Undergraduate Research Opportunity Program
(UROP) Project Report

Analysis of the role of cis-elements in the spatial regulation of the Arabidopsis salt stress response

By

Nguyen Duc Phong

Department of Computer Science

School of Computing

National University of Singapore

2009/10

Undergraduate Research Opportunity Program
(UROP) Project Report

Analysis of the role of cis-elements in the spatial regulation of the Arabidopsis salt stress response

By

Nguyen Duc Phong

Department of Computer Science

School of Computing

National University of Singapore

2009/10

Project No: U077353E
Advisor: Dr. José R. Dinneny and Prof. Wong Limsoon
Deliverables:
Report: 1 Volume
Source Code: 1 CD

Abstract

The computational search for cis-elements is a promising approach to enhance our understanding in gene regulation. However, no known computational method is satisfactory. Even worse, the performances of the methods vary to a great extent among different species. Our study focuses on Arabidopsis Thaliana, which has one of the smallest plant genomes. Firstly, we conduct a survey on certain biological assumptions which are crucial to the accuracy of any computational methods. The survey then lets us indirectly evaluate the available tools when used on Arabidopsis. Although we failed at creating a new computational method with higher accuracy, the study revealed the need of a system that utilizes available tools to conduct different analysis pipelines. We finalize this report with a description of the system we built and insights into future studies.

Subject Descriptors:
J.3 Biology and Genetics
I.5.5 Interactive systems

Keywords:
Microarray Data, Motif Analysis, Computer System Implementation

Implementation Software and Hardware:
Ubuntu 9, g++ 3.4, GNU Scientific Library, Perl 5 (CGI)

Acknowledgement

I would like to thank Dr. José R. Dinneny for introducing me to motif analysis. I thank him for all the discussions, guidance, and experience I received at his lab. Many ideas in this study are suggested by José.

I would like to thank Prof. Wong Limsoon for spending time to listen about my project and giving valuable advices.

Last but not least, my family have given me continuous support.

Without them, I would not have been able to complete this project.

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

1.1 Motivation

A computational method for finding cis-elements greatly benefits studies in biological systems.

The public expected a breakthrough in the understanding of biological systems and ultimately human body right after Human Genome Project finished. However, the book of life turned out to be more complicated than anticipated.

A lot of work has dedicated to genome annotation of different species. Many of the loci have their gene and function associated. However, when those genes are turned on (or off) is largely unknown. Such understanding is needed to understand a dynamic biological system, such as those in development process or those responding to a stimulus. In fact, every biological system is dynamic.

Gene regulation is mostly a result of the interaction of trans-elements and cis-elements. Trans-elements are transcription factors (TFs) that bind to regions of DNA-sequence to exert their regulation effect. Those regions are called cis-elements. It is worth noted here that other mechanisms do step in the way. RNA degradation, protein deactivation , siRNA... allows for more complicated patterns of expression. However, for many genes we know, fusing their regulation sequence with a reporter exon would be sufficient to replicate the expression pattern. This assures us of the validity of our aforementioned generalization.

From this realization, we see that cis-element finding is important to study gene regulation.

Knowing cis-elements allow us to effectively change the expression pattern of a gene in our way. Cis-elements also allow us to trace back to the transcription factors, and decode the gene regulation network. With microarray data, and hopefully transcription factor database, we will be able to visualize a complete picture of gene regulation. For example, among hundreds of genes involved in mitosis, we will know which one is up-regulated first, which one triggers which one, and if one gene is mutated, whether it will cause a cancerous cell.

1.2 Current Barriers

Experimental methods are costly. Computational methods are not accurate. Computational studies do not take species information into concern.

We differentiate motifs from cis-elements. A motif is a nucleotide pattern of binding sites of a transcription factor. A DNA sequence region following such a pattern may or may not be a functional binding site. Cis-elements are the binding sites that are functional. Introducing mutations to cis-elements will result in a change of expression.

Experimentally, the search for motifs is easier than cis-elements, thanks to recent proceedings in experimental molecular biology. However, low throughput methods are time consuming, while high throughput methods are expensive and require expert knowledge to set up. Researchers have added to our database of known cis-elements one by one, but those known cis-elements can only be used as a guide to find new cis-elements, not as a universal rule (Badis, Berger, & others, 2009).

Computationally, the search for motifs/cis-elements is not accurate enough. The performance of available systems varies among datasets (with $sensitivity \leq 0.222$ and $precision \leq 0.307$ in yeast(Tompa & others, 2005)). A user is puzzled when having to choose a particular method for use, since their results are different, and hard to combine. This is partly due to the degenerative nature of motifs. That is, a transcription factor can bind to more than one exact nucleotide stretch.

Most previous studies of computational methods are done on yeast, thanks to the database of genome, gene expression, and protein-DNA interactions available. However, such results might not hold when tested in another species. To an extreme, our human genome is notoriously complicated in regulation patterns.

1.3 Our approaches

Focus to one species. Survey the species characteristics. Build an integrated system of available tools.

Our study focuses on Arabidopsis genome with its expression data in salt-stressed environment. A plant can be grown in 2 different conditions: normal or salt stressed. For each type of cell, we would then measure the expression data of genes. That way, we obtained a spatial dataset of gene expression with and without salt-stressed.

Firstly, we study the genome and microarray data to test several biological assumptions. Background information in nucleotide percentage, gene spacing, clustering of expression data, positional bias... is explored. If one of these characteristics is unusual, this would directly leads to a more reliable way of detecting a subset of cis-elements. For example AlignACE looks for positional bias in cis-elements. If a genome does not follow this assumption, such an approach will lose its accuracy.

An important characteristic that we investigated was the complexity of protein-DNA interaction in Arabidopsis. There are different scenarios that can happen. To the simple extreme, each gene contains one or two cis-elements, and if another gene also contains the same element in its upstream sequence, that gene follows the same expression pattern. To the complicated extreme, the expression of a gene is rendered by competitive binding of transcription factors. The motifs that they recognize are highly degenerative. Also, whether a transcription factor can bind to a regulatory sequence depends on the chromatin structure and enhancers that loop back from far regions of the chromosome. This is the situation in human genes. If Arabidopsis follows such patterns, it will be extremely difficult to infer any logics of gene expression out of motif enrichment profile, due to the high number of false positive cases. We will later show with our study that fortunately this is not the case.

Testing different hypotheses allows us to have some ideas of the difficulty we are facing, and the reliability of each available computational method. In the other hand, while conducting the study we noticed a lack for an integrated system. Most available tools are built to serve a specific job. Some of them take in different kinds of input and return different kinds of output. To an extreme, FIRE (Elemento, Slonim, & Tavazoie, 2007) provides an end-to-end solution where users only have to input their microarray data, and few cis-elements will be returned in the end. Users who want to integrate different tools for their analysis will have to do some coding, and suffer hours of frustration because of the ill-defined input/output specification. We develop a system that integrates a few tools that can be combined to support different analysis pipelines. This will benefit biology researchers without background in computing, and overall, encourage the exploration of different analysis pipelines.

1.4 Problem Formulation

Aim: find regions in genes that determine expression pattern. If a mutation is introduced into such a region, we expect the expression pattern of the corresponding gene to be altered.

Input:

- Genomes of a species and its relative. In this study we focus at Arabidopsis Thaliana.
- Microarray data of the species

Output: pairs of (cis-element, gene)

This problem formulation deviates from many previous studies.

- Microarray data is utilized, instead of genome sequences only.
- The cis-elements found must be linked to certain genes. This is due to the fact that a cis-element in gene A may not works in gene B although they share the same transcription environment.
- We look for cis-elements instead of motifs only.

Let us explore this problem formulation further. All genes share the same intra-nuclear environment. Why are some genes expressed but not others? It is either because of the different cis-elements they possess, or the different localized environment they have.

Differential expression = different cis-element OR different trans-element OR BOTH

How cis-elements affect the differential expression depends on trans-elements. This relationship differs among families of species. For example, in mammals, reporter genes do not seem to work well. Simply put, their regulation sequences do not lie in the immediate neighborhoods, or their transcription factors create a loop in the DNA sequence to collect more transcription factors into the complex. We chose Arabidopsis Thaliana with the hope to avoid such complicated logics.

With the above equation in mind, we can reformulate the problem: given the differential expression, without prior knowledge of different trans-elements, can we find the different cis-elements which are causing the differential expression.

1.5 Report Organization

Firstly, we explore the previous approaches to the problem. These methods give us some idea of the difficulty of the problem, and which assumptions we should confirm if we are to use any of them. Then we move one to test several interesting assumptions with our dataset. This step is worthwhile because previous studies are mostly on yeast, not plant. Out of those assumptions, one was confirmed. While conducting the tests, we found it time consuming to create pipelines through different available tools. This leads us to creating a system combined of available tools, picking those that we believe to work judging from our prior tests.

Chapter 2

Related Work

2.1 Molecular Approaches

2.1.1 Introducing mutations

We make several $5'$ end deletion constructs of a promoter and create reporter genes with those. By observing the expression pattern in different constructs, we get an approximate of the cis-element location. Further site-directed mutagenesis is then required to confirm the location of cis-element

This approach does not work well with complicated situation of regulation, where cis-elements may not be upstream of a gene, or there are more than one cis-element. However, the underlying concept is clear: 1. Limit the range of search, and 2. Use site-directed mutagenesis to confirm the cis-element.

2.1.2 Electrophoretic mobility shift assay (EMSA)

Radioactive labeled DNA fragments are exposed to the cell extract. Fragments that are bounded to by transcription factors will have their mobility in gel reduced. This method allows analysis of in-vitro protein-DNA interaction.

2.1.3 ChIP-chip assay

Firstly, a living cell is freezed with formaldehyde, so that proteins that bind DNA will be fixed there. DNA strands would then be treated with sonication so that we obtain freely moving transcription factors with their binding DNA. Those DNA-TF complex belonging to the same transcription factor would the grouped together using immunoprecipitation. To know which DNA motifs one transcription factor binds to, we perform reverse cross-linking and then expose the motifs to DNA microarray.

This method confirms in-vivo protein-DNA interaction. However, the technique is costly. Moreover, biologically, the method is not perfect, since it only gives interactions, not functional interactions. Further test or computational prediction must be conducted to confirm cis-elements.

2.2 Computational Approaches

2.2.1 Group based

AlignACE(Hughes & others, 2000) and MEME (Bailey & Elkan, 2000) are two popular methods that look for significantly enriched motif in an input set. Both of them require the input set to be coregulated genes. AlignACE provides grouping of motifs, and can focus at positional bias. MEME is more suitable for protein motifs.

These two methods rely on the assumption that if a group of genes is coregulated, they should share the same motif. This assumption is biologically plausible, except for two points. First, the problem of finding co-regulated genes now belongs to biologists. Low quality grouping will affect the performance of the method. Second, they tend to see cis-elements acting individually, hence leaving group of functional motifs that act together statistically insignificant.

2.2.2 Conservation based

Conserved regions in the non-coding region of a gene are returned as cis-elements(Haberer & others, 2006). The assumption here is that any mutation in the expression pattern will be lethal. Another assumption is that we can find a species that is not too far from the species of examination, so that we can still find gene homologs; but it should not be too close, so that unimportant genes manage to deviate from its homolog.

Although this is a strong assumption, the underlying idea is valuable: not only comparing sequences intra-genome, we may also compare sequences inter-genome, as long as the difference in trans-elements and differential expression is taken into account. Meaning, in a relative, the gene of examination may not conserve its expression pattern; or the transcription factor and the cis-element have gone through co-evolution so that the cis-element is now changed and cannot be recognized when compared to our original sequence.

2.2.3 Preference based

Suppose we want to investigate an expression pattern (down-regulated in normal condition but up-regulated in hair root cells in salt-stressed condition...). We want to find motifs of the cis-elements that render this expression pattern. cERMIT (Georgiev & others, 2010) handles this kind of problem.

For each gene in the genome, cERMIT takes in an evidence value. This value represents our belief on whether the gene follows the type of regulation in investigation. cERMIT then returns a few motifs which are likely to cause such regulation.

If one wants to use cERMIT to replace AlignACE or MEME, a simple choice is to assign a score of 1 to selected genes, and 0 to unselected genes. This way, cERMIT will find motifs that are over-represented in selected genes (compared to unselected genes). However, when we consider if a gene is selected or not, we usually use some distance measure such as Eucledian or Pearson Correlation and a cut-off value. With cERMIT, we do not need to use a cut-off value to map a continous range into binary values (selected or not). A better scoring scheme is to channel the distance measure directly to cERMIT as evidence values.

In short, when compared to AlignACE and MEME, cERMIT saves the user from the burden of finding a cut-off value (to form groups).

2.2.4 Unified system

A few other methods try to combine algorithms to give a single final result. motifVoter (Wijaya & others, 2008) is a system from NUS that uses a voting system to combine results from single motif finding tools. Each motif finder returns a set of motifs. Motifs that are returned by different motif finders will be chosen. Chosen motifs are then aligned to regulatory sequences to extract possible binding sites.

The performance of such method depends on the performance of individual motif finder. In the case of motifVoter, it was observed that each motif finder performs well over different datasets. If one wants to use motifVoter on Araidopsis, one might want to ask if the individual motif finders work reasonably well on Arabidopsis.

Chapter 3

Our Study

This chapter describes our study of the cis-elements and gene expression in Arabidopsis. In each section, we also include the findings we have, and analyze the results obtained.

Firstly, we prepare input data for the study. Microarray data is then used for clustering analysis. Clustering analysis allows us to collect different patterns of spatial expression in salt-stress response. AlignACE, MEME and cERMIT then use the clustered genes as input to find enriched motifs. Up until here, this is a common work flow to find motifs.

We then investigate several assumptions. We test if similarly expressed clusters share similar motifs and vice versa, if clusters that share similar motifs would be similar in their expression. The last two assumptions tested are positional bias of motif occurrences and coexpression of neighboring genes. These two assumptions do not stand in line with the previous analysis, but are observations that can possibly increase the confidence of motif finding.

Lastly, we review the investigated assumptions, and analyze the implication our study has on motif finding in Arabidopsis.

3.1 Data preparation

3.1.1 Genome sequence data

3kb upstream of Arabidopsis gene contains most cis-elements

In Arabidopsis, the popularity of gene reporter method suggests that upstream sequences can be used to replicate the expression pattern. For most transcription factors (80%), 3kb upstream sequences would be enough (Lee & others, 2006). We collect the whole chromosome data together with gene annotation from TAIR9 (arabidopsis.org) and generate our own upstream sequences instead of directly use 3kb upstream sequences prepared by TAIR9. Firstly, we do not want to include exons in our regulatory sequences. Exons have to serve their structural role, hence unlikely to support any regulatory role. The binding of transcription factors to exon sites may affects transcription by RNA polymerase. Secondly, even if we use TAIR9 upstream sequences, we still need to parse the annotation information and the whole genome for later use (background nucleotide content, gene proximity...)

It is interesting to note that there are few degenerative characters in Arabidopsis genome. A degenerative character may stand for more than one of the nucleotides (A,C,T,G). Without noticing this, a motif finder will fail in rare cases, since they usually assume that they are working with sequences generated from the 4 basic letters.

Another side observation is that Chargaff's second rule holds in Arabidopsis genome. That is, %C = %G, %A = %T in each strand. This may be due to the fact that beside point mutation, duplication and inversion are the next common mutations.

3.1.2 Microarray data

.csv file with possible missing values from different file system

Parsing a large table is apparently simple. However, because the data usually goes through manual editing such as normalization, it is prone to errors. Such preprocessing and/or old systems may also introduce missing values.

To make the parser error-proof, we read in the file line by line. However, this introduces another problem, where different file system gives different end of line delimiters. At least 3 ways can be named (LF in Unix, CR+LF in Windows, CR in Mac OS). Our solution is to convert all CR into LF and then use LF as the newline delimiter.

3.1.3 Visualizing matches of motifs in promoters

We developed motifLocator as a tool to visualize location and orientation of matches.

There are two common ways to represent motifs: by position-specific scoring matrix (PSSM) or by degenerative characters. For example, if a motif is 5 characters long, its PSSM will be a 4x5 matrix p where p[c,i] = probability that the i-th character is c. Degenerative character is a different representation where N stands for [A,C,T,G], R stands for [A,G]...

We chose degenerative character since it is easier to stored, displayed and calculate. In fact, most of the available tools return motifs in degenerative form.

We then develop motifLocator as a web service that takes in gene IDs and a motif list. A Perl script would then convert motifs to their corresponding regular expression and search them against the specified promoters. Results are available in text format colored by HTML code, or image. We generated images using PostScript::Simple, a module in Perl. Different motifs results in different color of the arrows, and different orientation results in different direction of the arrows (Figure 3.1).

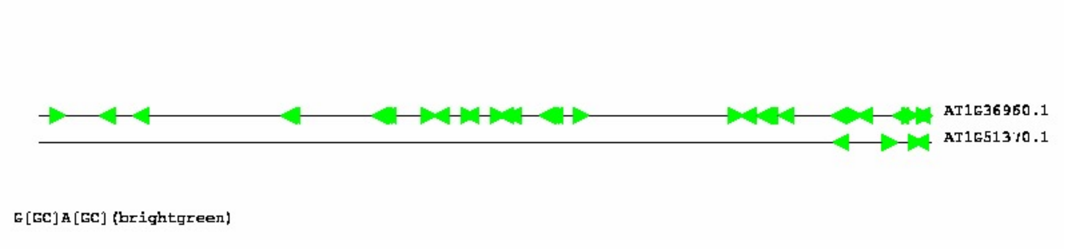

Figure 3.1: Location of motifs found with motifLocator

3.2 Clustering microarray data

To group genes into expression patterns and to find co-regulated genes

To start finding cis-elements, a researcher must gather a group of genes having the same cis-element (in AlignACE and MEME) or a measure of evidence for each gene to belong to such a group (in cERMIT). In either case, clustering would be a useful tool to explore the distribution gene expression. Only the clustering of gene expression is already informative to researchers. It gives them a hint on different modes of expression according to a set of conditions, and if the expression patterns are diverse or not.

Here we use Pearson Correlation as a measure of distance in clustering, and Affinity Propagation (Frey & Dueck, 2007) as a clustering algorithm. The choice of clustering algorithm will be discussed later in this report.

In around 21000 genes of Arabidopsis, we selected 4532 genes which are differentially expressed across different salt stressed conditions. Clustering these genes returns 49 clusters with the number of genes ranging from 25 to 218.

For each cluster, we would then run different motif finding algorithms to generate cis-elements. For AlignACE and MEME, the input is the genes in a cluster. For cERMIT, the evidence for each gene is the distance of the gene from the center (average) of a cluster. We combine the results from all clusters to give a single motif list that is predicted to cause differential expression in the input microarray dataset. At this stage of the study, we obtain 4 sets of motifs: known motifs in plants, motifs obtained by cERMIT, by AlignACE, and by MEME.

3.3 Motif enrichment profile for each cluster

Genes with similar expression pattern will be likely to possess the same cis-elements. Is this true? This step tests the possibility of inferring the logic of cis-elements. If such inferring can be done, the cis-elements found are more likely to be real.

This is a continued step of an analysis in Arabidopsis (Dinneny & others, 2008), where they found that similar clusters tend to share similar motifs. We try to observe the same pattern using our 4 sets of motifs.

To obtain a heatmap as in figure 3.2, we need to find the enrichment value of each motif over each cluster.

Given a motif M and a cluster, we want to find a p-value of enrichment of the motif in that cluster. Knowing those p-values, we can create a table where each value is used to fill a cell in the table (as in figure 3.2). We would then see if clusters similar in expression data would be similar in motif enrichment.

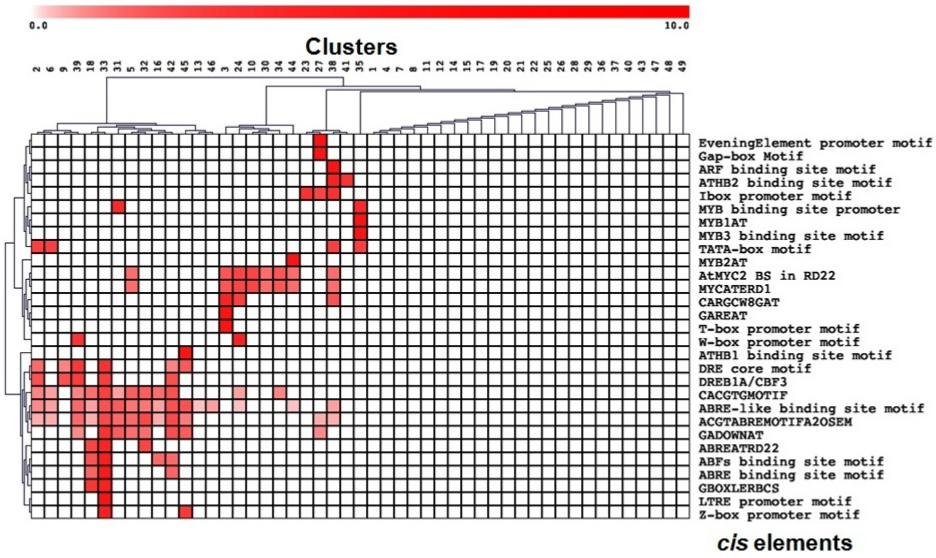

Figure 3.2: Clustering result from Dinneny's study

We initially used the count of occurrences instead of p-value, and found the result to be not meaningful. Then we have to use the p-value calculated from hypergeometric distribution as suggested in (O'Connor, Dyreson, & Wyrick, 2005). We use GSL library to calculate the p-value using this call:

$$gsl\_cdf\_hypergeometric\,Q(x, s1, BIGN - s1, s2)$$

where s1 is the number of genes that contain M, BIGN is the number of open reading frame, s2 is the number of genes in the cluster.

Let us further elaborate the use of hypergeometric distribution.

A cluster is seen as a sample of size s2 over our universal gene set of size BIGN. In BIGN genes, we have s1 genes containing M, and the rest (BIGN-s1) do not. We want to calculate the probability of having at least x genes containing M in such a sample. Although hypergeometric distribution is not guaranteed to be the true underlying distribution, it provides a mean of normalizing x over s2. If we only use x instead of the p-value, very large clusters would tend to have large values of enrichment, and this brings bias into our analysis.

While Dinneny's group only found the similarity between cluster 18 and 33 (figure 3.3), we now detect a new pair of clusters with similar cis-enrichment: (32,42) and (18,27) from the full set of known motifs. Those pairs of similarity are also confirmed by expression data.

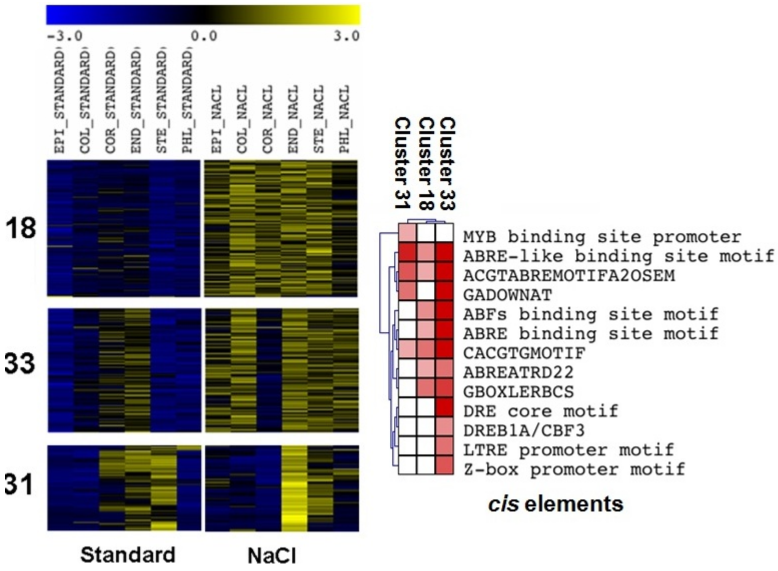

Figure 3.3: Cluster 18 and 33 have high similarity in expression pattern and motif-enrichment pattern

Beside individual comparison, we also made a quantitative overall comparison between cis-enrichment profile and gene expression profile. Each time we compare two clusters, there are two choices, either by taking Pearson correlation over expression profile, or Pearson correlation over enrichment profile. This leaves us 2 NxN distance matrices that capture the distance between pairs of clusters, generated by 2 methods. The two matrices are then used to calculate Pearson correlation. For this step of analysis, we use the list of known motifs to construct the enrichment profile. The Pearson correlation between this enrichment profile and the expression profile is 0.4. This value is not especially high, but it suggests that generally there is a relationship between cis-elements enrichment and expression profile.

3.4 Use motif enrichment to predict expression pattern

We create a predictor of the expression pattern of a cluster by returning another cluster that is closest in cis-enrichment profile to that cluster. If such a predictor works well, it suggests that the motif list we are using to profile the clusters captures the expression patterns well.

For a cluster, we find its approximation by iterating through the rest of the clusters and choose the one with the greatest Pearson correlation in cis-enrichment pattern. We measure the ”goodness” of this approximation by taking Pearson correlation between the original and the suggested cluster. For a list of cis-elements, we summarize the overall ”goodness” by taking the average of the correlation obtained for each cluster.

After getting the goodness measure for four different motif lists cERMIT, MEME, known, AlignACE, we need to know how big (small) those values are compared to control cases. We introduce two control cases. First, we generate random lists of motifs and measure their goodness. Second, we achieve an upperbound of this goodness measure regardless of motif list by using the best approximation for each cluster. Since the overall goodness is avaraged out of the individual goodness; and the individual goodness are Pearson correlation values, zero correlation will be counted as 0, while maximum correlation returns 1. We calculated and get the upperbound to be 0.81.

Following are the values we obtained for different motif lists, so call predictive power:

- cERMIT: 0.442392
- Known: 0.305368
- MEME: 0.251597
- Small random motifs: 0.291948
- Medium random motifs: 0.257711
- Long random motifs: 0.0925249

(Random motif lists are generated with size = 500, length = 5, 7, 12)

These are measured against the upperbound ”goodness” of 0.810605 where we approximate each cluster by its closest cluster in expression data.

This replicates the results from Dinneny’s group where known motifs perform relatively well. cERMIT performs best, partly because it is created from microarray data, or possibly because it is indeed a good method. It is also interesting that control cases perform relatively well when the length of motif is reasonable.

As cERMIT performs well in this test, we will later use it in our web service, GeneTable.

3.5 Motif position analysis

For known motifs, we study to see if positional bias is common in Arabidopsis

Experience from the biology side suggests that the majority of positional bias tends to happen closer to the promoter. Here we survey all the occurrences of known motifs in Arabidopsis and see if this is the case.

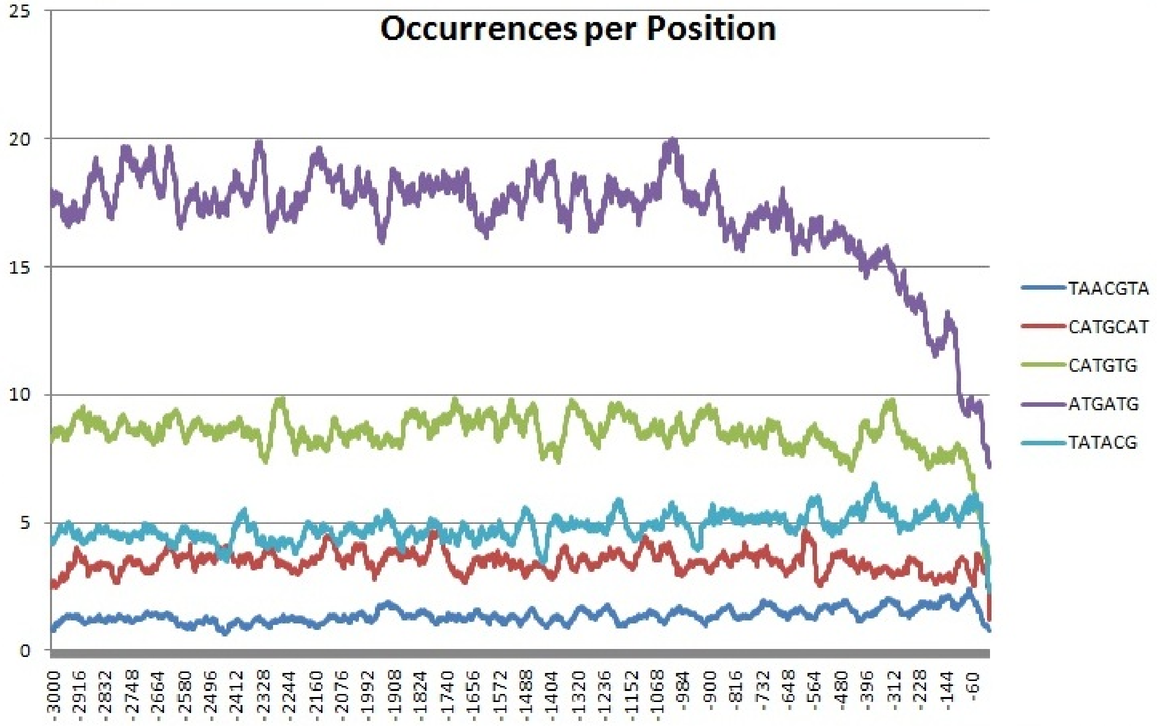

Figure 3.4: Occurences of motifs in different positions. The first three motifs are known motifs in plants, while the last two are randomly generated motifs. y axis is the number of occurences. x axis is the position relative to transcription initiation site -3000..-1

We expected that known motifs will tend to appear more in the region closer to the transcription start site, but there are no clear difference between known and random motifs in this aspect (figure 3.4).

This result is not satisfactory partly due to the fact that not all considered occurrences are functional, hence not restricted to positional bias. Analysis with FIRE suggests that even if a motif follows positional bias, it only does so in a group of genes where it is likely to be functional.

3.6 Correlation of expression among neighboring genes

We test if this is a common phenomenon. If yes, our motif finding strategy should focus at different regions of the genome.

In prokaryote, operon provides a mean of grouping neighboring genes under the same expression. As soon as the shared promoter is bound to, all genes in the operon will be transcribed. We are interested in looking for such phenomenon in eukaryotes, where neighboring genes may follow the same expression pattern due to shared localized environment such as chromatin structures... If this is common in Arabidopsis, we would want to merge the regulatory sequences of similarly expressed neighboring genes and treat the resulted sequence as the representative.

For most genes, there is one neighboring gene to the 3' end, and another to the 5' end of the DNA strand. We find the mean correlation of all such pairs of neighboring genes. The mean correlation is 0.05, suggesting that this phenomenon is not universal. Out of 20000 pairs of neighboring genes examined, about 400 pairs have Pearson correlation greater than 0.9. Individual studies may look into those pairs to examine whether that correlation is by chance or by some biological determinant.

3.7 Analysis

Our study confirms an important property of Arabidopsis genome: the logic of cis-elements are relatively simple. If two genes share some predicted motifs, they would be likely to behave similarly in expression data. This is a good news, because old approaches that worked in yeast would still be portable to Arabidopsis. AlignACE and cERMIT would be two candidates that stand out.

We also searched for similar expression among neighboring genes, and conclude that this is not common in Arabidopsis. Hence, the property does not help much in a genome scale study of finding cis-elements. However, in individual studies, there may still be chance that neighboring genes do follow the same expression pattern. It would then require more analysis to see the reason behind, whether it is because of gene duplication or localized environment.

The last property of Arabidopsis genome that we tested was positional bias. Although we did not observe any pattern, we cannot conclude that the phenomenon does not exist. This is due to a large amount of noise introduced into our calculation. The known motifs are not definitive description of cis-elements, but more as a suggestive description. Hence, many occurrences of the motifs are false positive, and dilute the positional bias signal. We think that this approach is worth examining after one has obtained a better description for cis-elements. That is, if we can obtain all functional binding site of a transcription factor, positional bias may be checked to increase confidence.

Chapter 4

GeneTable

4.1 Introduction

While conducting our analysis, we came up with several pipelines that deviate from the common workflow - clustering, find enrichment, find binding sites. Below are some interesting pipelines.

- Take a conserved region C in a gene G, find those that contain C, and select for genes with high correlation to G in salt expression. /*Trying to confirm whether that conserved region is related to salt-stress response.*/
- Use cERMIT to find enriched cis elements in a cluster, select for genes with those enriched-cis, and repeat the process over the new restricted cluster. /*Collect genes that seem to be co-regulated, and search for cis-elements. Those cis-elements are then used to refine the set of co-regulated genes.*/
- Find transcription factors that are differentially expressed in salt stress conditions.

The code that supports these pipelines are simple, but boring to write and requires background in computing. At times we were frustrated by the difficulty of getting the input right for available tools. This suggests the need of an integrated system.

GeneTable helps biologists use different combination of bioinformatics tools to mine for cis-elements of their interest. The outcome of each tool is human friendly, and can be used for further analysis.

To use the collection, one does not have to install anything. Setting up a computer system alone is a burden for many biologists. We provide the service through a website, so that the requirement of hardware from the client side is minimal.

Beside the sequence data that is readily available in the host, users will need to input microarray data and/or their motif lists. Given such data, users can proceed with different pipelines of analysis, as mentioned above.

Here we will discuss the architecture of the system as a way of introducing features, and explain the reason why we included those features. Finally, we discuss on what more can be done to improve GeneTable.

4.2 Architecture

To be able to support the aforementioned use cases, GeneTable keeps track of a shared pool of numeric data. The database can be visualized as a large table, with each row being a gene, and each column being numerical. A column can contain Boolean values to represent a set, real values to represent preferences to a model of interest, integers to indicate clustering, or possibly p-values to represent cis-element enrichment. Since we are collecting data from experimental outcome, missing values should be catered for.

Table 4.1: Example of a database

GeneID	ClusteringID	Gene Distance	Enrichment value	Membership	...
AT1G36960	1	0.34	3	0	...
AT3G16360	2	0.12	4	0	...
AT5G23150	3	-0.23	5	1	...
...	...	...	...	...	

When a session initially starts, the table is empty. The user can input more columns. After data has been input, users can start their analysis using different tools in the system.

4.2.1 Clustering

Clustering is an important step to gain knowledge of expression pattern. Although a series of clustering algorithms is available, not all of them fit well into the role of clustering microarray data. Density-based clustering is not used for microarray data, since clusters are expected to be of spherical shape. Not only a clustering method has to work well, it also has to be able to explain its decision well. This brings more confident to the biologists. With respect to this criteria, hierarchical clustering is among the more widely used methods, especially agglomerative hierarchical clustering. In the class of partitioning methods, k-means and derivatives are quite popular as well.

In our collection, we use Affinity Propagation, which is a recently popular method. One good feature of Affinity Propagation is that a user does not have to decide the number of clusters to be made.

4.2.2 Find distance

Given an expression pattern, we may be interested in finding genes that follow the expression. We may also be interested in ranking the genes according to that pattern, and search for cis-elements that correlate with the expression pattern.

The distance might be Pearson correlation, or Euclidean distance... The distance measurement for all genes are stored in the shared database, available for further use by other tools.

4.2.3 Find enrichment

Biologists might want to validate the results returned by different motif finders. To do this, they have to obtain the enrichment value for each gene over the motif list to be tested.

We count the number of occurrences of the set of motifs in each gene, and use the counting to select for significantly enriched genes. If the number of enriched genes in the original cluster is no greater than the background enrichment, we can conclude that the motif list is not good.

4.2.4 Find correlated motifs

Beside tools that find motif-enrichment over a genomic background like AlignACE and MEME, cERMIT provides a new way of utilizing inputs. For each sequence it takes in one more preference value, and the output motifs will be the ones that are likely to account for the preference gradient.

4.3 Possible extension

- The learning curve exists when using the system, since it is not an end-to-end solution. To solve this, we intend to build a layer on top that utilizes underlying tools. Biologists may either use that shortcut, or start with their new design of analysis.
- For the ease of use, more prepared data should be collected. We may want to prepare a list of transcription factors, or a list of histone modification proteins.
- We may also allow more choices for the sequence data. Conserved regions in Arabidopsis would be a good candidate.

Chapter 5

Conclusion

Cis-element finding is a hard problem with no satisfactory solution yet. Beside the current trend of conducting genomic scale ChIP-chip experiments tailored by computational post-processing, our study suggests that there is still hope for a good computational solution utilizing widely available genome data and microarray data. In the meantime, biologist may try using available tools in different ways. GeneTable is built to ease the analysis in common workflow, and facilitate the thinking of new workflows.

This problem is hard to biologists because it lies deeply in the field of statistics and data mining. It is hard to computer scientists because the amount of noise in the data, and the different assumptions with different strength lie in the field of molecular biology.

If a future solution manages to solve this problem, we expect that it utilizes more biological assumptions. It may use different types of data, such as conserved regions, orthologous genes, transcription factor expression patterns... With the current biological assumptions used, data mining methods seem to have reached their limit. In Arabidopsis, comparative genomics with Arabidopsis lyrata and Capsella rubella would be interesting. Another situation is that cheaper and more accurate experiments come out and remove the burden from computational approaches.

References

Badis, G., Berger, M., et al. (2009). Diversity and complexity in dna recognition by transcription factors. , June, 2009, 1720–3.

Bailey, T. L., & Elkan, C. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. , March, 2000, 28–36.

Dinneny, J., et al. (2008). Cell identity mediates the response of arabidopsis roots to abiotic stress. , May, 2008, 942–5.

Elemento, O., Slonim, N., & Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. , October, 2007, 337–50.

Frey, B., & Dueck, D. (2007). Clustering by passing messages between data points. , February, 2007, 972–6.

Georgiev, S., et al. (2010). Evidence-ranked motif identification. , February, 2010.

Haberer, G., et al. (2006). Large-scale cis-element detection by analysis of correlated expression and sequence conservation between arabidopsis and brassica oleracea. , December, 2006, 1589–602.

Hughes, J., et al. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. , March, 2000, 337–50.

Lee, J., et al. (2006). Transcriptional and posttranscriptional regulation of transcription factor expression in arabidopsis roots. , April, 2006, 6055–60.

O'Connor, T., Dyreson, C., & Wyrick, J. (2005). Athena: a resource for rapid visualization and systematic analysis of arabidopsis promoter sequences. , December, 2005, 4411–3.

Tompa, M., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. , January, 2005, 137–44.

Wijaya, E., et al. (2008). Motifvoter: a novel ensemble method for fine-grained integration of generic motif finders. , October, 2008, 2288–95.

Appendix A

Glossary

Gene: unit of our genetic information, which will be translated to give a protein

Expression: whether one gene would be transcribed to yield mRNA, and whether its mRNA would be translated to give a protein.

Promoter: DNA sequence that lies in front of a gene

Transcription Factor (TF): proteins that bind to regions around genes to determine when one gene should be expressed

in vitro: in test-tubes

in vivo: in the living organism

cis-element: DNA sequences that are recognized by transcription factor as a docking object

motif: cis-elements that are relatively short and consists of concensus nucleotides

co-expressed: being expressed together

co-regulated: being bound to and control when to express by the same set of TF

orthologous: similar genes in different species

microarray: set of cells that will highlight when they are bound to. They can be recognized by proteins/mRNAs...