experiment was performed to determine if a predicted target genome transcription factor, mapped using sequence similarity, was of the same type as the source genome transcription factor used for mapping. A high percentage of predicted transcription factors were mapped correctly at very low e-value thresholds, although this percentage decreased with the increase in e-value threshold.

Chapter 5

Predicting Target Genes in an Organism Related to a Model

5.1 Overview

Chapter 3 and 4 described the different methods employed to predict transcription factors, the first step of regulatory network mapping. The second step is to predict target genes in the target genome using information from the source genome, including its target genes and its transcription factors. This chapter [60, 63] [1] [2] presents the different methods proposed for predicting target genes and a comparative study of the prediction accuracies of these different mapping methods. The different methods used to predict target genes are based on searching TFBS motifs, sequence similarity, active binding sites and identification of TFBS motifs. BLAST [5], the Nucleosome prediction tool [58] and the GALF-P tool [14], discussed in Chapter 2, are used in the various target gene mapping methods described in this chapter. The datasets of *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are used as source and target

[1] Sharma, R.; Evans, P.A.; Bhavsar, V. C., Regulatory link mapping between organisms, BMC Systems Biology 2011, 5(Suppl 1): S4.

[2] Sharma, R.; Evans, P.A.; Bhavsar, V. C., Mapping a Regulatory Network Between Organisms, Proceedings of the Fourth International Conference on Computational Systems Biology 2010, 204-218.

genome test data, respectively, for target gene prediction.

5.2 Methodology

Predicting target genes in the target genome using source genome information, for a particular transcription factor in a regulatory network, involves finding similar target genes in the target genome. These similar target genes should have the same function as the source genome target genes. Similar function is important in order to have the same type of regulatory relationship in the target genome as the regulatory relationship being mapped from in the source genome. Since target genes may or may not produce proteins, they cannot be grouped based on similar protein signatures, but the BLAST tool can be used to find highly similar nucleotide sequences that will tend to have similar function.

Additionally, being part of the gene regulation process, the transcription factor of a regulatory link binds to the binding site containing the TF binding site (TFBS) motif, which is generally located upstream of the target gene sequence [39]. The transcription factors have specificity for motifs (patterns) in the binding site regions of the target genes based on the type and family of the transcription factor. Different transcription factors have different sets of TFBS motif specificities in order to regulate their target genes. Hence, the target gene being regulated by a certain transcription factor in the source genome will tend to have one of the specific TFBS motifs of that transcription factor in its binding site region. The binding sites of the mapped TGs in the target genome will also tend to contain one of the TFBS motifs of the source genome TF from the regulatory link being mapped. If the TFBS motif information is not available for the source genome, then the TFBS motifs can be located using one of the available methods used for finding TFBS motifs. The current genetic algorithms for identifying TFBS motifs use position-led and consensus-led

representations independently whereas GALF-P (Genetic Algorithm with Local Filtering and adaptive post-processing techniques) [14] combines both representations and uses local filtering to decrease false positives and improve efficiency. The binding sites also need be active, that is, located on the linker DNA to be available for interactions with proteins. Nucleosome Position Prediction Tool [58] is used to determine nucleosome occupancy in a sequence.

The six different methods propsed in this chapter that predict target genes in a target genome based on source genome information are: TG_Seq, TG_Prom, TG_Nuc, TG_BL, TG_BLProm, and TG_GALF. The result sets are evaluated in the next section by comparing the predicted target genes with the available target genes data for *Arabidopsis thaliana* to determine the number of mapped target genes predicted correctly as target genes of the target genome. A binary classifier is used to analyze the result sets from different methods.

5.2.1 Method 1: TG_Seq

Researchers develop motifs to describe the patterns of the features, and classify factors into types and families based on these patterns. There are several commonly used tools developed for locating binding sites for a genome such as MATCHTM [35], TFSEARCH [1], TFBIND [68], and CREAD [18]. These tools use the position weight matrix library from the TRANSFAC [28] database for searching binding sites; however, the full database is not freely available and thus could not be used for testing in this work. As there is not much data available for a non-model organism, the transcription factor binding site motifs of the source genome can be used to search binding sites in the nucleotide database of target genome. Therefore, in this thesis work, Yeast is used as the source genome with publicly available TFBS motif data. Different transcription factors have different sets of TFBS motifs that describe features of the binding sites that appear to affect gene regulation. Hence, the target

gene being regulated by a certain transcription factor in the source genome will tend to have one of the specific TFBS motifs of that transcription factor in its binding site region. The binding sites of the mapped target genes in the target genome will also tend to contain one of the TFBS motifs of the source genome transcription factor from the regulatory link being mapped. Figure 5.1 shows the second step of regulatory link mapping where the target genes set TGbs is predicted for the target genome. The binding site motif locator searches for the TFBS motifs of the source genome transcription factor in the target genome nucleotide sequence data. When a motif in the nucleotide sequence is an exact match to one of the TFBS motifs of the source genome transcription factor, then that gene is determined to be the predicted target gene for the target genome. The set of these predicted target genes are termed TGbs.

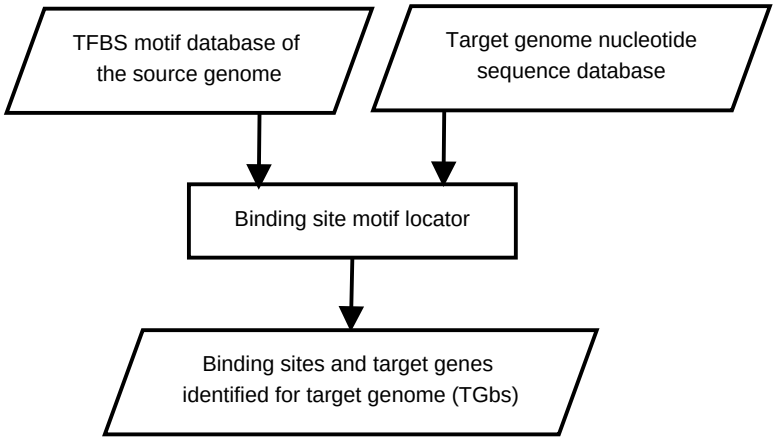


Figure 5.1: Method TG_Seq: based on searching TFBS motifs in target genome sequence data

5.2.2 Method 2: TG_Prom

Being part of the gene regulation process, the transcription factor of the regulatory link binds to the transcription factor binding site (TFBS) motifs of the target gene, generally upstream of the gene sequence in the promoter region. The promoter

sequence is a specific region of a DNA sequence that is generally upstream of the gene and tends to contain the TFBS motif for regulating that gene. These transcription factor binding site motifs can also be used to search the promoter sequence data of the target genome, if available, as the primary cis-regulatory elements tend to reside in the promoter regions. There are numerous tools developed for identifying the promoter regions in a sequence such as Promoter 2.0 [36], Dragon Promoter Finder [8], WWW Promoter Scan [51], Neural Network Promoter Prediction [74]; however, of the tools that are freely available for research purposes, some tools are only usable through a web interface for individual sequences and are not available for high-throughput use, while the others are designed only for animal genomes. Therefore, available promoter sequence data for the target genome is used in this technique for identifying target genes.

Hence, another set of predicted target genes named TGpr is determined by searching the TFBS motifs in the promoter sequence data instead of the nucleotide sequence data (compared to previous method) of the target genome, as shown in Figure 5.2.

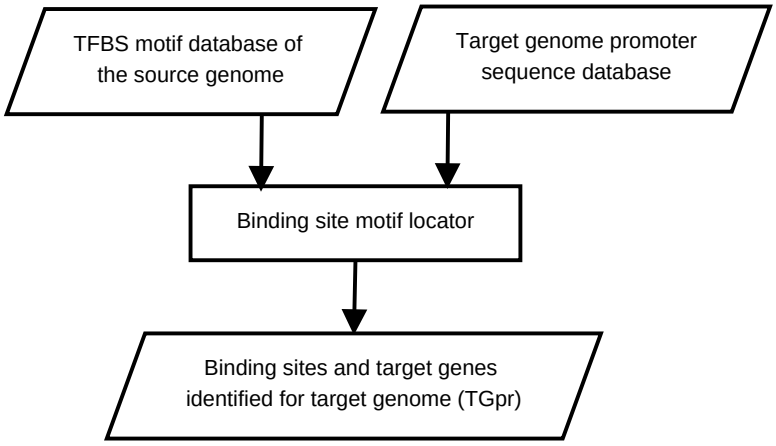


Figure 5.2: Method TG_Prom: based on searching TFBS motifs in target genome promoter sequence data

5.2.3 Method 3: TG_Nuc

The result set TGbspr from the previous target gene prediction method can be further filtered by identifying the likely active binding sites in the predicted target genes of the target genome. Using nucleosome position information might be useful in predicting active binding sites. Most of the functional binding sites, transcription start sites, coding and intergenic regions are located in the low nucleosome occupancy region. Therefore, nucleosomes tends to affect regulation of a gene and it is reasonable to ignore binding site motifs identified in a region of high nucleosome occupancy.

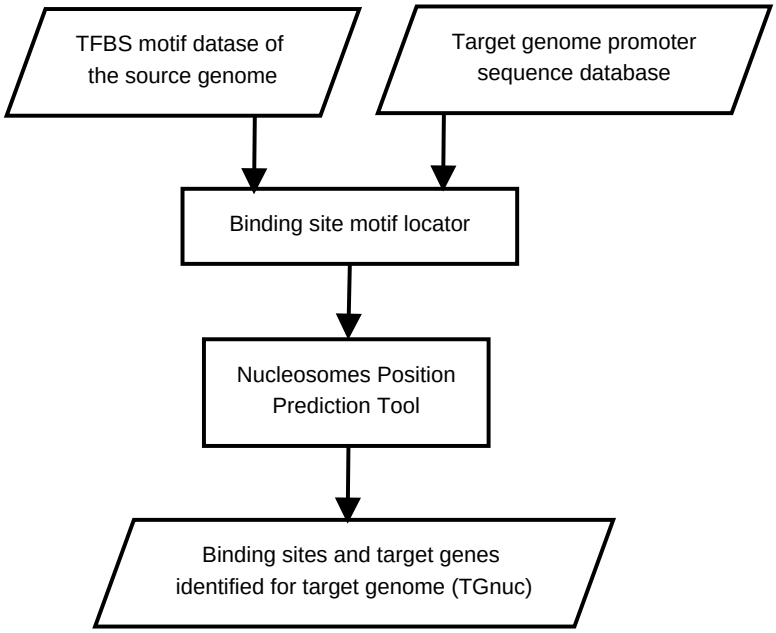


Figure 5.3: Method TG_Nuc: based on searching TFBS motifs and identifying active binding sites

The recently developed Nucleosomes Position Prediction Tool predicts the probability of nucleosome occupancy and nucleosome start probability at every position in a nucleotide sequence [58]. This tool, based on the yeast nucleosome-DNA interaction model and other default parameters, is used to predict nucleosome occupancy on the promoter sequences of the predicted target genes of the target genome as shown in Figure 5.3. The predicted target genes, in TGbspr result set, whose binding sites

have more than 50 percent probability of being located in the nucleosome occupancy region are discarded in this method. The result set of predicted target genes, TGnuc, consists of genes that contain active binding site regions.

5.2.4 Method 4: TG_BL

Finding target genes in a target genome using target gene information of the source genome can involve finding similar target genes in the target genome. Functionally similar target genes are likely to be regulated similarly, but the functional similarity is not needed by itself. Therefore, similarly functioning genes may have the same type of regulatory relationship in the target genome as the regulatory relationship being mapped from in the source genome. Since target genes may or may not produce proteins, they cannot be grouped based on similar protein signatures, but the BLAST tool can be used to find highly similar nucleotide sequences that will tend to have similar function. Sequence similarity is a commonly used way to find sequences with similar function.

The target genes, in the source genome regulatory network data, are query sequences for running the BLAST tool, with the nucleotide sequence data of the target genome as the BLAST database, to identify similar nucleotide sequences in the target genome, as shown in Figure 5.4. These similar sequences in the target genome are predicted as target genes for the target genome and are named TGbl.

5.2.5 Method 5: TG_BLProm

As discussed earlier, the TFBS motifs are useful in locating the correct target genes for a transcription factor. Therefore, in this method, sequence similarity between source genome target genes and target genome nucleotide sequence data is combined with searching the TFBS motifs to predict the target genes in the target genome. Target genes predicted in the previous method TGbl are filtered by searching for

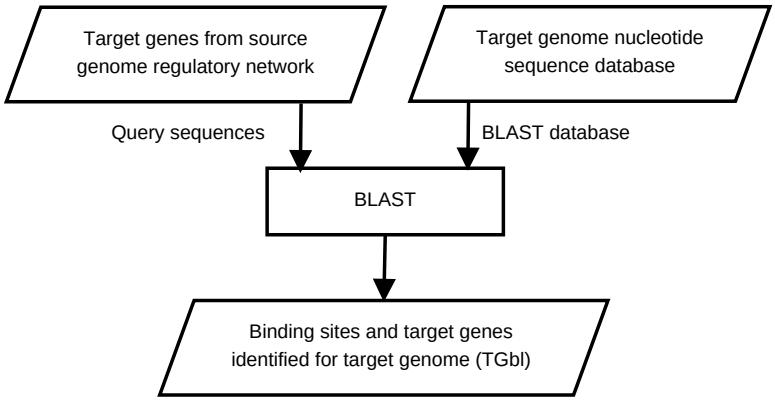


Figure 5.4: Method TG_BL: based on sequence similarity

specific TFBS motifs in these similar target gene sequences of the target genome, as shown in Figure 5.5.

For a particular target gene in a regulatory link, initially the BLAST tool is used to find the similar target genes in the target genome. Target genome nucleotide sequence data is used as the BLAST database, and the source genome target genes from the regulatory network being mapped are the query sequences. The output of BLAST is a set of similar target genes that are then the input for the next step of locating binding sites. The role of the binding site motif locator is to search for the TFBS motifs of the source genome transcription factor in promoter sequences of the similar target genes found in the target genome. Finally, the genes that are similar and have the right TFBS motifs are predicted as target genes for the target genome and named TGblbs.

5.2.6 Method 6: TG_GALF

There are computationally determined motifs that are based on experimentally determined TFBS motifs available for various transcription factors in a source (model) genome. These TFBS motifs have been used in the previous methods to identify target genes in a target genome containing these TFBS motifs. However, the infor-

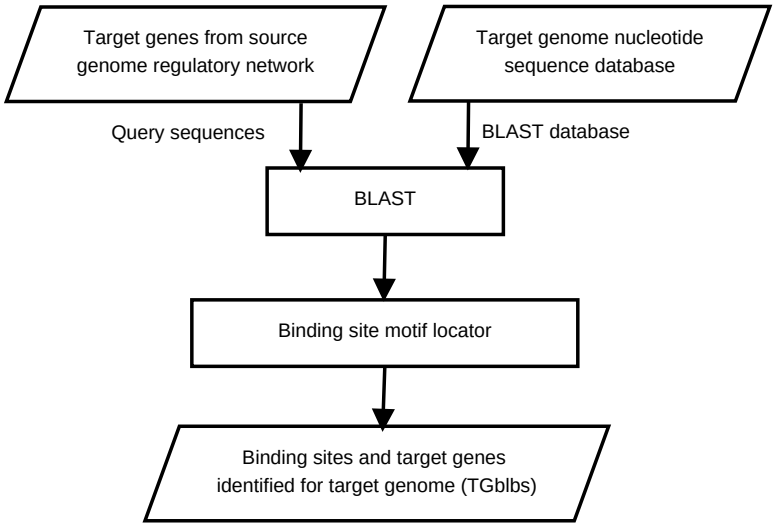


Figure 5.5: Method TG_BLProm: based on sequence similarity and searching TFBS motifs in target genome promoter sequence data

mation available about TFBS motifs for some model organisms may be incomplete or missing entirely. These TFBS motifs might also not be completely conserved in the target genome. Therefore, target genes can be missed, resulting in many false negatives in the results. Therefore, there is a need to first find the TFBS motifs in source genome and then use these motifs to predict target genes. The current genetic algorithms for identifying TFBS motifs use position-led and consensus-led representations independently whereas GALF-P (Genetic Algorithm with Local Filtering and adaptive post-processing techniques) [14] combines both representations and uses local filtering for decreasing false positives and getting better efficiency. The results based on synthetic and real datasets using current genetic algorithms such as GAME [75], BioOptimizer [32], MEME [7], and BioProspector [41] compared to using GALF-P demonstrated weaker performance [14].

Therefore, GALF-P, with motif length parameter as 7, maximal sequence length parameter as 20,000 and rest of parameters as default, is used to identify TFBS in the target genes regulated by a given transcription factor of the source genome as shown in Figure 5.6. As a result, a number of common TFBS are identified among these

target genes that may have the tendency to interact with the regulating transcription factor. Target genes are then predicted for the target genome by searching all the TFBS motifs found using GALF-P for a target gene in the promoter sequence database [19] of the target genome. The result set identified using this method is named as TGgalf.

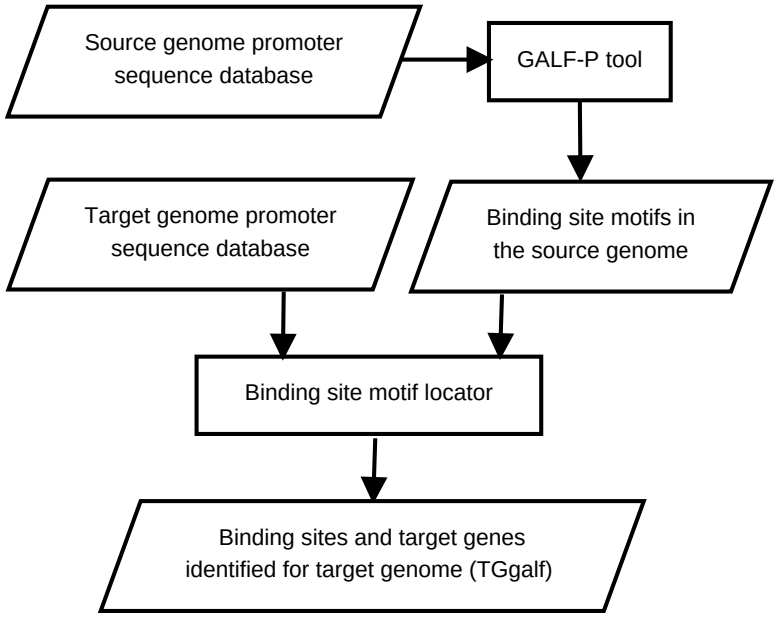


Figure 5.6: Method TG_GALF: based on identification of TFBS motifs

5.3 Results and Analyses

The results and analysis for experiments conducted using the six methods for predicting target genes of *Arabidopsis thaliana* using target gene information of *Saccharomyces cerevisiae* [26] are discussed in this section. 25191 target genes were identified in the *Arabidopsis thaliana* [50] genome that are used for validation of the predicted target genes. Six sets of predicted target genes are identified for *Arabidopsis thaliana* using the target gene mapping methods described in the previous section.

Five out of the six methods are based on searching TFBS motifs alone or combined with sequence similarity, active binding sites or identification of TFBS motifs. The binding site locator, used in these five methods to search TFBS motifs, predicts the binding sites in the target genome. The number of binding sites identified using the five methods, TG_Seq, TG_Prom, TG_NUC, TG_BLProm and TG_GALF are shown in Table 5.1. Method TG_BLProm identifies the lowest number of binding sites, whereas method TG_Seq predicts the highest number of binding sites in the target genome. Searching TFBS motifs in the nucleotide sequence data compared to promoter sequence data predicts more binding sites in the target genome. Additionally, similar numbers of binding sites are identified as would be expected by methods TG_Prom, TG_Nuc, and TG_GALF that are based on searching TFBS motifs in the promoter sequences. However, method TF_BLProm predicts fewer binding sites since the TFBS motifs are searched in the promoter sequences of only highly similar target genes of the target genome.

Table 5.1: Number of binding sites identified for *Arabidopsis thaliana* using methods TG_Seq, TG_Prom, TG_Nuc, TG_BLProm and TG_GALF

Method used	Number of binding sites identified
TG_Seq	1087795
TG_Prom	587249
TG_Nuc	424145
TG_BLProm	82678
TG_GALF	492975

The total number of distinct predicted target genes containing the identified binding sites for all six target gene prediction methods are shown in Table 5.2. The six sets of predicted target genes are: TGbs, TGpr, TGnuc, TGbl, TGblbs and TGgalf. Similar to the previous result of binding sites, the highest number of target genes is identified in the result set TGbs. This suggests that searching TFBS motifs in the complete

nucleotide sequences leads to more predicted target genes identified than searching for TFBS motifs in the promoter sequences as would be expected. Similar numbers of target genes are located in the result sets TGpr, TGnuc and TGgalf. Additionally, the result sets TGblbs and TGbl contain the same number of predicted target genes. This suggests that all the highly similar target genes in the target genome contain the TFBS motifs in their promoter sequences for the corresponding source genome transcription factor. To look at it another way, it appears the binding site motif locator did not add anything to the the search.

Table 5.2: Number of target genes identified for *Arabidopsis thaliana* using TG_Seq, TG_Prom, TG_Nuc, TG_BL, TG_B:Prom and TG_GALF

Method used and Result Set	Number of target genes identified
TG_Seq: Result set TGbs	35181
TG_Prom: Result set TGpr	25259
TG_Nuc: Result set TGnuc	19976
TG_BL: Result set TGbl	2252
TG_BLProm: Result set TGblbs	2252
TG_GALF: Result set TGgalf	25388

The six result sets are divided into two groups based on the sequence data used in the first step of each method. The sequence data used is either nucleotide sequence data or promoter sequence data. Therefore, Group 1 contains the result sets TGbs, TGbl and TGblbs that used nucleotide sequence data. The result sets TGpr, TGnuc and TGgalf, determined using promoter sequence data, belong to Group 2. The result sets are compared and analyzed using a binary classifier [45] in the following subsections. The results are classified into four groups: true positives (TP), false

positives (FP), true negatives (TN) and false negatives (FN). The TPs are the predicted target genes that are present in the available target genes data of the target genome. FPs are the predicted target genes that are not present in the available binding database. The target genes that are present in the available target genes data but have not been identified by the method are the FNs. The sequences that are not present in the available target genes data and are discarded by the method are TNs.

5.3.1 Comparison of Group 1 result sets

The result sets TGbs, TGbl and TGblbs are given in Table 5.3 and Figure 5.7. The true positives decrease from 25180 to 1717 by 93.18 percent and the false positives decreases from 10001 to 535 by 94.65 percent for set TGbs to set TGblbs. Therefore, set TGblbs is more accurate than set TGbs since it loses very few true positives and many false positives. However, a large number of true target genes are predicted in result set TGbs compared to set TGbl and set TGblbs. Comparing set TGbl and set TGblbs, the same set of target genes are identified in both result sets. These results show that the TFBS motifs were found in all the highly similar target genes identified using BLAST. Hence, no true target gene is lost when the predicted target genes (set TGbl) determined using BLAST are refined further by searching the TFBS motifs in the promoter sequences, resulting in set TGblbs.

Table 5.3: True positive, false positive, false negative and true negative values for target genes identified in result sets TGbs, TGbl and TGblbs

	TGbs	TGbl	TGblbs
Number of target genes identified	35181	2252	2252
True Positives	25180	1717	1717
False Positives	10001	535	535
True Negatives	159	9625	9625
False Negatives	11	23474	23474

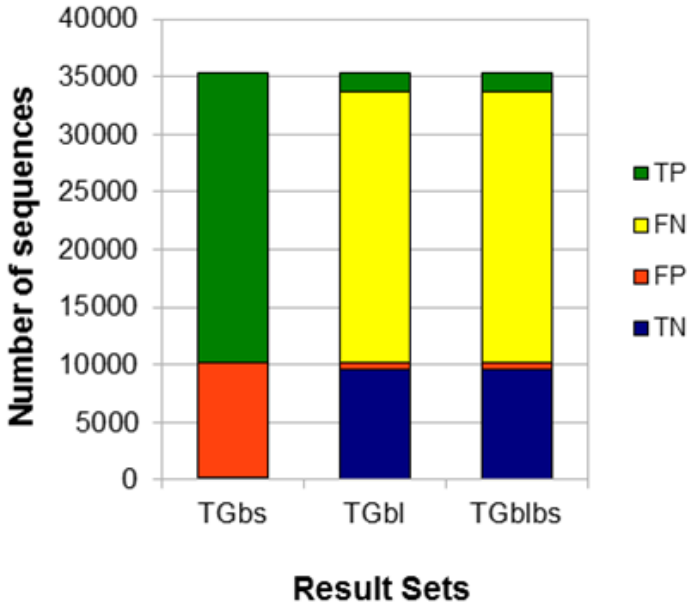


Figure 5.7: True positive, false positive, false negative and true negative values for target genes identified in result sets TGbs, TGbl and TGblbs

5.3.2 Comparison of Group 2 result sets

The result sets TGpr, TGnuc, and TGgalf with true positives, false positives, true negatives and false negatives are shown in Table 5.4 and Figure 5.8. Comparing result sets TGpr and TGgalf, both sets have similar numbers of true positives. Comparing result sets TGpr and TGnuc, true positives decrease from 24973 to 19702 by 21.1 percent and false positives decrease from 286 to 274 by only 4.2 percent. Therefore, result sets TGpr and TGgalf have more accurate results than TGnuc with much more true positives and only a little more false positives. Hence, refining the result set TGpr using the Nucleosomes Position Prediction tool, to get the result set TGnuc, does not improve the accuracy of the results. The result sets TGpr and TGgalf are very close and hence both methods, TG_Prom and TG_GALF, can be used to predict target genes.

Table 5.4: True positive, false positive, false negative and true negative values for target genes identified using TGpr, TGnuc, and TGgalf

	TGpr	TGnuc	TGgalf
Number of target genes identified	25259	19976	25388
True Positives	24973	19702	25077
False Positives	286	274	311
True Negatives	49	61	24
False Negatives	218	5489	114

5.3.3 Comparison of Group 1 and Group 2 result sets

The first three sets, TGbl, TGbs and TGblbs (shown in Table 5.5), use the nucleotide sequence data of *Arabidopsis thaliana* as the target genome database for searching the target genes. The first set TGbl is obtained as a result of using BLAST only and searching TFBS motifs is used to identify the second set TGbs. To refine these results, BLAST is used first to find highly similar TGs before searching for the

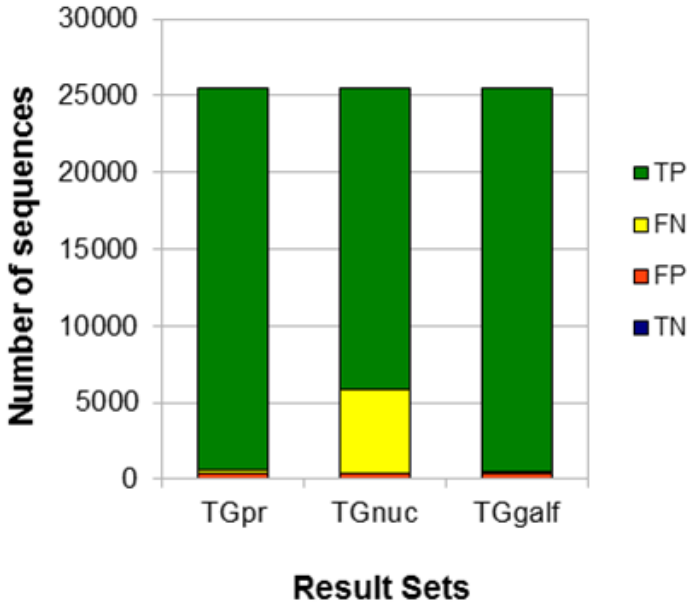


Figure 5.8: True positive, false positive, false negative and true negative values for target genes identified using TGpr, TGnuc, and TGgalf

TFBS motifs. Hence, the third set TGblbs is determined using both BLAST and searching TFBS motifs. Results for set TGbs show that most of the *Arabidopsis thaliana* sequences are predicted as target genes containing TFBS motifs in the target genome. True positives are the highest for this set, among the first three sets, along with lowest false negatives. Further, it does have the highest false positive count.

Table 5.5: True positive, false positive, false negative and true negative values for target genes identified in result sets TGbs, TGbl and TGblbs, TGpr and TGgalf

	TGbs	TGbl	TGblbs	TGpr	TGgalf
Number of target genes identified	35181	2252	2252	25259	25388
True Positives	25180	1717	1717	24973	25077
False Positives	10001	535	535	286	311
True Negatives	159	9625	9625	49	24
False Negatives	11	23474	23474	218	114

Comparing set TGbl and set TGblbs, the same set of target genes are identified in both result sets. These results show that the TFBS motifs were found in all the highly similar TGs identified using BLAST. Therefore, refining the result set from TGbl to TGblbs, based on searching TFBS motifs, did not lose any true target genes. The last two sets, TGpr and TGgalf (shown in Table 5.5), use the promoter sequence data of *Arabidopsis thaliana* as the target genome database for searching the target genes. The fourth set TGpr is determined by searching TFBS motifs in the target genome database. The last set TGgalf is obtained by using the GALF-P tool (Genetic Algorithm with Local Filtering and adaptive post-processing techniques) to identify TFBS motifs, followed by searching these TFBS motifs in the target genome database. A large number of target genes are identified in set TGpr and set TGgalf with lower false positives as compared to set TGbs. This decrease in false positives reflects the increased selectivity of the promoter sequence data used for sets TGpr and TGgalf instead of the nucleotide sequence data used for set TGbs. This im-

provement in accuracy over TGbs, TGbl and TGblbs makes TGpr and TGgalf the best sets for mapping target genes. There is little difference between the results for these two sets.

5.4 Conclusion

We have proposed six methods for target gene mapping, based on sequence similarity, TFBS motifs, active binding sites, and sequence similarity along with TFBS motifs. Using three methods, most of the target genes are identified correctly for *Arabidopsis thaliana* by searching TFBS motifs only. These methods have better results with many more true positives than using only sequence similarity and than using sequence similarity with TFBS motifs. The same set of target genes is predicted using sequence similarity and using sequence similarity along with TFBS motifs. Therefore, the methods using TFBS motifs only are the preferred method for general target gene mapping. Among these three methods, the two methods using the target genome promoter sequence database for searching TFBS motifs are better with much lower false positives.

Chapter 6

Mapped Regulatory Element Integration in a Related Organism

6.1 Overview

In Chapters 3, 4 and 5, the methodology and results for predicting transcription factors and target genes for a target genome (non-model organism) using information from a source genome (model organism) have been discussed. The final step of mapping a regulatory network from one genome to another is to integrate the mapped regulatory elements to obtain the regulatory links of the target genome. It is crucial to find out if the predicted target gene in the target genome is correctly linked to the right transcription factor by the link mapped from the source genome. This chapter describes the methods used to integrate these regulatory elements into regulatory links [60, 63] [1] [2], refine these links and validate this refined dataset of regulatory links using gene expression data. Two different sets of publicly available gene expression data for the target genome are used to refine and validate the integrated

[1] Sharma, R.; Evans, P.A.; Bhavsar, V. C., Regulatory link mapping between organisms, BMC Systems Biology 2011, 5(Suppl 1): S4.

[2] Sharma, R.; Evans, P.A.; Bhavsar, V. C., Mapping a Regulatory Network Between Organisms, Proceedings of the Fourth International Conference on Computational Systems Biology 2010, 204-218.

links due to regulatory network data (transcription factors, target genes and their relationships) not being available for testing for a non-model organism. A set of rules is established to use the expression information in the gene expression experiments to refine the regulatory relationships predicted for the target genome. The dataset of *Saccharomyces cerevisiae* is used as source genome test data and *Arabidopsis thaliana* dataset as target genome test data for regulatory links mapping.

6.2 Methodology

The regulatory links for the source genome are mapped to the target genome to predict the regulatory links for the target genome. The regulatory link mapping involves integration of regulatory elements identified for the target genome in the previous chapters and use of gene expression data, if available, for the target genome. The gene expression data of the target genome is used to filter the predicted regulatory links and analyze the predicted regulatory links for the target genome. The methods for integrating regulatory elements to form links, refining regulatory links and verifying the predicted regulatory links are described in the following sections.

6.2.1 Mapped Regulatory Element Integration

A regulatory relationship in a gene regulatory network consists of a transcription factor, target gene and the type of regulatory relationship between them, which can be either positive or negative. Two regulatory links from two different genomes tend to be similar if the two transcription factors from these links bind to the same motifs in the target genes, implying that these transcription factors might regulate the target genes in a similar way. Hence, the TFBS motif of the transcription factor present in the target gene of the source genome should also be present in the target gene of the target genome. In this final step of regulatory network mapping shown in Figure 6.1,

for every link in the source genome regulatory network, the corresponding predicted transcription factors are located among the mapped transcription factors identified in the first step of network mapping. The TFBS motifs of the source genome transcription factor are searched in its target gene. The TFBS motifs that are present in the source genome target gene are then searched in the corresponding mapped target gene nucleotide sequences of the target genome. Finally, these transcription factors and the target genes, which contain the specific TFBS motifs, are combined to obtain the target genome regulatory links.

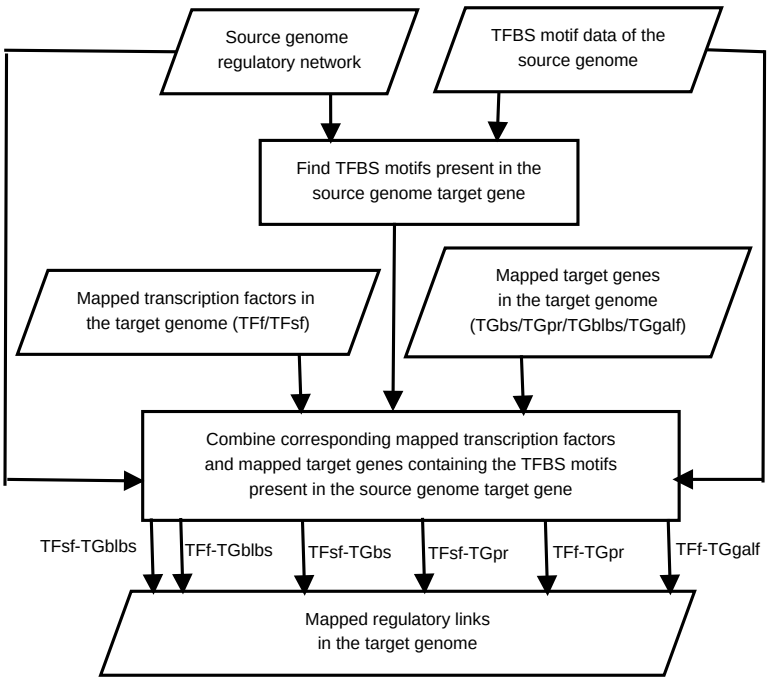


Figure 6.1: Method to integrate mapped transcription factors (TFf and TFsf) and target genes (TGblbs, TGbs, TGpr and TGgalf) for the target genome

Six sets of predicted regulatory links are identified using this method out of the possible eight sets. Out of these six sets of predicted regulatory links considered, the first set TFsf-TGblbs is obtained by integrating the predicted transcription factor set TFsf, combining BLAST similiarity and subfamily classification, and the predicted target gene set TGblbs, combining BLAST similarity and TF binding sites. The set TFf found using BLAST similarity and family classification is integrated with the

set TGblbs to identify the second predicted regulatory links set TFf-TGblbs. The third set TFsf-TGbs is identified by combining the mapped transcription factors from the set TFsf and the mapped target genes from the set TGbs, found using binding sites alone. The fourth and fifth sets, TFsf-TGpr and TFf-TGpr, are identified by combining the target genes in TGpr from searching promoter regions for binding site motifs with the transcription factors found with protein subfamily and family classification. The sixth and final set TFf-TGgalf is identified by combining the transcription factors using family classification with the target genes TGgalf found through locating motifs determined by GALF-P in the promoter sequences. The possible set TFf-TGbs, combining sets TFf and TGbs, is not included for further refinement and analysis because set TGbs with high false positives is already being combined with a low false positive transcription factor set TFsf to reduce false positives in predicted regulatory links, so TFf-TGbs should be worse than TFsf-TGbs. Additionally, another possible set TFsf-TGgalf, combining sets TFsf and TGgalf, is not included because it generates very few regulatory links resulting in high false negatives.

6.2.2 Regulatory Link Refinement

Only small, potentially related and arbitrary sets of biased gene expression data are usually available for non-model organisms. It can be used to determine how a target gene is going to react in the presence and absence of a transcription factor, even though the data would not be sufficient for link prediction using statistical techniques. Gene expression data is used to analyze the mapped regulatory links of the target genome and predict expression level relationships between transcription factors and target genes. It is used for both refining the regulatory links and testing the results.

Gene expression data contains expression levels of some of the genes investigated for

a genome in any gene expression experiment. Different experiments for the genome contain different expression levels of the genes depending on the experimental conditions. On the basis of the expression value, a gene is considered expressed or not in any given experiment. The expression levels of the transcription factor and the target gene from a predicted regulatory link in different experiments can be investigated to determine if the expression data supports the regulatory relationship in the predicted link.

In this thesis, gene expression data from The Arabidopsis Information Resource (TAIR) is used for the target genome, *Arabidopsis thaliana*. This expression data is already normalized with information determined about a gene being expressed or not in an experiment, which is required for refining the integrated regulatory links in this thesis. Therefore, only the expression information (expressed or not expressed) for the genes was used from this data and every experiment in this data was treated independently.

In the first step of regulatory links refinement, the predicted regulatory links and the gene expression data from some experiments on the target organism are used to determine if the expression values of the regulatory elements are present in the experiments. For any predicted regulatory link, the information of the regulatory elements, as expressed (Yes), not expressed (No) or absent (ab), in every gene expression experiment is collected. A set of rules is established to verify the regulatory links based on regulation type, transcription factor expression and target gene expression in the gene expression experiments as shown in Table 6.1. Using these rules, the gene expression experiments for every predicted regulatory link are classified into three groups: Confirming (C), Contradictory ($\overline{C}$) and Neutral (N). The number of elements in the Confirming group (c) for a predicted regulatory link represents the number of gene expression experiments that verify that the predicted regulatory link is correctly mapped in the target genome. The number of gene expression exper-

iments that contradict the regulatory relationship in the predicted regulatory link is the number of elements in the Contradictory group ($\overline{c}$) for that regulatory link. The number of elements in the Neutral group (n) for a predicted regulatory link corresponds to the number of gene expression experiments that neither confirm nor contradict that regulatory link but can provide additional information. If the expression information of a transcription factor or target gene from a regulatory link is absent in a gene expression experiment, then that experiment is ignored for that regulatory link, since nothing can be inferred unless the expression information of all regulatory elements is available.

For a positive gene regulation type (represented as 1) in a regulatory link, the target gene should be expressed if the transcription factor is expressed in an experiment. Therefore, if both the transcription factor and target gene from a predicted regulatory link are expressed in an experiment then this experiment is classified as Confirming. If the transcription factor is expressed but the target gene is not, then the experiment is considered to be Contradictory because the expressed transcription factor is not able to express the target gene as it should in the case of positive gene regulation. For a regulatory link, if the target gene is expressed but the transcription factor is not, then the experiment is marked as Neutral because it does not confirm or contradict the regulatory link. This experiment does allow that the target gene may be regulated by other transcription factors as well. The transcription factor not expressed in the experiment might be expressed in another experiment supporting the regulatory link. Additionally, if either the transcription factor or the target gene from a predicted regulatory link is not present in an experiment, then nothing can be inferred and that experiment is ignored for that regulatory link.

In the case of negative gene regulation (represented as -1), the target gene should not be expressed if the transcription factor is expressed in the experiment. Accordingly, if both the transcription factor and the target gene from a predicted regulatory link

Table 6.1: Rules for distributing the predicted regulatory links into three groups, Confirmed, Contradictory and Neutral

Regulation Type	TF expression	Gene expression	Result
1	Yes	Yes	C
1	Yes	No	$(\overline{C})$
1	No	Yes	N
1	No	No	N
-1	Yes	Yes	$(\overline{C})$
-1	Yes	No	C
-1	No	Yes	C
-1	No	No	$(\overline{C})$

are expressed or are not expressed, then the experiment is marked Contradictory. But for a predicted regulatory link, if the transcription factor is expressed and the target gene is not, or if the transcription factor is not expressed and the target gene is expressed, then the experiment is considered to be Confirming that regulatory link. Finally, each regulatory link is analyzed by identifying the number of times that link has been Confirmed, Contradicted and Neutral in all the gene expression experiments used. These values are then compared in different ways to evaluate the results for the regulatory links. This counting-based approach is needed because the limited number of experiments normally available for non-model organisms are usually insufficient to predict regulatory links from the expression data using statistical methods.

6.3 Results and Analyses

The two sets of predicted transcription factors based on the same family (TFf) and subfamily (TFsf), identified in transcription factor mapping (the first step of regulatory network mapping) are integrated with the four sets of predicted target genes (TGbs, TGblbs, TGpr, TGgalf) determined in the second step (target gene predic-

tion). The results of regulatory elements integration consist of six sets of regulatory links mapped from *S. cerevisiae* to *A. thaliana* as shown in Table 6.2. There are 14254 regulatory links present in the available regulatory network [38] of *S. cerevisiae* used for regulatory links mapping.

The predicted regulatory links for *A. thaliana* are then refined using gene expression data (an arbitrary first 43 experiments out of 513 that represent initial experiments) [56] to classify the gene expression experiments for each regulatory link into Confirming (C), Contradictory ($\overline{C}$), and Neutral (N) groups based on the rules described in the previous section. The total number of Confirming, Contradictory and Neutral values for a regulatory link to be analyzed should be equal to the number of gene expression experiments of *A. thaliana* used. Therefore, the number of links analyzed are the predicted links that have both transcription factor and target gene present in all the 43 gene expression experiments used. Then the Confirming (c), Contradictory ($\overline{c}$), and Neutral (n) values are compared in different ways in Table 6.2 to refine the predicted regulatory links using the 43 gene expression experiments for *A. thaliana*. The 43 gene expression experiments are of various different types: Mutants response to IAA (6 experiments), Zinc finger (2 experiments), Elevated CO2 (2 experiments), Auxin-induced lateral root formation (3 experiments), TMV infection (2 experiments), Response to systemic virus infection (2 experiments), T-DNA disruption of TTR1 (2 experiments), AGL15 overexpressor vs wildtype (2 experiments), CO2 stress (3 experiments), Cauliflower curd (2 experiments), Oleate hydroxylase overexpressor (2 experiments), target tissues and growth conditions for cDNA cloning (2 experiments), DST-mediated MRNA degradation (8 experiments) and Salicylic acid-dependent coupled regulation of cell growth and cell death (5 experiments).

In the fifth row, weights of two and one are assigned to the Contradictory and Neutral values to obtain a threshold for comparing it to the Confirming value. If the Confirming value is higher than this threshold for a regulatory link, then the link is

considered to be a verified true regulatory link mapped in the target genome, otherwise it is thrown out. This criterion is used to make sure that less than 25 percent of the gene expression experiments contradict any regulatory link. This is done to ignore any error due to outliers in the gene expression data. It is also evaluated based on the ratio between the Confirming and Contradictory values for the regulatory links, to determine for which links there is much greater evidence to support it than there is to contradict it. The Confirming value is compared with thrice and twice the value of Contradictory in the sixth and seventh rows of Table 6.2. The regulatory links satisfying these conditions are well supported by the gene expression experiments. In the eighth row, the number of regulatory links with a Confirming value higher than or equal to their Contradictory value but lower than twice the Contradictory value are determined. The regulatory links that meet this condition are not as well supported by the expression data. The Ninth row shows the number of regulatory links that have a Contradictory value higher than their Confirming value. These regulatory links are not supported by the gene expression data.

Comparing set TFsf-TGblbs and set TFf-TGblbs, about half of the analyzed regulatory links are verified and refined as true regulatory links for set TFsf-TGblbs, whereas only one-third of the analyzed regulatory links are verified for set TFf-TGblbs. The percentage of regulatory links analyzed with a Contradictory value more than the Confirming value increases from approximately 50 percent to about 66 percent from set TFsf-TGblbs to set TFf-TGblbs. This indicates that about two-thirds of the regulatory links analyzed in set TFf-TGblbs are definitely not supported by the gene expression experiments. Therefore, the results from the TFsf-TGblbs set are better before refinement, with a much higher percentage of verified mapped regulatory links in the target genome.

Comparing set TFsf-TGblbs and set TFsf-TGbs, the percentage of regulatory links analyzed with a Contradictory value more than the Confirming value is about 50

Table 6.2: Regulatory links refined for *Arabidopsis thaliana* using gene expression data for different sets of predicted regulatory links

	TFsf-TGblbs	**TFf-TGblbs**	**TFsf-TGbs**	**TFsf-TGpr**	**TFf-TGpr**	**TFf-TGgalf**
Number of mapped TFs	118	434	118	118	434	434
Number of mapped TGs	2252	2252	35181	25259	25259	25388
Number of links mapped	43423	480524	536154	274400	3182551	242030
Number of links analyzed	691	8621	6085	4465	61929	4380
$c \geq 2\bar{c} + n$ (filter threshold)	339	2628	2995	2343	18187	3727
$c \geq 3\bar{c}$	339	2375	2968	2323	16196	2012
$2\bar{c} \leq c < 3\bar{c}$	0	344	30	26	2651	1715
$\bar{c} \leq c < 2\bar{c}$	8	215	100	66	2221	503
$c < \bar{c}$	344	5687	2987	2050	40861	150

percent for both sets. This indicates that half of the regulatory links analyzed in these sets are definitely not supported by the gene expression experiments. The correctly mapped regulatory links are about half of the analysed links in the set TFsf-TGblbs and about one-third in the set TFf-TGpr. The percentage of regulatory links analyzed with a Contradictory value more than the Confirming value increases from approximately 50 percent to about 66 percent from set TFsf-TGblbs and set TFf-TGpr. Hence, set TFsf-TGblbs has better results than the other regulatory link sets, before refining the links, based on having a higher percentage of true regulatory links identified and a lower percentage of links contradicted more than they are confirmed.

About half of the regulatory links have been verified as true regulatory links in the result sets TFsf-TGblbs, TFsf-TGbs and TFsf-TGpr, about 85 percent in the result set TFf-TGgalf, 40 percent in the set TFf-TGpr and 30 percent in the set TFsf-TGblbs. This shows that, even though *S. cerevisiae* and *A. thaliana* are two organisms that are evolutionarily far apart, they still do share a significant amount of regulatory information among them. The transcription factor mapping result set TFf comprises the best results when we only consider how many transcription factors are mapped, but it does not produce the best regulatory links set when integrated with the predicted target gene set TGblbs. This indicates that the set TFsf containing the mapped transcription factors based on sequence similarity and subfamily classification contains the most efficiently and correctly mapped TFs for the purpose of mapping regulatory links.

The TGbs set, TGpr set and TGgalf set contain the best results of mapped target genes from the previous chapter, but they do not all work well when used in the integration of regulatory elements to predict regulatory links. The additional predicted target genes in these sets lead to too many false regulatory links in the sets TFsf-TGbs, TFf-TGpr, TFsf-TGpr, and TFf-TGgalf. These false links indicate that

many of the true target genes identified for the target genome in set TGbs, set TGpr and set TGgalf are, however, not the correctly mapped target genes linked to the right transcription factor in their corresponding regulatory link result sets. All the predicted TGs do contain the TFBS motifs for some transcription factor but these target genes need to correspond to the correct transcription factor, identifying the right regulatory link in the target genome. These results suggest that, in order to be used to map regulatory links, the target genes identified in the target genome using TFBS motifs also need to be similar in sequence to the source genome target genes, as identified in set TGblbs. Therefore, the results of other possible sets, TFf-TFbs and TFsf-TGgalf, are not included as their sets of regulatory elements are already determined to lead to many false positives when integrated to produce a predicted regulatory link.

6.4 Testing of filtered predicted regulatory links

The final six sets of predicted regulatory links have been determined by integrating the different sets of predicted transcription factors and target genes, followed by refinement of these predicted links using a small group of gene expression experiments as described in the previous sections. These predicted and refined result sets need to be validated as the regulatory links of the target genome. A large group of gene expression experiments is used to validate the predicted and refined links; this is an appropriate approach for validation since predicting gene expression is a key use of regulatory networks. This set of gene expression data is distinct from the group of gene expression experiments used in the previous section to refine the predicted links. The links are considered confirmed based on the following two rules when verified against every gene expression experiment as shown in Figure 6.2. A predicted regulatory link, with positive regulation type, is considered confirmed if

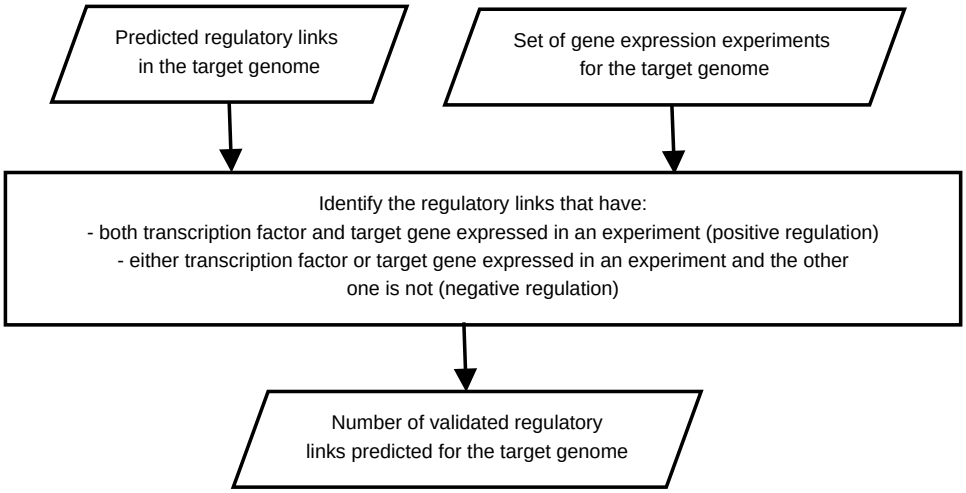


Figure 6.2: Method to validate the integrated and refined regulatory links for the target genome

both the transcription factor and the target gene from that link are expressed in the experiment. A negative regulation type of predicted regulatory link is considered confirmed if either the transcription factor or the target gene from that link is expressed in the experiment and the other one is not. These rules are used to determine if a target gene predicted is the right target gene for the corresponding predicted transcription factor in a predicted regulatory link. The neutral regulatory links for positive regulation type are not included as confirmed regulatory links in this testing but they also are not contradictory to the gene expression data.

The predicted and refined regulatory links from the previous section are validated as confirmed expression level relationships for *A. thaliana*. The validation is done based on the rest of the publicly available gene expression data (470 out of 513 experiments) [56] for *A. thaliana*, using the same rules for confirmation used in the refinement process in the previous section. The number of regulatory links that were validated in all 470 experiments (result A) and in at least 100, 200, 300, 350 or 400 experiments (result B) out of the 470 experiments are determined. The result set B factors in possible noise in the gene expression data to validate the predicted

regulatory links. Also, not every regulatory link tested will have both the transcription factor and target gene present in every gene expression experiment. Results for testing the predicted and refined regulatory links against 470 gene experiments are shown in Table 6.3.

Comparing links confirmed in all 470 experiments (result A) to links confirmed in at least 400, 350, 300, 200 and 100 experiments (result B), many more regulatory links are validated in the latter. This shows that it is quite likely that some regulatory links are not validated in result A due to noise in the gene expression data used.

Most of the predicted regulatory links have been validated when confirmed by at least 300 experiments for the result sets TFsf-TGblbs, TFsf-TGbs and TFsf-TGpr and about 75 percent are validated for the result sets TFf-TGblbs and TFf-TGpr. However, very few regulatory links are predicted and validated for result set TFsf-TGblbs compared to result sets TFsf-TGbs and TFsf-TGpr. Going from 100 up to 200, to 300, to 350 experiments, the number of confirmed regulatory links do not change much for all the result sets, except TFf-TGgalf, although there is a huge change in the confirmed regulatory links from 350 to 400 experiments. For result set TFf-TGgalf, most of the predicted regulatory links have been confirmed for at least 200 experiments but none for at least 300 or more experiments. Therefore, similar numbers of predicted regulatory links in result sets TFsf-TGbs and TFsf-TGpr are very well validated by the gene expression data used, and are thus accurately predicting a large proportion of the gene expression results. It implies that not just the correctly predicted but also the right target genes have been mapped along with the right transcription factors to form regulatory links in these result sets, and that the overall method has been very successful at accurately predicting regulatory links.

The TGbs set, TGpr set and TGgalf set contain the best results of mapped target genes from the previous chapter, but TGgalf set does not work well when used in the integration of regulatory elements to predict regulatory links. The additional

Table 6.3: Confirmation of filtered predicted regulatory links from the six different result sets for *Arabidopsis thaliana* using a group of 470 gene expression experiments

	TFsf-TGblbs	**TFf-TGblbs**	**TFsf-TGbs**	**TFsf-TGpr**	**TFf-TGpr**	**TFf-TGgalf**
Number of mapped TFs	118	434	118	118	434	434
Number of mapped TGs	2252	2252	35181	25259	25259	25388
Number of filtered predicted links	339	2628	2995	2343	18187	3727
Links confirmed in all 470 experiments	0	0	0	0	0	0
Links confirmed in at least 400 experiments	28	228	189	132	1123	0
Links confirmed in at least 350 experiments	290	1535	2219	1707	9473	0
Links confirmed in at least 300 experiments	337	2019	2845	2215	13417	0
Links confirmed in at least 200 experiments	339	2494	2986	2340	17359	3457
Links confirmed in at least 100 experiments	339	2628	2995	2343	18187	3727

predicted target genes in all these sets lead to too many false regulatory links in the sets TFsf-TGbs, TFf-TGpr, TFsf-TGpr, and TFf-TGgalf. These false links indicate that many of the true target genes identified for the target genome in set TGbs, set TGpr and set TGgalf are, however, not the correctly mapped target genes linked to the right transcription factor in their corresponding regulatory link result sets. Refinement of regulatory links using gene expression data helps in getting rid of most of these false regulatory links.

The transcription factor mapping result set TFf comprises the best results when we only consider how many transcription factors are mapped, but it does not produce the best regulatory links set when integrated with the predicted target gene set TGblbs. This indicates that the set TFsf containing the mapped transcription factors based on sequence similarity and subfamily classification contains the most efficiently and correctly mapped TFs for the purpose of mapping regulatory links.

6.5 Testing of predicted regulatory links thrown out during filtration step

In the previous section, the filtered predicted regulatory links from Table 6.2 were tested against the gene expression data for validation. Most of the links were validated well in three result sets TFsf-TGblbs, TFsf-TGbs and TFsf-TGpr as shown in Table 6.3. This shows that refinement of predicted links eliminates most of the false positives but there is also a need to determine the number of true positives lost during the refinement process as part of the thrown out regulatory links. The purpose of testing the thrown out predicted links is to determine the significance of the filtering process for predicted regulatory links. The lower the number of links validated among the thrown out predicted regulatory links, the better the results of the filtering process. Therefore, in this section, the predicted regulatory links

that were thrown out during filtration are tested against the same gene expression data percentage of regulatory links validated here as compared to the percentage of regulatory links validated for the filtered predicted links for the same result set that was used for testing the filtered regulatory links. The method used for testing is the same as in the previous section as shown in Figure 6.2.

The thrown out regulatory links from the highly validated result sets TFsf-TGblbs, TFsf-TGbs and TFsf-TGpr are used, and the results are shown in Table 6.4. For all the three result sets, less than 11 percent of regulatory links are confirmed in at least 100 experiments, less than 2.5 percent are confirmed in at least 200 experiments, and none in 400 or more experiments. This is a very small percentage of regulatory links validated here as compared to the percentage of regulatory links validated for the filtered predicted links for the same result set in the previous section. This illustrates that it is extremely effective to filter the predicted links to get rid of most of the false positives since so few true positives are lost.

6.6 Conclusion

A two step approach has been proposed to determine the regulatory links in the target genome. This includes integrating the predicted transcription factors and the target genes of the target genome into predicted regulatory links and then refining these regulatory links using the gene expression data set of the target genome. Finally, the refined regulatory links are validated using another set of gene expression data to verify if the predicted transcription factor is integrated with the correct target to form a regulatory link. Six sets of regulatory links are obtained after regulatory elements integration and regulatory links refinement. The first set combines the mapped transcription factors based on sequence similarity and protein family classification with the target genes based on sequence similarity and searching TFBS

Table 6.4: Validation of the thrown out predicted regulatory links during filtration from result set TFsf-TGpr for *Arabidopsis thaliana* using a group of 470 gene expression experiments

	TFsf-TGblbs	**TFsf-TGbs**	**TFsf-TGpr**
Number of predicted links thrown out in filtration	352	3090	2122
Links confirmed in all 470 experiments	0	0	0
Links confirmed in at least 400 experiments	0	0	0
Links confirmed in at least 350 experiments	1	9	9
Links confirmed in at least 300 experiments	5	38	38
Links confirmed in at least 200 experiments	8	52	50
Links confirmed in at least 100 experiments	34	323	231

motifs in the nucleotide sequence database. The second set combines the mapped transcription factors based on sequence similarity and protein subfamily classification with the target genes based on sequence similarity and searching TFBS motifs in the nucleotide sequence database. The mapped transcription factors based on sequence similarity and protein family classification are integrated with the target genes based on searching TFBS motifs in the nucleotide sequence database in the third set. The fourth set combines the mapped transcription factors based on sequence similarity and protein subfamily classification with the target genes based on searching TFBS in the promoter sequence database. The mapped transcription factors based on sequence similarity and protein family classification are integrated with the target genes based on searching TFBS motifs in the promoter sequence database in the fifth set. The sixth set combines the mapped transcription factors based on sequence similarity and protein family classification with the target genes based on

identifying and then searching TFBS motifs in the promoter sequence database. In the results, the large number of target genes identified using the preferred method for target gene mapping produce many false regulatory links, since, while they are target genes, they are not linked to the correct transcription factor.

Additionally, the transcription factors from the preferred method of transcription factor mapping also contribute to many false regulatory links when used in the regulatory elements integration step.

Hence, the predicted regulatory links obtained by integrating the mapped transcription factors based on sequence similarity and protein sub-family classification and mapped target genes based on searching TFBS motifs in the nucleotide sequence database or promoter sequence database followed by the refinement process contain the most regulatory links for the target genome that are verified during testing using the gene expression data. This implies that more correctly mapped target genes that link to the right transcription factor are determined by searching for TFBS motifs. Also, the correctly mapped transcription factors are obtained using the method based on sequence similarity and protein sub-family classification. This suggests that regulatory relationships are conserved between different genomes and can be mapped between them using the relationship between the gene sequences. Therefore, for a newly sequenced organism, a related model organism can be used to determine some regulatory information for the lesser explored organism, avoiding the need to complete significant expression experiments for the target organism.

The use of TFBS information in finding target genes, while showing the best final results, does cause some errors. Even when the target genes contain the regulatory motifs of certain transcription factors, these genes may not be regulated by these transcription factors at all. This is because only a very small percentage of binding sites are available to bind to interact with the transcription factors, as mentioned earlier in the limitation of regulatory networks section. False positives in our results

could then be reduced by integrating information about the active binding sites in the target genes. Considering alternative sources of TFBS information could also decrease the false negatives, as we discovered that the lack of an instance of the appropriate binding site motif did not always correlate to lack of regulation in the source genome.

The validation of refined and predicted regulatory links illustrated that most of the regulatory links are highly confirmed in four out of six result sets. About half of the regulatory links were validated in the result set that integrated mapped transcription factors based on sequence similarity and protein family classification with the target genes based on searching TFBS motifs in the promoter sequence database. There were no regulatory links validated based on at least 300 gene expression experiments for the result set obtained by combining the mapped transcription factors based on sequence similarity and protein family classification with the target genes based on searching for TFBS motifs identified using GALF-P in the promoter sequence database. The highest percentage of predicted regulatory links that were confirmed were from two result sets: combination of mapped transcription factors based on sequence similarity and protein sub-family classification with the predicted target genes based on searching TFBS motifs in the sequence database, and combination of mapped transcription factors based on sequence similarity and protein sub-family classification with the predicted target genes based on searching TFBS motifs in the promoter sequence database.

Very few predicted and validated regulatory links were thrown out during filtration. This suggests that filtration of predicted regulatory links is a critical step in the prediction of regulatory links. This step helps to eliminate most of the false positives and loses very few true positives, even though very little gene expression data were used to filter. More biological and regulatory information can also be integrated further into the regulatory network mapping method as more data become available

for non-model organisms.

Chapter 7

Conclusion

7.1 Summary

The goal of this thesis is to infer gene regulatory links for a non-model organism. Determination of these regulatory links that can provide in-depth knowledge about the new organism being studied produces a major challenge. Over time, many statistical methods [27, 76] have been developed that use gene expression data to determine regulatory networks for an organism. Since these methods need a very large amount of data to get reliable results, they work well for model organisms that have large amounts of data available. Non-model organisms, on the other hand, have very limited information available, which is not enough to be used by these methods to get definitive regulatory network information. Therefore, integrating the limited information available for the non-model organism along with the large experimental data available for a related model organism to predict regulatory links for the non-model organism is a potential solution to this issue, but itself is a significantly challenging problem.

To address this challenge with the proposed method in this thesis, regulatory links data, transcription factor data and target gene data for a model organism are used

along with the sequence data and limited gene expression data available for the target genome. First the transcription factors and target genes are mapped from the model organism to the non-model organism. Next, the predicted transcription factors and target genes for the non-model organism are then integrated together to form regulatory links. And finally, these integrated regulatory links for the non-model organism are filtered based on a set of rules using a very small set of gene expression data (43 out of 513 experiments) for the non-model organism. To test the proposed method, experiments are performed to map the regulatory network data of *Saccharomyces cerevisiae* to *Arabidopsis thaliana* and the results are analyzed. For testing the predicted and filtered regulatory links, the refined regulatory links are validated using a very large set of gene expression data (470 out of 513 experiments) to verify if the predicted transcription factor is integrated with the correct target to form a regulatory link. The thrown out links from the filtration process for the three best regulatory link sets are also tested against the same set of gene expression data. This test is performed to determine the number of true regulatory links lost in the filtration step.

The step of mapping transcription factors using three different techniques is based on sequence similarity, protein family classification and protein sub-family classification. The results show that the technique based on sequence similarity and protein family classification maps transcription factors most effectively from *Saccharomyces cerevisiae* to *Arabidopsis thaliana*; therefore, it is the preferred method for general transcription factor mapping.

Experiments are also performed on bacterial organisms based on evolutionary distance to compare the predicted transcription factors determined using two similar tools, based on protein sequence similarity alone and distant evolutionary relationships among sequences. Analysis of the results suggests that transcription factors can be mapped from one bacterial organism to another, as transcription factor motifs

are well preserved among these organisms. Results are also analysed to determine the most suitable threshold for the e-value parameter of the tools that can be used to map transcription factors and the appropriate tool to use. Both the e-value threshold of the tool and evolutionary distance from the model organism used for mapping have significant impact on the quality of results.

For mapping target genes from the model to the non-model organism, six techniques are used based on sequence similarity, TFBS motifs, active binding sites and sequence similarity along with TFBS motifs. Out of the six methods, the one based on searching only TFBS motifs is the preferred method for predicting target genes since most of the target genes are determined correctly for *Arabidopsis thaliana*. However, two out of the three methods using the *Arabidopsis thaliana* promoter sequence database for searching TFBS motifs have much lower false positives. If a non-model organism does not have promoter database available, then one of the available promoter prediction tools [**?**] can be used to determine this information. Additionally, the GALF-P tool, determined to be the best tool in a recent comparative analysis research, did not give the best results in the overall process of mapping target genes.

In the integration step, six sets of regulatory links are obtained after combining the predicting sets of transcription factors and target genes for the target genome, *Arabidopsis thaliana*. These integration results show that the large number of target genes identified using the preferred method for target gene mapping produce many false regulatory links, since, while they are target genes, they are not linked to the correct transcription factor. Additionally, the GALF-P tool, determined to be the best tool in a recent comparative analysis research, did not give the best results in the overall process of mapping target genes. This illustrates that it is very important to choose the techniques for predicting transcription factors and target genes that produce the best results for the overall method instead of choosing the techniques that work best for an individual step. Finally, most of the false links are then suc-

cessfully separated from the true regulatory links by filtering them using a small amount of gene expression data.

The validation of refined and predicted regulatory links illustrated that most of the regulatory links are confirmed in three out of six result sets. About half of the regulatory links were confirmed in the result set that integrated mapped transcription factors based on sequence similarity and protein family classification with the target genes based on searching TFBS motifs in the promoter sequence database. There were no regulatory links confirmed based on at least 300 gene expression experiments for the result set obtained by combining the mapped transcription factors based on sequence similarity and protein family classification with the target genes based on using GALF-P to identify TFBS motifs in the promoter sequence database. The highest percentage of predicted regulatory links that were comfirmed were from these three result sets: combination of mapped transcription factors based on sequence similarity and protein sub-family classification with the predicted target genes based either on searching TFBS motifs in the sequence database, or on searching TFBS motifs in the promoter sequence database, or on sequence similarity and searching TFBS motifs in the sequence database.

Finally, when the links thrown out during filtration are tested against the same very large set of gene expression data used earlier for testing filtered regulatory links, very few regulatory links are validated. This demonstrates that filtration of predicted regulatory links is a significant step in the prediction of regulatory links. This step gets rid of most of the false positives in the predicted regulatory links and still does not lose the true positives. This use of a small arbitrary set of expression data, together with the filtration rules, is thus demonstrated to be very effective at determining which links should be kept for the target organism.

An integrated workflow has been designed and developed for regulatory network mapping from a model organism to a non-model organism. Our workflow for identi-

fying regulatory links for a non-model organism includes three steps: transcription factor mapping, target gene prediction, and regulatory elements integration and filtration. Three different techniques were designed for transcription factor mapping and six different techniques were devised for prediction of target genes and are implemented using currently available bioinformatics tools and databases. For both these phases, the result sets are analyzed and compared to identify the best technique for the individual step and for the complete workflow. When deciding on choosing a specific technique, it is important that the ones that work the best overall should be picked and not the techniques that are shown best in individual steps. Now that the transcription factors and target genes have been predicted for the non-model organism in the first two steps of the workflow, the transcription factors need to be combined with the correct target gene to get similar regulatory relationship in the target genome. This is the final step in the workflow to integrate the mapped transcription factors and target genes into predicted regulatory links and filter out the false positives. In this step, rules are constructed that are applied on very limited gene expression data of the non-model organism in the filtration technique to get rid of most of the false positives without losing much true positives. This final step of the workflow is computationally intensive, so was implemented in parallel on cluster computing facilities. The complete workflow for regulatory network prediction is then tested using the publicly available data of Yeast and *Arabidopsis thaliana*. We get a substantial proportion of predicted regulatory links for the target organism that are very consistent with the complete gene expression data used for testing. The thrown out regulatory links from the filtration process, part of the final step of the workflow, are not consistent with the same set of gene expression data, illustrating the significance of the filtration step to eliminate false positives. This proposed workflow thus makes very effective use of an extremely limited amount of experimental data for the target organism, alleviating the problems faced when investigating

non-model organisms whose limited data prevents the use of statistical techniques. In the proposed workflow, any model organism with an available experimentally confirmed regulatory network and transcription factor binding site (TFBS) motif information can be used as the source genome. However, a model organism closer to the target genome is preferable, since evolutionarily closer organisms will tend to have more similar regulatory relationships. The target genome can be any non-model organism with available nucleotide sequence data. For experimentation and testing purposes *Saccharomyces cerevisiae* and *Arabidopsis thaliana* have been used in this thesis as source and target genomes, respectively. *Saccharomyces cerevisiae* is a model organism with all the required information required for a source genome. *Arabidopsis thaliana*, not being a non-model organism, is used as the target genome for experimentation purposes only, so that the gene expression data available for *Arabidopsis thaliana* can be used for analysis and verification of the mapped regulatory relationships determined using the proposed method. *Arabidopsis thaliana* is also a suitable test organism to investigate regulatory inference for plants, for which there are far fewer well-investigated examples and thus are likely targets of the technique presented in this thesis. While closer organisms are preferred, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are used as the source and target genomes since there is no suitable pair of organisms that are evolutionarily close that have both sufficient regulatory network and gene expression information available for verification. This non-ideal situation can nevertheless be common for many non-model organisms, due to the limited availability of well-investigated model organisms, and the results in this thesis thus identify the proportion of regulatory relationships that can be mapped from an evolutionarily distant model organism.

To further understand the gene regulation process, these predicted transcription gene regulatory links can be integrated with the translation regulatory data in the gene expression process and any additional biological information for non-model organism

can be incorporated as well. Finally, the combined regulatory gene expression links for a non-model organism can be integrated later with the organism's metabolic network information to understand the two systems together which are mostly studied separately.

7.2 Thesis Contributions

The goal of this thesis work is to demonstrate how a model organism can be used as a model to infer regulation in other organisms with minimal data available. Understanding gene regulation at the genome level is a very significant challenge for biologists in the systems biology field. It can help in identifying abnormal gene variants that can be the potential cause of various diseases. Studying a very small set of genes in an organism is performed by conducting large scale lab experiments which take a lot of time and effort. Due to the work required, it is not feasible to perform all these experiments at the genome level with thousands of sequences. Until now, researchers have been able to design and use many statistical techniques to determine the regulatory information for the model organisms using a very large amount of experimental data (gene expression data) that is easily available, although it is not possible to use the same statistical techniques to identify regulatory information for non-model organisms that have very limited data available. For example, about 3,75,000 plant species are known, with more being discovered and the researchers cannot use the currently available statistical techniques to understand the regulatory relationships in all these plants species. Therefore, it would be extremely useful to have a method that can use a model organism to identify many regulatory relationships in all the related species without doing all the experimentation for every gene in every species.

The proposed method in this thesis provides a new ability to use a model organism's

data to determine the regulatory information for a non-model organism. Proposed method as a whole is novel, and the different stages in this method are unique as well. A new insight is found as to which technique works better for different stages of this method. In the first stage, three different techniques are designed using bioinformatics tools to map transcription factors from the model organism to the non-model organism. A large number of transcription factors for the target genome are predicted correctly using these techniques designed here. However, the technique that gives the best results for transcription factor mapping stage does not turn out to be best technique for the complete method of mapping regulatory links.

Additionally, the most suitable parameter value is identified for one of the tools and this tool parameter is mostly used by researchers with a default value. The experiments performed for mapping transcription factor between bacterial genomes also identified the suitable e-value threshold parameter for the BLAST tool along with the suitable model organism to be used. It demonstrated the importance of using the correct e-value threshold over choosing the evolutionarily closer model organism although researchers tend to use the evolutionarily closer model organism instead of changing default parameter value of the tool.

In the next stage of the proposed method, the target genes of the non-model organism are predicted using different techniques that are devised based on using sequence similarity and binding site motifs. Many mapped target genes are predicted correctly for the target genome using three out of the six techniques designed and compared here. The best technique identified in this individual stage does not turn out to be the best for the overall process of mapping regulatory links. Therefore, it was important to identify the best techniques for transcription factor and target gene mapping that gave the best results for the overall process of mapping regulatory links.

In the final stage of integration and filtration, the regulatory elements are combined based on the binding site motif information available for the model organism and

then regulatory links are filtered based on the new rules using limited gene expression data of the non-model organism. The testing and analysis of the results confirmed that the final set of integrated and filtered regulatory links identified as predicted regulatory links for *Arabidopsis thaliana* were consistent with the very large set of gene expression data used for testing. It also confirmed that the regulatory links thrown out in filtration process were not consistent with the same gene expression data set. This demonstrates that the predicted transcription factors are the true transcription factors that are combined with the right predicted target gene to obtain the validated regulatory links for the target genome getting rid of a large number of false positives without losing much true positives. This new capability enables the biologist to investigate a non-model organism and use a model organism to infer regulatory information for a non-model organism despite having very little expression data.

References

[1] Y. Akiyama, *TFSEARCH: Searching Transcription Factor Binding Sites*, http://www.rwcp.or.jp/lab/pdappl/papia.html.

[2] T. Akutsu, S. Kuhara, and S. Miyano, *Identification of genetic networks from a small number of gene expression patterns under the Boolean network model*, Proceedings of the 4th Pacific Symposium on Biocomputing 4 (1999), 17-28.

[3] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology (Second Edition)*, Garland Science Publishing and Sumanas, Inc., 2004.

[4] E. Almasri, P. Larsen, G. Chen, and Y. Dai, *Incorporating Literature Knowledge In Bayesian Network For Inferring Gene Networks With Gene Expression Data*, Bioinformatics Research and Applications: Lecture Notes in Computer Science 4983 (2008), 184-195.

[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool*, Journal of Molecular Biology 215 (1990), 403-410.

[6] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Research 25 (1997), 3389-3402.

[7] T.L. Bailey, and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology 2 (1994), 28-36.

[8] V.B. Bajic, S.H. Seah, A. Chong, G. Zhang, J.L. Koh, and V. Brusic, *Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters*, Bioinformatics 18 (2002), 198-199.

[9] W.C. Barker, J.S. Garavelli, P.B. McGarvey, C.R. Marzec, B.C. Orcutt, G.Y. Srinivasarao, L.L. Yeh, R.S. Ledley, H. Mewes, F. Pfeiffer, A. Tsugita, and C. Wu, *The PIR-International Protein Sequence Database*, Nucleic Acids Research 27 (1999), 39-43.

[10] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler, *Genbank*, Nucleic Acids Research 36 (2008), D25-D30.

[11] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, *The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003*, Nucleic Acids Research 31 (2003), 365-370.

[12] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G.G. Lara, A. Oezcimen, P. Rocca-Serra, and S.A. Sansone, *Arrayexpress-a public repository for microarray gene expression data at the EBI*, Nucleic Acids Research 31 (2003), 68-71.

[13] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil,A. Santos-Zavaleta, M.I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado, C. Bonavides-Martínez, C. Abreu-Goodger, C. Rodríguez-Penagos, J. Miranda-Ríos, E. Morett, E. Merino, A.M. Huerta, L.

Treviño-Quintanilla, and J. Collado-Vides, *RegulonDB version 6.0: gene regulation model of Escherichia coli k-12 beyond transcription, active annotated promoters and textpresso navigation*, Nucleic Acids Research 36 (2008), D120-D124.

[14] T. Chan, K. Leung, and K. Lee, *TFBS identification based on genetic algorithm with combined representations and adaptive post-processing*, Bioinformatics 24 (2008), 341-349.

[15] T. Chen, H.L. He and G.M. Church, *Modeling gene expression with differential equations*, Proceedings of the 4th Pacific Symposium on Biocomputing 3 (1999), 29-40.

[16] *Chromosomes and Cell Division: Genes*, December 2011, Graphics Gallery: http://www.accessexcellence.org/RC/VL/GG/genes.php.

[17] G. Cochrane, P. Aldebert, N. Althorpe, M. Andersson, W. Baker, A. Baldwin, K. Bates, S. Bhattacharyya, P. Browne, A. van den Broek, M. Castro, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, C. Kanz, T. Kulikova, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, M. McHale, H. McWilliam, G. Mukherjee, F. Nardone, M.P. Pastor, S. Sobhany, P. Stoehr, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler, *EMBL Nucleotide Sequence Database: developments in 2005*, Nucleic Acids Research 34 (2006), D10-D15.

[18] *CREAD: Comprehensive regulatory element analysis and discovery*, January 2008, http://rulai.cshl.edu/cread/.

[19] R.V. Davuluri, H. Sun, S.K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, and E. Grotewold, *AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors*, BMC Bioinformatics 4 (2003).

[20] M.J.L. de Hoon, S. Imoto and S. Miyano, *Inferring gene regulatory networks from time-ordered gene expression data using differential equations*, Proceedings of the 5th International Conference on Discovery Science 5 (2002), 267-274.

[21] A. Gambin, S. Lasota, and M. Rutkowski, *Analyzing stationary states of gene regulatory network using petri nets*, Silico Biology 6 (2006), 93-109.

[22] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, and W.M. Gelbart, *An Introduction to Genetic Analysis. 7th edition.*, W. H. Freeman, New York, 2000.

[23] C. Guan, C. Ye, X. Yang, J. Gao, *A review of current large-scale mouse knockout efforts.*, Genesis 48 (2010), 73-85.

[24] A. Guo, K. He, D. Liu, S. Bai, X. Gu, L. Wei, and J. Luo, *DATF: a Database of Arabidopsis Transcription Factors*, Bioinformatics 21 (2005), 2568-2569.

[25] A.Y. Guo, X. Chen, G. Gao, H. Zhang, Q.H. Zhu, X.C. Liu, Y.F. Zhong, X. Gu, K. He, and J. Luo, *PlantTFDB: a comprehensive plant transcription factor database*, Nucleic Acids Research 36 (2008), D966-D969.

[26] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young, *Transcriptional regulatory code of a eukaryotic genome*, Nature 431 (2004), 99-104.

[27] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, *Gene regulatory network inference: data integration in dynamic models-a review*, Biosystems 96 (2009), 86-103.

[28] T. Heinemeyer, E. Wingender, I. Reuter, H. Hermjakob, A.E. Kel, O.V. Kel, E.V. Ignatieva, E.A. Ananko, O.A. Podkolodnaya, F.A. Kolpakov, N.L.

Podkolodny, and N.A. Kolchanov, *Databases on Transcriptional Regulation: TRANSFAC, TRRD, and COMPEL*, Nucleic Acids Research 26 (1998), 362-367.

[29] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, *Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks*, Journal of Bioinformatics and Computational Biology 2 (2004), 77-98.

[30] *Iterated profile searches with PSI-BLAST*, 2013, http://www.ncbi.nlm.nih.gov /BLAST/tutorial/Altschul-2.html.

[31] Jakubowski, *Chapter 5 - Binding, D: Binding and the control of gene transcription*, April 2013, http://employees.csbsju.edu/hjakubowski/classes/ch331/bind/olbindtransciption.html.

[32] S.T. Jensen, and J.S. Liu, *BioOptimizer: a Bayesian scoring function approach to motif discovery*, Bioinformatics 20 (2004), 1557-1564.

[33] C. Jiang, Z. Xuan, F. Zhao, and M.Q. Zhang, *TRED: a transcriptional regulatory element database, new entries and other development*, Nucleic Acids Research 35 (2007), D137-D140.

[34] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, *The KEGG databases at GenomeNet*, Nucleic Acids Research 30 (2002), 42-46.

[35] A.E. Kel, E. Gößling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, and E. Wingender, *MATCHTM: a tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Research 31 (2003), 3576-3579.

[36] S. Knudsen, *Promoter 2.0: for the recognition of PolII promoter sequences.*, Bioinformatics 15 (1999), 356-361.

[37] W.B. Langdon, *Receiver Operating Characteristics (ROC)*, 2006, http://www.cs.ucl.ac.uk/staff/W.Langdon/roc/.

[38] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, and R.A. Young, *Transcriptional regulatory networks in saccharomyces cerevisiae*, Science 298 (2002), 799-804.

[39] B. Lewin, *Genes VII*, Oxford University Press Inc., New York, 2000.

[40] S. Liang, S. Fuhrman, and R. Somogyi, *Reveal, a general reverse engineering algorithm for inference of genetic network architectures*, Proceedings of the 3rd Pacific Symposium on Biocomputing 3 (1998), 18-29.

[41] X. Liu, D.L. Brutlag, and J.S. Liu, *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.*, Pacific Symposium on Biocomputing 6 (2001), 127-138.

[42] G. Liua, W. Fenga, H. Wanga, L. Liua, and C. Zhou, *Reconstruction of Gene Regulatory Networks Based on Two-Stage Bayesian Network Structure Learning Algorithm*, Journal of Bionic Engineering 6 (2009), 86-92.

[43] H. Matsuno, M. Nagasaki, and S. Miyano, *Hybrid petri net representation of gene regulatory network*, Proceedings of the 5th Pacific Symposium on Biocomputing 5 (2000), 341-352.

[44] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P.S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J.

McDowall, J. Mistry, A. Mitchell,, A.N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J.D. Selengut, C.J.A. Sigrist, P.D. Thomas, F. Valentin, D. Wilson, C.H. Wu, and C. Yeats, *New developments in the InterPro database*, Nucleic Acids Research 35 (2007), D224-D228.

[45] S. Nargundkar, and J.L. Priestley, *Assessment of evaluation methods for binary classification modeling*, Proceedings of the 2003 Decision Sciences Institute National Conference (2003), 1-6.

[46] I. Nasidze, J. Li, D. Quinque, K. Tang, and M. Stoneking, *Global diversity in the human salivary microbiome*, Genome Research 19 (2009), 636-643.

[47] *What is Nucleosomes*, 2011, http://www.newmediway.com/what-is-nucleosomes.

[48] O. Ogunseitan, *Phylogenetic Tree*, December 2008, http://bioweb.uwlax.edu/bio203/s2007/falk_pete/phylogenetic_tree.htm.

[49] V. Olman, H. Peng, Z. Su, and Y. Xu, *Mapping of microbial pathways through constrained mapping of orthologous genes*, Proceedings of the IEEE Computational Systems Bioinformatics Conference (2004), 363-370.

[50] S.K. Palaniswamy, S. James, H. Sun, R.S. Lamb, R.V. Davuluri, and E. Grotewold, *Agris and AtRegNet: A platform to link cis-regulatory elements and transcription factors into regulatory networks*, Plant Physiology 140 (2006), 818-829.

[51] D.S. Prestridge, *Predicting Pol II Promoter Sequences Using Transcription Factor Binding Sites.*, Journal of Molecular Biology 249 (1995), 923-932.

[52] R.C. Périer, T. Junier, and P. Bucher, *The Eukaryotic Promoter Database EPD*, Nucleic Acids Research 26 (1998), 353-357.

[53] *Promoters and Terminators*, 2013, http://molbiol-tools.ca/Promoters.htm.

[54] P. Puntervoll, and R. Aasland, *Basic Sequence Analysis Using BLAST Tutorial*, 2008, http://hpc.ilri.cgiar.org/Tutorials/Blast_analysis.html.

[55] S.A. Rensing, D. Lang, E. Schumann, R. Reski, and A. Hohe, *EST sequencing from embryogenic Cyclamen persicum cell cultures identifies a high proportion of transcripts homologous to plant genes involved in somatic embryogenesis*, Journal of Plant Growth Regulation 24 (2005), 102-115.

[56] S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang, *The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.*, Nucleic Acids Research 31 (2003), 224-228.

[57] A. Sandelin, W. Alkema, P. Engström, W.W. Wasserman, and B. Lenhard, *JASPAR: an open-access database for eukaryotic transcription factor binding profiles*, Nucleic Acids Research 32 (2004), D91-D94.

[58] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thaström, Y. Field, I.K. Moore, J.Z. Wang, and J. Widom, *A Genomic Code for Nucleosome Positioning*, Nature 442 (2006), 772-778.

[59] C.E. Shannon, and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1963.

[60] R. Sharma, P.A. Evans, and V.C. Bhavsar, *Mapping a Regulatory Network Between Organisms*, Proceedings of the Fourth International Conference on Computational Systems Biology (2010), 204-218.

[61] R. Sharma, P.A. Evans, and V.C. Bhavsar, *Mapping Transcription Factors from a Model to a non-Model Organism*, Proceedings of the International Joint Con-

ferences on Bioinformatics, Systems Biology and Intelligent Computing (2009), 160-166.

[62] R. Sharma, P.A. Evans, and V.C. Bhavsar, *Transcription factor mapping between bacteria genomes*, International Journal of Functional Informatics and Personalised Medicine 2 (2009), 424-441.

[63] R. Sharma, P.A. Evans, and V.C. Bhavsar, *Regulatory link mapping between organisms*, BMC Systems Biology 5 (June 20, 2011), Supplement 1:S4, available at http://www.biomedcentral.com/1752-0509/5/S1/S4, 12 pages.

[64] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, M. Kaloper, S. Weng, H. Jin, C.A. Ball, M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, and J.M. Cherry, *The Stanford Microarray Database*, Nucleic Acids Research 29 (2001), 152-155.

[65] D. Swigon, *Genome Sizes*, December 2010, http://www.math.pitt.edu/∼*swigon/Lectures/Lecture*1*.pdf*.

[66] M.C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A.R. Fernandes, N.P. Mira, M. Alenquer, A.T. Freitas, A.L. Oliveira, and I. Sá-Correia, *The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae*, Nucleic Acids Research 34 (2006), D446-D451.

[67] P.D. Thomas, M.J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania, *PANTHER: a library of protein families and subfamilies indexed by function.*, Genome Research 13 (2003), 2129-2141.

[68] T. Tsunoda, and T. Takagi, *Estimating Transcription Factor Bindability on DNA.*, Bioinformatics 15 (1999), 622-630.

[69] R. Twyman, *The human genome: Gene expression*, 2003, http://genome.wellcome.ac.uk/doc_WTD020757.html.

[70] M. Vilela, I.C. Chou, S. Vinga, A.T. Vasconcelos, E.O. Voit, and J.S. Almeida, *Parameter optimization in S-system models*, BMC Systems Biology 16 (2008).

[71] J. Villard, *Transcription regulation and human diseases*, Swiss Medical Weekly 134 (2004), 571-579.

[72] E.O. Voit, *Modelling metabolic networks using power-laws and S-systems*, Essays in Biochemistry 45 (2008), 29-40.

[73] A. Wagner, *A computational genomics approach to the identification of gene networks*, Nucleic Acids Research 25 (1997), 3594-3604.

[74] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Transactions on Acoustic, Speech, and Signal Processing 37 (1989), 328-339.

[75] Z. Wei1, and S.T. Jensen, *GAME: detecting cis-regulatory elements using a genetic algorithm*, Bioinformatics 22 (2006), 1577-1584.

[76] L.F. Wessels, E.P. van Someren, and M.J. Reinders, *A comparison of genetic network models*, Proceedings of the 6th Pacific Symposium on Biocomputing 6 (2001), 508-519.

[77] D. Wilson, V. Charoensawan, S.K. Kummerfeld, and S.A. Teichmann, *DBD - taxonomically broad transcription factor predictions: new content and functionality*, Nucleic Acids Research 36(Database issue) (2008), D88-D92.

[78] E.M. Zdobnov, and R. Apweiler, *InterProScan - an integration platform for the signature-recognition methods in InterPro.*, Bioinformatics 17 (2001), 847-848.

Vita

Candidate's full name:

Rachita Sharma

University attended:

State University of New York, USA.

Masters in Computer and Information Science, 2003.

Nagpur University, India.

Bachelor of Engineering in Computer Technology, 2000.

Publications:

Sharma, R.; Evans, P.A.; Bhavsar, V. C. Regulatory link mapping between organisms, BMC Systems Biology June 20, 2011, Vol. 5, Supplement 1:S4, available at http://www.biomedcentral.com/1752-0509/5/S1/S4, 12 pages.

Sharma, R.; Evans, P.A.; Bhavsar, V. C. Mapping a Regulatory Network Between Organisms, Proceedings of the Fourth International Conference on Computational Systems Biology 2010, 204-218.

Sharma, R.; Evans, P.; Bhavsar, V. C. Transcription Factor mapping between Bacteria Genomes, International Journal of Functional Informatics and Personalised Medicine 2009, Vol. 2, 4, 424-441.

Sharma, R.; Evans, P.; Bhavsar, V. C. Mapping Transcription Factors from a Model

to a non-Model Organism, Proceedings of the International Joint Conferences on Bioinformatics, Systems Biology and Intelligent Computing 2009, 160-166.

Sharma, R.; Evans, P.; Bhavsar, V. C. Prediction of Regulatory Networks for Non-Model Organisms, Proceedings of the Fifth Annual Research Exposition 2008, 99-104 and poster presentation.

Sengupta, S.; Andriamanalimanana, B.; Card, S.W.; Das, K.; Sharma, R.; Gunasekaran, A. Data mining a high-speed bursty stream on a limited buffer in pseudo-stationary states, Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Proceedings of the Second IEEE International Workshop, 2004, 201-204.

Workshops and Presentations:

Participated in 'Workshop on Comparative Genomics: Special Session at the Smithsonian Institution', Washington DC 2011.

Successfully completed the requirements for Advanced Computational Proteomics: Structure, Imaging and Control certification from Lipari, International Summer School on BioInformatics and Computational Biology, Italy 2007.

Sharma, R. Bioinformatics Research for the Canadian Potato Genome Project, Oral Presentation, Rendez-Vous BioSciences N-B, Sackville, NB, Canada, Nov. 3-4, 2004.