Hypothesis-Driven Specialization-based Analysis of Gene Expression Association Rules

by

Dharmesh Thakkar

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

May 2007

APPROVED:

Professor Carolina Ruiz, Thesis Advisor

Professor Elizabeth Ryder, Co-Advisor

Professor Murali Mani, Thesis Reader

Professor Michael Gennert, Head of Department

Abstract

During the development of many diseases such as cancer and diabetes, the pattern of gene expression within certain cells changes. A vital part of understanding these diseases will come from understanding the factors governing gene expression. This thesis work focused on mining association rules in the context of gene expression. We designed and developed a tool that enables domain experts to interactively analyze association rules that describe relationships in genetic data. Association rules in their native form deal with sets of items and associations among them. But domain experts hypothesize that additional factors like relative ordering and spacing of these items are important aspects governing gene expression.

We proposed hypothesis-based specializations of association rules to identify biologically significant relationships. Our approach also alleviates the limitations inherent in the conventional association rule mining that uses a support-confidence framework by providing filtering and reordering of association rules according to other measures of interestingness in addition to support and confidence. Our tool supports visualization of genetic data in the context of a rule, which facilitates rule analysis and rule specialization. The improvement in different measures of interestingness (e.g., confidence, lift, and p-value) enabled by our approach is used to evaluate the significance of the specialized rules.

Acknowledgments

I would like to express my gratitude to my advisor, Prof. Carolina Ruiz for her guidance and support throughout the course of this thesis which has helped me grow and be better prepared for the professional challenges I would encounter. But for her constant encouragement, patience and availability it would have been impossible for me to go on. I would like to thank my co-advisor Prof. Liz Ryder who did not get bugged by all the biology questions I had but instead went to great lengths to ensure that even a computer science student could understand them. Their (both my advisor's and co-advisor's) enthusiasm in the 8 a.m. meetings forced upon them by my schedule always kept me motivated. I thank Prof. Murali Mani for agreeing to be my reader and also for reading the thesis at such short notice. Last but not the least, I would like to thank my family, that is, my folks back home in India and my wife Vandna. Their confidence in me inspires me to scale new heights and they form a strong support system whenever I need it.

Contents

List of Figures

Chapter 1

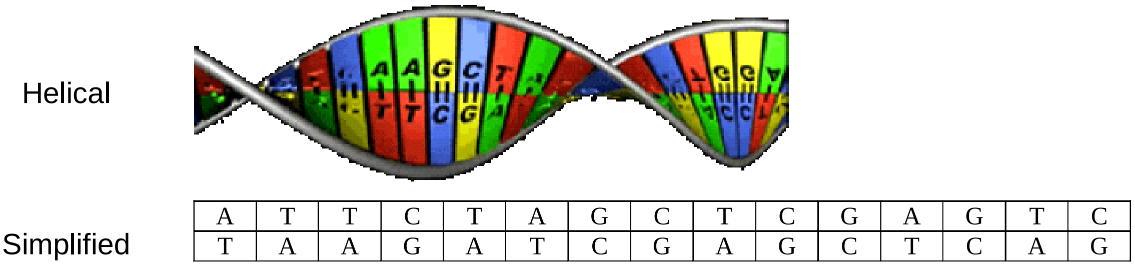
Introduction

During the development of many diseases such as cancer and diabetes, the pattern of gene expression within certain cells changes. A vital part of understanding these diseases will come from understanding the factors governing gene expression. This thesis work focused on association rules mined in the context of gene expression. We designed and developed a tool that enables domain experts to interactively analyze association rules that describe relationships in genetic data. Association rules in their native form deal with sets of items and associations among them. But domain experts hypothesize that additional factors like relative ordering and spacing of these items are important aspects governing gene expression.

We proposed hypothesis-based specializations of association rules to identify biologically significant relationships. Our approach also alleviates the limitations inherent in the conventional association rule mining that uses a support-confidence framework by providing filtering and reordering of association rules according to other measures of interestingness in addition to support and confidence. Our tool supports visualization of genetic data in the context of a rule, which facilitates rule analysis and rule specialization. The improvement in different measures of interestingness (e.g., confidence, lift, and p-value) enabled by our approach is used to evaluate the significance of the specialized rules.

1.1 Biological Motivation

One of the central questions in modern biology today is what controls gene expression. Deoxyribonucleic acid (DNA) is a complex molecule which encodes genetic information unique to an organism. Every cell in an organism contains the same set of instructions encoded in the DNA, and this information is arranged into regions called genes. Still, a brain cell is very different from a heart cell and performs an



A	T	T	C	T	A	G	C	T	C	G	A	G	T	C
T	A	A	G	A	T	C	G	A	G	C	T	C	A	G

Figure 1.1: Structure of a DNA molecule and corresponding linear simplification.

Gene expression is the process by which the information encoded in a gene is copied (transcribed) into RNA which may further be translated into a protein. The promoter region of a gene is the portion of the DNA sequence upstream of the gene. The process of making an RNA copy of the relevant portion of the genes DNA is called transcription, and the point where the promoter region ends and the gene begins is called the start of transcription (SoT). RNA is chemically slightly different from DNA, but contains the relevant information for a particular gene, and it can move to a part of the cell where that information can be translated into protein.

Proteins are the basic chemicals that make up the structure of cells and direct their activities. Each protein has a specific function that is determined by the blueprint stored in DNA, specifically the gene.

Deeper understanding of gene expression would not only help in the functional classification of genes but would also be instrumental for developing cures to diseases where the gene expression patterns within a cell change. With technological advances more genetic data is being collected today than ever before. This is creating an increasing gap between the rate of data collection and the rate of data analysis.

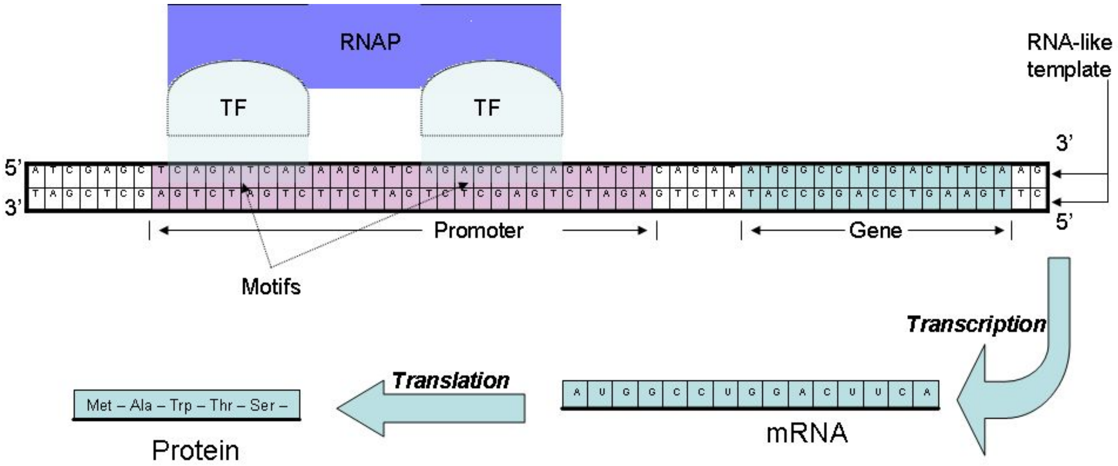

RNAP
TF
TF
RNA-like template
5'
3'
A T C G A G C T C A G A T C A G A A G A T C A G A G C T C A G A T C T C A G A T A T G G C C T G G A C T T C A A G
T A G C T C G A G T C T A G T C T T C T A G T C T C G A G T C T A G A G T C T A T A C C G G A C C T G A A G T T C
3'
5'
Promoter
Gene
Motifs
Transcription
Translation
Met – Ala – Trp – Thr – Ser –
Protein
A U G G C C U G G A C U U C A
mRNA

Figure 1.2: Central dogma of biology: DNA→RNA→Protein.
During transcription, RNA polymerase (RNAP) copies DNA to RNA using the template strand. The RNA transcribed is identical to the RNA-like strand except that U's are substituted for T's. A transcription protein (TP) binds to either enhance or repress transcription of a gene by assisting or blocking RNAP binding. During translation, the RNA encodes proteins.

Domain experts attribute the selective activation of genes in any cell to:

1. The presence of a particular set of proteins controlling transcription called Transcription Proteins (TP) in a given cell.

2. The presence of certain repeated sequences of DNA (motifs) in the promoter

region, which is the section of the DNA sequence upstream of the gene. These motifs are bound by transcription proteins during transcription.

Throughout this thesis we focus on identifying control patterns, defined in terms of the presence of a motif, to model gene expression. Motifs control gene expression, as they are putative binding sites which bind the transcriptional proteins. Expression of each gene may require the presence of a combination of motifs. Domain experts also hypothesize that inter-motif distances and order of occurrence of motifs in the promoter region are additional factors that control this regulatory interplay of motifs and thus also influence gene expression.

1.2 Computational Motivation

Computational processes to identify and shortlist *interesting* relationships (associations) between motifs and gene expression, which could then be analyzed biologically in detail, are gaining importance, as these would help reduce the growing gap between data collection and analysis. Associations that are statistically significant may yield biologically valid connections between associated variables. Association rule mining, introduced in [AIS93], provides a useful mechanism for discovering relationships between variables in a dataset. Relationships are represented in an if-then format with statistical measures to indicate the strength of the relationship. Association rules are of the form:

$$Antecedent \Rightarrow Consequent[Support = S, Confidence = C]\text{such that } S, C \in [0, 1]$$

A rule of this form in a market basket analysis of customer purchases could be:

$$Bread, Butter \Rightarrow Eggs[Support = 0.45, Confidence = 0.80] \tag{1.1}$$

The **Support** of the rule is the probability of finding both the antecedent (the *if* part) and consequent (the *then* part) of the rule in a data instance (i.e., a row in the dataset). For instance, in rule 1.1, the Support of the rule signifies that 45% of the customers in the database bought all three items, that is, Bread, Butter, and Eggs.

The **Confidence** of the rule is the conditional probability of the consequent given the antecedent. Again, in the context of sample rule 1.1, the Confidence of the rule signifies that 80% of the customers who bought Bread and Butter bought Eggs as well.

Previous work at WPI ([MPPT01], [BLT02], [BFG$^+$03], [Ice03], [IRR03]) has provided the foundation for gene expression association rule mining. [BLT02] focused on creating a computational biology tool, CAGE, that built association rule based models to predict gene expression. [BFG$^+$03] implemented a more elaborate methodology for discovering potential motifs and concentrated on improving the predictive accuracy of the models. However, none of these systems provide an interface that enhances the ability of a domain expert to analyze the resulting rules. This work focuses on facilitating visualization of the mined rules with respect to location of motifs on promoter regions of interest, since this is essential to interpretation of the rules by domain experts.

Besides setting the base methodology, [MPPT01] also attempted to address the biological hypothesis - “Does distance between motifs matter?”. [Ice03], [IRR03] extended the idea and focused on incorporating distance information in the mining process itself. But they do not provide a way to verify alternative biological hypotheses. Consider the following scenario, which brings forth one of the shortcomings of incorporating more constraints in the mining process, and suggests why it might not be the best approach in an environment where we intend to perform exploratory analysis. A batch of gene sequences was mined for gene expression related association rules with a support and confidence of 0.5 or greater. If the user now wants to find out rules with support and confidence of 0.4 or greater, the only way to achieve this is to mine again, which is a very time consuming process. From an exploratory analysis perspective, it is imperative to facilitate visualization of data in the context of a rule in real-time. So one could mine for association rules with lowest bounds of support and confidence. Then this work lets the user visualize the data which may allow the user to quickly perceive a pattern in the data, suggesting a specialization that would greatly increase the confidence (or other measures of interestingness). This work also computes the different measures of interestingness of the specialization over the training data for each such identified specialization and thereby provides an instantaneous estimate on the statistical strength of the rule. This work provides the ability to test a few biological hypotheses, as such a provision is instrumental in the development of an effective data-mining algorithm for gene expression.

Moreover all the above mentioned efforts were based on the basic support-confidence framework of rule generation. Several other measures of interestingness have been proposed to measure the relative importance of association rules (e.g., gain [Mor98], chi-squared value [Alv03], and lift [BMS97]). But there is no one good measure that is applicable to all domains. This work provides a way to analyze important rules according to several measures and thereby observe the applicability of each of these measures to this problem domain.

[PR05] introduced an algorithm to mine expressive positional relationships from complex sequential data. We adapted the data to define motifs as events and then utilized this algorithmic approach to mine for statistically significant association rules with positional relationships. This work also provides a facility to visualize genetic data in the context of such positional specializations.

1.3 Problem Statement

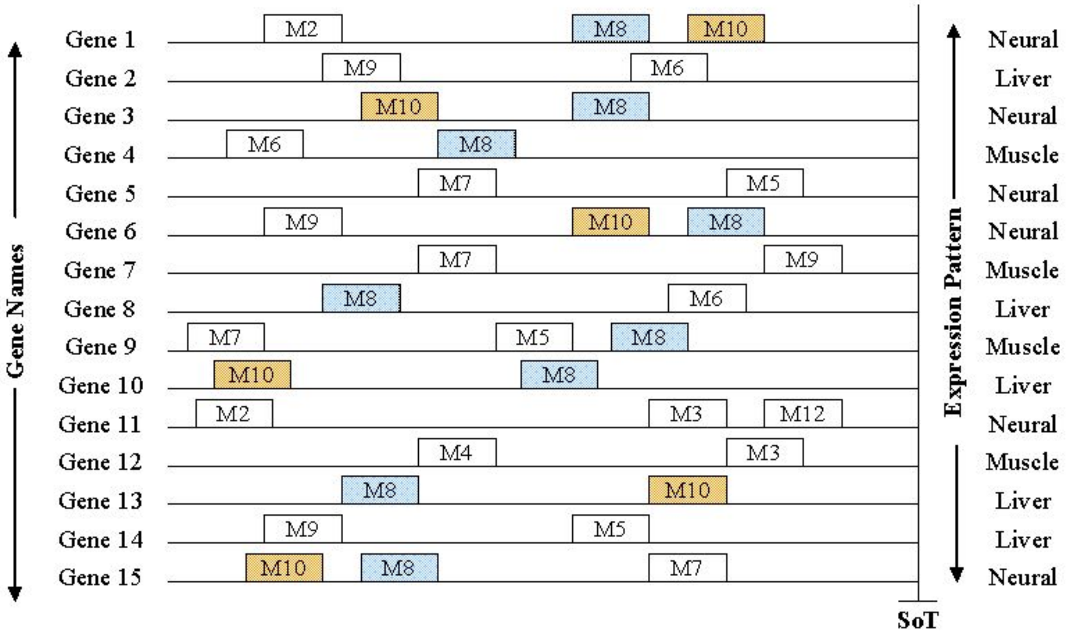


Figure 1.3: Hypothetical dataset of 15 genes and 10 motifs.

We designed and developed a tool to facilitate the post-mining analysis of rules for verifying biological hypotheses and also to aid the visualization of the rules generated. The tool has been integrated with the WPI-Weka system, a local version of the open source data mining tool. Mining of a hypothetical gene expression dataset as shown in Figure 1.3 would produce association rules of the form:

$$\text{M8 \&\& M10} \;\Rightarrow\; \text{neural } [Support = 0.27, Confidence = 0.67] \tag{1.2}$$

which states that the presence of Motif M8 and Motif M10 in the promoter region of a gene implies that there is a 67% likelihood that the gene is expressed in cells of type neural. Also, 27% of our data instances contain M8, M10 and are expressed in cells of type neural. The support and confidence statistics are computed from the hypothetical data in Figure 1.3.

This work will provide the necessary functionality for a domain expert to interactively analyze genetic data in the context of the following biological hypotheses:

1. *Inter-motif distance is important in characterizing gene expression.* DNA consists of linearly linked nucleotides. Subsequences of the DNA sequence, like a gene or a promoter region, can be represented for example as ATTTCCCGGT. By representing DNA as a sequence, the number of bases could be used to imply a notion of distance between motifs. In the sample sequence **ATT**CGGGGGG**TAT** we could say that the Motif ATT is at a distance of 7 bases from the Motif TAT. Now consider the following specialized form of rule 1.3:

   $$\text{M8 (0-250) M10} \;\Rightarrow\; \text{neural } [Support = 0.20, Confidence = 1.0] \tag{1.3}$$

   That is, the presence of Motif M8 within a distance of 250 bases from Motif M10 implies that the gene is likely to be expressed in cells of type neural. Again the support and confidence statistics are computed from the hypothetical data in Figure 1.3. Notice the change in the statistical measures with respect to (1.2). Every time a rule is specialized, it may be applicable to fewer data-instances (genes) in the dataset. This explains the reduction in the Support value. The increase in Confidence is indicative of the classification accuracy. A value of 1.0 signifies that of the data instances (genes) to which the rule could be applied (i.e., those that match the antecedent of the rule), the expression type predicted by the consequent of the rule is correct(i.e., it matches the known expression type of the data instance.

2. *Distance of a motif from the start of transcription (SoT) affects gene expression.*

   Since many of the known putative regulatory elements are found close to the Start of Transcription (SoT), we want to find out if the distance of the occurrence of a motif to the start of transcription has any effect on gene expression. This work facilitates visualization of gene sequences, along with the motifs involved and the start of transcription, in the context of the generic rule of the form (1.2). This enables the user to form and visualize specializations of the form:

   $$\text{M8 (0-500) SoT \&\& M10} \;\Rightarrow\; \text{neural } [Supp = 0.13, Conf = 1.0] \qquad (1.4)$$

   where the presence of Motif M8 within a distance of 500 bases or less from the Start of Transcription (SoT) and presence of motif M10 anywhere in the promoter region imply that the gene is likely to be expressed in cells of type neural. Again the support and confidence statistics are computed from the hypothetical data in Figure 1.3. Observe again the alteration in support and confidence as compared to (1.2).

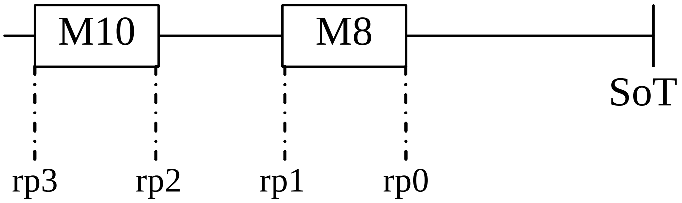


Figure 1.4: Syntax description of order of occurrence based specialized rules.

3. *The order of occurrence of motifs affects gene expression.* Knowledge of gene expression regulation is not complete. Domain experts hypothesize that the order of occurrence of motifs in the regulatory regions could also affect gene expression. This work provides the facility to visualize the gene sequences in the context of specializations of the following form which were either mined directly using the approach in [PR05] or visually observed and enhanced as a part of the exploratory analysis.

   $$\text{M8 (rp0-rp1) M10 (rp2-rp3)} \;\Rightarrow\; \text{neural } [Supp = 0.20, Conf = 0.75] \tag{1.5}$$

   This rule states that the presence of an occurrence of Motif M8 in the promoter region of a gene in between an occurrence of Motif M10 and the Start of Transcription (SoT) implies that the gene is likely to be expressed in cells of type neural. *rp* in the rule above refers to the relative position of the motif with respect to the Start of Transcription (Figure 1.4). Each motif has a begin point and an end point and hence requires both a begin index and an end index to capture the relative positioning of the motifs on the gene sequence. The support and confidence statistics are computed from the hypothetical data in Figure 1.3. Here again we notice that the specialization process produces an improvement in the confidence of the rule.

For each of the form of specializations discussed above, different measures of interestingness are computed to estimate if the hypothesis-based specialized rule better explains the underlying regulatory mechanism as compared to its generic counterpart.

1.4 Summary of the contributions of this work

This work focused on the development of a computational tool for exploratory specialization of association rules predicting gene expression in the context of the above mentioned biological hypotheses. Also it provides for filtering/sorting association rules based on measures of interestingness beyond the conventional measures of confidence and support.

The main contributions of this work are a framework and a tool that can:

1. Facilitate exploratory rule analysis (specialization) by providing a user-interface for rule visualization in the context of different biological hypotheses.

2. Select and present rules according to different interestingness metrics.

3. Test hypotheses relating the order of motifs, inter-motif distance and the distance of motifs from the start of transcription to gene expression control.

4. Provide updated genetic data, an important resource for any further research in the domain.

5. Be integrated seamlessly with the WPI Weka system. That is, gene expression association rules mined with the WPI-Weka system can be visualized and specialized using this work. Also, a model consisting of interesting rules and their specializations could be used by the WPI Weka to measure the model's classification accuracy over novel data.

Chapter 2

Background

2.1 Gene Expression

Simply put, a gene is a segment of DNA and is the physical and functional unit of heredity information. Gene expression is the process of using the information encoded in a gene to manufacture protein in the cell. Each cell of an organism, for instance neural or muscle, has the same DNA, but still performs completely different functions. This phenomenon is often referred to as the "The central dogma of biology" and is described in more detail in Section 1.1.

With technical advances in all fields, more and more data is being generated today than ever before, and the field of gene expression is no exception. Thus there is an imperative need for methods to analyze data at an equivalent rate.

The focus of this work was to design and develop a tool that helps a biologist to visualize genetic data to explore interesting regulatory patterns governing gene expression. As discussed in Section 1.1, motifs (repeating DNA segments) control gene expression, as they are putative binding sites which bind the transcriptional proteins. Hence, it is central to our work to determine groups of motifs that are likely to be real binding sites collectively controlling expression, which in turn is contingent upon the quality of the data used to discover (elicit) motifs. This was one of the prime reasons why we decided to collect data from scratch. We compiled a database using the Wormbase database [Wor] and RSA database [RSA] that consists of 164 genes, from nine different cell types with at least 30 genes known to be expressed in each cell type. Furthermore, we conducted a pilot experiment that elicited motifs from these sequences using both MEME [MEM] and GIBBS [JAC95] to observe the cost vs quality analysis of both algorithms. Since there was no perceived benefit in terms of quality of motifs we opted for lower cost (less time-consuming) elicitation algorithm MotifSampler (GIBBS). The data collection as well as the motif elicitation process is covered in detail in Section 3.

2.2 Association rules

Association rules were introduced in [AIS93]. Consider a database (D) in relational format where each record (data instance) consists of n boolean attributes. Association rules model relationships of the form: presence of a set A $(A \subset D)$ implies the presence of the another set C $(C \subset D)$ where the two sets A and C are disjoint (i.e., $A \bigcap C = \emptyset$). The most common statistical measures to estimate the strength of the rule are support and confidence and these have been covered in depth in Section 1.2.

Apriori is the traditional algorithm used to mine association rules [AS94]. Even a small dataset could yield a large number of rules and so the support-confidence framework is utilized to identify relationships which are statistically interesting. It follows an inductive approach to find *itemsets* (i.e., sets of items belonging to the data) that occur together *frequently*, within a dataset. An itemset is frequent if the support of the itemset is greater than the minimum support, a threshold provided as an input parameter. The apriori principle basically states that an itemset is frequent only if all its subsets are frequent and this principle is utilized in the inductive approach to search for frequent itemset.

Even with the popularity of the support-confidence framework in the association rule mining literature there is no one good measure that is applicable to all domains. Several other measures of interestingness have been proposed to measure the relative importance of association rules. The lift value of an association rule is another measure to try to quantify the interestingness of the rule. It is defined as the ratio of the confidence of the rule and the support of the consequent of the rule. The p-value of the rule is the probability that the correlation between the antecedent and the consequent is due to chance by using the chi-square test.

2.3 Prior work at WPI

As discussed in Section 1.2 previous work at WPI has provided the foundation for gene expression association rule mining. AprioriSetsAndSequences [PR05] introduced an algorithm to mine expressive temporal relationships from complex sequential data in addition to the regular association rule mining and provides enhanced pruning mechanism, which is of significant value especially when mining large sequence databases. It takes preprocessed sequences, that is sequences of events (e.g., the price of a company's stock recorded every hour or the repeating patterns in a gene sequence) as input with a minimum support and confidence threshold and produces association rules with temporal relationships between events. [MPPT01] attempted to address the biological hypothesis - "Inter-motif distance influences gene expression". [Ice03], [IRR03] extended the idea and focused on incorporating distance information in the mining process itself. Integrating this work with the WPI-Weka system has been a shared goal with [Rudss], a work in progress.

Chapter 3

Data: Motif Elicitation and Sequence Annotation

The data in the domain of interest (genetics) is sequential in nature. In this section we present the process followed in order to transform sequential genetic data to a relational format. That is, a format similar to the conventional database format that enables the use of existing data-mining algorithms to identify patterns of interest from a gene regulation perspective.

3.1 Data Collection

It cannot be stressed enough that no matter how good a mining algorithm may be, the information retrieved/discovered is only as good as the data. Adhering to this thought, we found an imperative need to collect genetic data from scratch.

C. elegans was our choice of organism. It is a well-studied organism often used as a model for genetic research because it is genetically tractable, that is, the entire *C. elegans* genome has been sequenced and the expression patterns of many genes are known. More than 60% of human genes have homologs in the *C. elegans* genome. The facts that it is small and easy to culture are some of the secondary reasons why this nematode is popular amongst biologists.

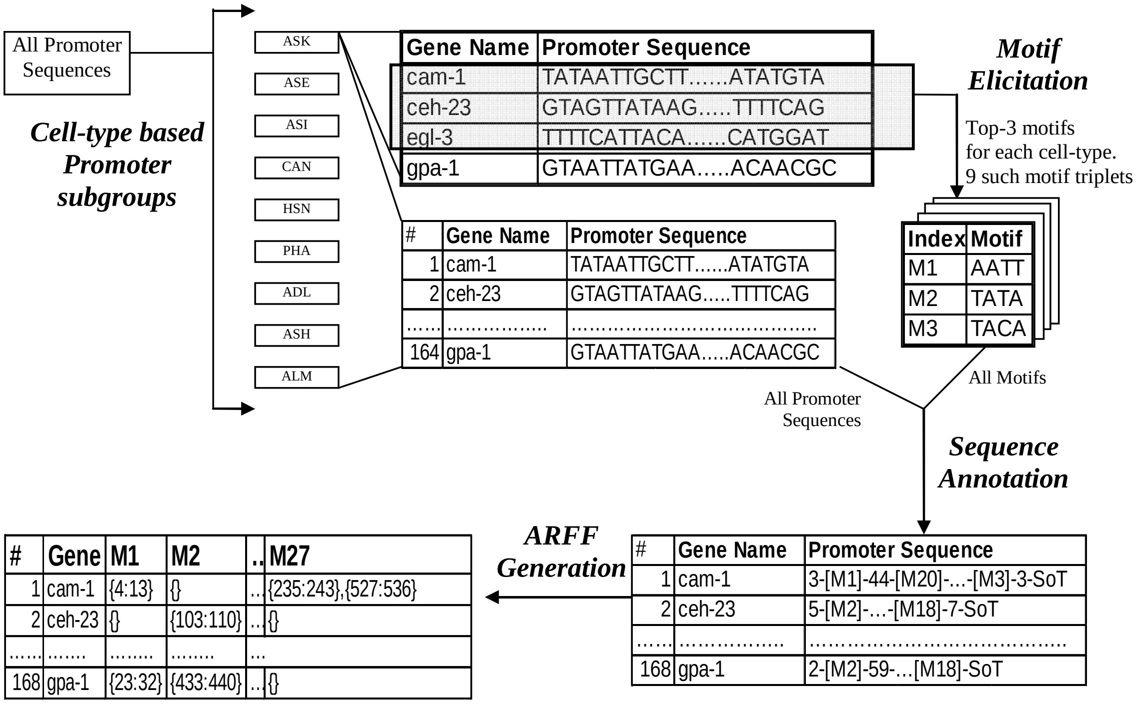
As a first step we identified cell types to include in our study. The primary intent was to identify cell types in which there were at least 30 genes known to be expressed. We require them to be “high-density” cell types, because the data sample should be large enough to derive statistically significant information. Also, this would provide us with substantial data to break down into a training set and a test set.

We used the WormBase [Wor] database and the RSA Database [RSA] for identifying the cell types based on the conditions delineated above. The next step was to gather the actual data, that is, the promoter regions for each of the 30 or more genes per cell type. This data was downloaded from the sources listed above and manually cross verified using BLAST [BLA]. At the end we were able to identify nine cell types of interest to us, each of which had at least 30 known genes. We created nine batches of promoter sequences, one per cell type (Figure 3.1). These nine batches of gene sequences or promoter subgroups were also the initial input for the data transformation process as depicted in Figure 3.2, which provides the graphical overview of the contents of this chapter. Each subsequent step in the data transformation process corresponds to a subsequent section of this chapter.

Another important decision was with regard to the length of the promoter region for the data collected. Based on expert opinion, the length of promoter region included in the data was 5000 base pairs (bp) upstream (5’ to the gene) of the gene. This choice was influenced by the fact that although the regulatory elements critical to gene expression are usually proximal to the initiation site, another type of regulatory elements called enhancers, that may influence expression, can be located

Cell Type	Genes Expressed
ASK	bra-1, cam-1, che-3, daf-11, eat-4, egl-4, gpa-14, gpa-15, gpa-3, ida-1, kin-29, kvs-1, nlp-10, nlp-14, nlp-8, odr-1, opt-3, osm-3, osm-6, osm-9, sra-7, sra-9, srg-2, srg-8, tax-2, tax-4, tax-6, unc-103, zig-4, zig-5
ASE	ceh-23, che-1, che-3, cog-1, csk-1, egl-2, egl-4, gcy-5, gcy-6, gcy-7, gpa-3, hen-1, kvs-1, lim-6, mps-1, ncs-1, nlp-14, nlp-3, nlp-7, npr-1, osm-3, osm-6, osm-9, src-1, tax-2, tax-4, tax-6, unc-5, nlp-1, flp-6
ASI	bra-1, cam-1, ceh-23, che-3, daf-11, daf-28, daf-7, gpa-1, gpa-10, gpa-14, gpa-3, gpa-4, gpa-5, gpa-6, gpc-1, ida-1, kal-1, kin-29, nlp-1, nlp-14, nlp-18, nlp-24, nlp-27, nlp-5, nlp-6, nlp-7, nlp-9, odr-1, opt-3, osm-10, osm-3, osm-6, osm-9, sra-6, srd-1, str-2, str-3, tax-2, tax-4, tax-6, ttx-3, unc-3, zig-3, zig-4
CAN	acy-1, acy-2, cam-1, ced-10, ceh-10, ceh-23, ceh-43, cle-1, ctl-2, dbl-1, ggr-2, goa-1, gpa-10, gpa-14, gpb-2, gsa-1, hbl-1, jkk-1, jnk-1, kal-1, lin-14, mig-2, nlp-10, nlp-15, pak-1, unc-129, unc-73, unc-76, vab-8, cat-1
HSN	cam-1, cdh-3, cha-1, clh-3, ctl-2, eat-16, egl-21, egl-3, egl-43, egl-44, egl-5, flt-1, gar-2, ggr-2, glr-5, goa-1, gpb-2, grd-6, gsa-1, ham-2, hbl-1, ida-1, inx-4, jkk-1, jnk-1, kal-1, mab-23, mec-6, mig-1, mig-2, nhx-5, nid-1, nlp-15, nlp-3, sax-3, sem-4, syg-1, tph-1, unc-103, unc-14, unc-17, unc-40, unc-51, unc-53, unc-73, unc-76, unc-8, unc-86, cat-1, lin-4
PHA	bra-1, ceh-14, che-2, che-3, egl-43, gcy-12, goa-1, gpa-1, gpa-13, gpa-14, gpa-15, gpa-2, gpa-3, ida-1, lin-11, ncs-1, nlp-14, nlp-7, npr-1, ocr-2, osm-10, osm-3, osm-6, osm-9, pkc-1, srg-13, tax-6, unc-103, flp-15, srb-6, tax-2
ADL	cam-1, ceh-23, ceh-32, che-3, cog-1, gpa-1, gpa-11, gpa-15, gpa-3, gpc-1, hlh-2, kvs-1, lin-11, nhr-79, nlp-10, nlp-7, nlp-8, ocr-1, ocr-2, opt-3, osm-3, osm-6, osm-9, qui-1, srb-6, sre-1, sro-1, tax-6, ttx-3, unc-103, ver-2
ASH	cam-1, ceh-23, che-1, che-3, eat-4, egl-3, egl-4, gpa-1, gpa-11, gpa-13, gpa-14, gpa-15, gpa-3, gpc-1, hlh-2, kin-29, kvs-1, mps-1, nhr-79, nlp-15, nlp-3, npr-1, ocr-2, odr-3, opt-3, osm-10, osm-3, osm-6, osm-9, qui-1, sra-6, srb-6, tax-6, unc-42, unc-8
ALM	cam-1, daf-1, deg-3, dyn-1, eat-4, egl-2, egl-21, egl-3, glr-8, goa-1, jkk-1, jnk-1, lin-14, mec-10, mec-2, mec-3, mec-4, mec-6, mec-7, mec-8, mig-2, mps-1, mtd-1, nid-1, pag-3, pat-4, pkc-1, ptl-1, tba-1, tol-1, unc-32, unc-73, unc-86, unc-97

Figure 3.1: List of high density cell types with the associated identified genes that are expressed in the worm's adult life stage.



Gene Name	Promoter Sequence
cam-1	TATAATTGCTT......ATATGTA
ceh-23	GTAGTTATAAG.....TTTTCAG
egl-3	TTTTCATTACA......CATGGAT
gpa-1	GTAATTATGAA.....ACAACGC

#	Gene Name	Promoter Sequence
1	cam-1	TATAATTGCTT......ATATGTA
2	ceh-23	GTAGTTATAAG.....TTTTCAG
……	………………	……………………………………
164	gpa-1	GTAATTATGAA.....ACAACGC

Index	Motif
M1	AATT
M2	TATA
M3	TACA

#	Gene Name	Promoter Sequence
1	cam-1	3-[M1]-44-[M20]-...-[M3]-3-SoT
2	ceh-23	5-[M2]-...-[M18]-7-SoT
……	………………	……………………………………
168	gpa-1	2-[M2]-59-...[M18]-SoT

#	Gene	M1	M2	..	M27
1	cam-1	{4:13}	{}	…	{235:243},{527:536}
2	ceh-23	{}	{103:110}	…	{}
……	…….	……..	……..	…	
168	gpa-1	{23:32}	{433:440}	…	{}

Figure 3.2: Overview of the process from data collection to the ARFF generation. Promoter regions were grouped based on the cell type in which their associated genes were expressed. DNA motifs common to promoters within a group were elicited for each group. All sequences were then annotated with all the elicited motifs. The gene expression information and the annotated sequences, that is, the sequences overlaid with the positional information of each motif, are transformed to ARFF format, Weka's input format.

Sample Promoter Sequence in FASTA format

```
>osm-6 25148409 upstream sequence, from -425 to -1, size 425
TTTTATAATTGCTTATATGTAGTAGTTATATTTTCAGTTTTCATTACATTTCATGGGTAT
TTATTTATTAACTATAATCTTGTATAAGACGATGTAATTATGAAACAACGATTTCACACT
TCCGGTTTTCATGTAAAATTTTTTTCGTTCCAAATAAATTGTTATAAAATTAATTACATC
TTTCATCAAACTTCAAAAATGAAATTGCATTTTTAATAATTAGGAGTCTATTACGGAATT
CATTAAATTTCAGAAAACAAAGTTAACTATATATTTCTCTAGTAGTTCCTTTCCCAGGAG
ACCCTTCCAAGATTTGTATCCACATGTTACCATAGTAACCACTCATTGCTTCTCGCTCAC
ATTGTCTGCTCCCTCTCTTGGGGCTTATATCTCTTTCAAGCTATTACCTTCATTAGTATA
CATCT
```

Figure 3.3: Sample promoter sequence from the data collected.

several thousand base-pairs from the gene. In case another gene was found within the 5000 bp upstream sequence of the promoter sequence, the length of the promoter being considered was truncated at the start of this gene.

The data collected consists of 164 promoter sequences and is a valuable resource for future work. The collected sequences were represented in the FASTA format for further processing by motif discovery and annotation tools. A sample sequence is shown in Figure 3.3.

3.2 Motif Elicitation

The next step in the data transformation process is motif elicitation(Figure 3.2). A motif is a sequence pattern that occurs repeatedly (ideally at least once per gene) in a group of related promoter sequences. Motif elicitation is the process of discovering significant motifs from a group of sequences. There are several tools available that use different statistical modeling techniques to discover significant motifs. The basic premise governing the elicitation process is that the regulatory controls governing the expression of genes in the cells of the same cell type might be in common to at least few of these sequences. The idea was to use motif search programs to identify cell type specific motifs. To avoid over fitting we decided to include only 90% of the sequences as input to motif elicitation programs and the remaining 10% of the sequence for testing.

Two different motif elicitation programs (algorithms), MotifSampler [Mot] and MEME [MEM] were used in the preliminary tests on the same dataset. MotifSampler is a motif finding algorithm that uses GIBBS sampling [JAC95] to find the position probability matrix that represents the motif. MEME discovers one or more motifs in a collection of sequences by using the technique of expectation maximization [BE94]. Based on the similarity of the outputs from both algorithms and considerably less computational time exhibited by MotifSampler, GIBBS was our choice of algorithm to be used for the motif elicitation stage. It is worth noting, however, that the preliminary tests were in no way a detailed comparative study of the two methods.

Based on the input from the domain expert it was decided to look for motifs of size 8, 10, or 12 base pairs. Thus, motif elicitation was performed by executing MotifSampler individually on each of the nine batches of promoter sequences (Figure 3.2), once for each size. The output of this process was in the form of a set of position probability matrices (Figure 3.4) each representing a motif. For each combination of a size and batch, the three best scoring matrices (motifs) were selected for future use. As a result we ended up with 81 motifs; nine motifs for each cell type.

3.3 Sequence annotation

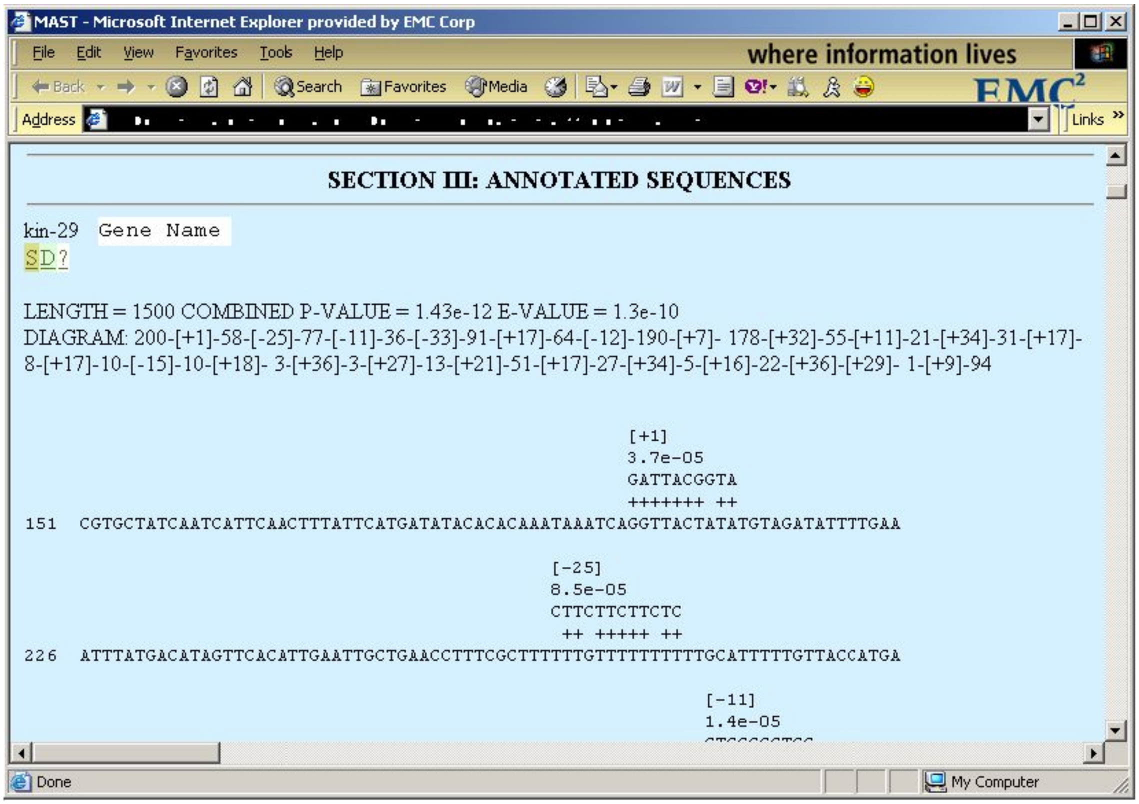
Motifs identified by MotifSampler are in the form of a probability-matrix. The next step was to annotate all promoter sequences with the elicited motifs. Annotation

Sample motif represented as a position probability matrix

```
# Width = 8
# Consensus = ATAACTAG
#       A          C          G         T
    0.996545  0.000890  0.000840  0.001726
    0.001520  0.000890  0.000840  0.996750
    0.499033  0.000890  0.000840  0.499238
    0.499033  0.000890  0.000840  0.499238
    0.001520  0.498402  0.498352  0.001726
    0.001520  0.000890  0.000840  0.996750
    0.996545  0.000890  0.000840  0.001726
    0.001520  0.000890  0.498352  0.499238
```

Figure 3.4: Sample motif represented as a position probability matrix. The commented row titled width provides the length of the motif. The consensus sequence of the motif is a sample occurrence of the motif built by using the most common base at each position. Each column of the position probability matrix corresponds to the bases A, C, G and T respectively. Each row corresponds to a position of a base within the motif. For instance a position probability matrix for a motif with width 8 would consist of 8 rows. The value at row i column j in the matrix is the probability of finding the base j at position i in the motif.

is the process of finding matches of a given motif(s) in a given set of sequences and also quantifies how good each match is. We used Motif Alignment and Search Tool (MAST) [MAS] for the annotation process. A PERL script (Appendix B) was developed to convert probability-matrices from MotifSampler output to corresponding MAST friendly format(log-odds matrices as shown in Figure 3.6). A master motif file was created which consisted of all 81 identified motifs irrespective of the cell type. A master gene sequences file was created which consisted of promoter sequences for all high-density genes listed in Figure 3.1. The master gene sequences file and the master motif file were fed to MAST as input to annotate all promoter sequences with all elicited motifs. The MAST output file is a HTML file consisting


MAST - Microsoft Internet Explorer provided by EMC Corp
File Edit View Favorites Tools Help
where information lives
Back Search Favorites Media
EMC²
Address [redacted]
Links »
Done
My Computer

SECTION III: ANNOTATED SEQUENCES

kin-29 Gene Name
SD?

LENGTH = 1500 COMBINED P-VALUE = 1.43e-12 E-VALUE = 1.3e-10
DIAGRAM: 200-[+1]-58-[-25]-77-[-11]-36-[-33]-91-[+17]-64-[-12]-190-[+7]- 178-[+32]-55-[+11]-21-[+34]-31-[+17]-8-[+17]-10-[-15]-10-[+18]- 3-[+36]-3-[+27]-13-[+21]-51-[+17]-27-[+34]-5-[+16]-22-[+36]-[+29]- 1-[+9]-94

```
                                                       [+1]
                                                       3.7e-05
                                                       GATTACGGTA
                                                       +++++++ ++
151  CGTGCTATCAATCATTCAACTTTATTCATGATATACACACAAATAAATCAGGTTACTATATGTAGATATTTTGAA

                                                [-25]
                                                8.5e-05
                                                CTTCTTCTTCTC
                                                 ++ +++++ ++
226  ATTTATGACATAGTTCACATTGAATTGCTGAACCTTTCGCTTTTTTGTTTTTTTTTTGCATTTTTGTTACCATGA

                                                              [-11]
                                                              1.4e-05
```

Figure 3.5: MAST annotated sequence sample.

During annotation each supplied sequence is searched for matches with each supplied motif. The four lines above each motif occurrence contain, respectively, the motif number of the occurrence, the position p-value (i.e., a measure of the match quality, lower is better) of the occurrence, the consensus sequence of the motif, and a plus sign ('+') above each letter in the occurrence that has a positive match score to the motif. MAST can automatically generate the reverse complement strand for each supplied sequence and search for motif occurrences on either the given strand or its reverse complement. The ('+') or ('-') sign alongside the motif number is used to distinguish whether the match occurred on the given strand or the reverse complement respectively.

Sample motif represented as a log odds matrix

```
# Width = 8
# Consensus = ATAACTAG
# A       C       G       T
166     -766    -757    -764
-770    -766    -757    154
66      -766    -757    54
66      -766    -757    54
-770    146     164     -764
-770    -766    -757    154
166     -766    -757    -764
-770    -766    164     54
```

Figure 3.6: Sample motif represented as a Log-odds matrix.
This matrix is a log-odds matrix calculated by taking the log (base 2) of the ratio p/f at each position in the motif where p is the probability of a particular letter at that position in the motif, and f is the average frequency of that letter in the training set.

of all annotated gene sequences. A relevant section of the output file is shown in Figure 3.5.

3.4 ARFF Generation

A Java module was developed to transform the union of annotated promoter sequences from the MAST output and the known gene expression information, that is, all of the cells in which the gene is expressed, to Attribute-Relation File Format (ARFF) format dataset. The ARFF format is a format similar to the relational database format. Each gene occurs as a tuple in the relation. Each motif is an attribute of the gene tuple. A set-value consisting of the known gene expression pattern is also an attribute of the gene. A sample of the same has been included in Appendix A. As mentioned in section 3.2, to avoid mining and exploring relationships that overfit the data, the sequences were divided into a training set (90% of the sequences) and a test set (10% of the sequence). Thus the ARFF generation was performed twice, one to create a training set ARFF file and the other for a test set ARFF file.

Chapter 4

Data Visualization and Mining

4.1 Mining Process

The sequence annotation process described in Chapter 3, transformed the gene sequences into a relational format. This provides us the ability to utilize algorithms from the WPI-Weka system to discover relationships or patterns that describe gene expression. These relationships are rules like "Genes whose promoter regions contain the motifs M10 and M16 are often expressed in Neural cells". The ability to uncover such relationships in an automated fashion is of prime importance, as they provide the domain experts with a relevant view of the data that demands further exploration.

[Ice03] investigated the problem of incorporating hypothesis based information into the mining process. We instead decided to follow a post mining exploratory approach for hypothesis testing, that is, to specialize (post-process) *interesting* rules produced by conventional association rule mining algorithms. The reasons that influenced this choice were:

1. It is very difficult, if not impossible, to determine in advance the right distance-

related and position-related mining parameters so that the mined rules will capture the desired patterns. This is due to the fact that the appropriate distance and position values vary for each subset of motifs in the context of each cell type under consideration. For instance, the appropriate values for the distance and relative position of motifs M10 and M16 in neural cells might be very different to those in muscle cells. Furthermore, these values might vary in the presence of other motifs. That is, the distance and relative position of motifs M10 and M16 in neural cells might vary even for the same cell type once that say motif M20 is added to the mix. Since mining association rules is expensive in terms of execution time, a trial and error approach in which possible values of the input parameters are guessed would require multiple executions of the mining algorithm, which would be too time consuming. Another alternative is to perform an automatic search for the right values of those parameters within the mining algorithm, but the time complexity of an exhaustive search is exponential in the size of the input data and would make the runtime of the mining algorithm prohibitive. It is unclear that good heuristics to prune the search are possible. Our selected post mining exploratory analysis approach is much faster since it is performed on one rule (i.e., one subset of motifs under one cell type) at a time, and only on such rules that are deemed interesting by the domain expert user.

2. Exploratory analysis often combines visualization with data mining tools that provides a *simple* sequence of steps(work flows) to identify and isolate *relevant* information. *Simplicity* is important because we need to bridge the gap between the domain experts and the tools usage. Equipped with an easy to use tool, the domain experts could utilize their knowledge to analyze and define the data patterns displayed by the tool in an intuitive manner. The tool is

not limiting but instead banks on the experts, and it could easily evolve to accommodate newer hypotheses.

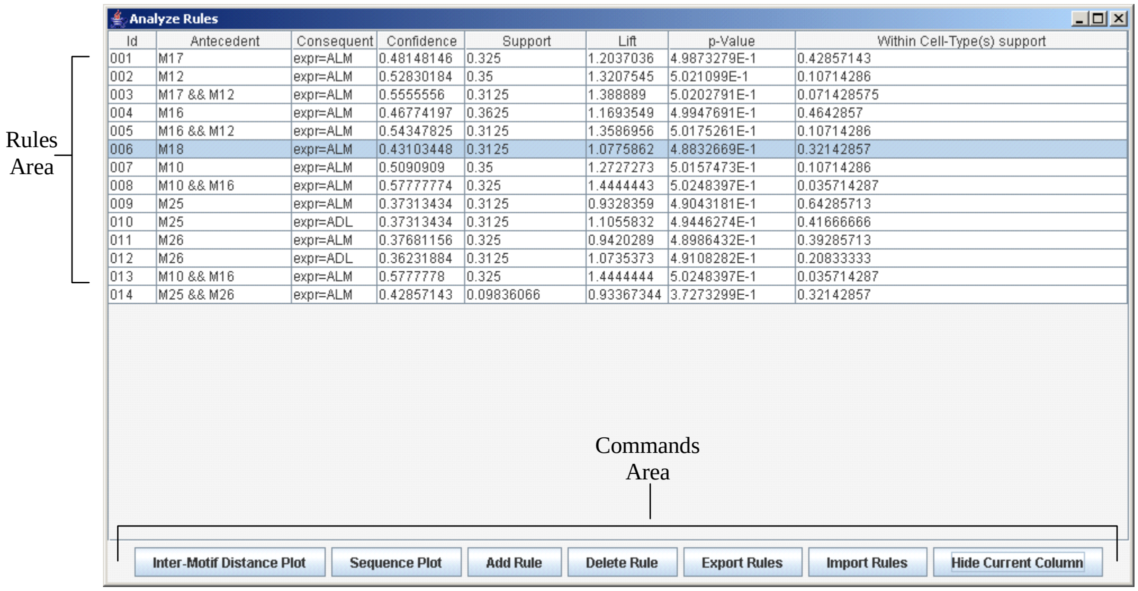
The association rule mining modules ([Sho01] and [PR05]) of the WPI-Weka system [WPI] were used to mine for basic gene expression patterns in the annotated sequences. The visualization system that we develop here facilitates exploratory analysis to specialize these mined patterns. [Rudss] contributions to WPI Weka helped integrate the visualization modules into the WPI Weka in a transparent fashion. The user now has the choice to either save the results of the mining for perusal at a later point in time or could directly invoke the visualization modules from the mining modules.

4.2 Visualization and Specialization Module

Some of the prime contributions of this work are the Visualization and Specialization Module (VSM) which is an extension to the WPI Weka system. This module enables visualization of the annotated promoter sequences in the context of a specific rule or a set of motifs. The primary interface of the VSM is the Analysis frame, which is the first screen to be displayed when VSM is invoked is depicted in Figure 4.1. We explain the Analysis frame below.

4.2.1 Analysis Frame

The analysis frame is the focal point for using visualization extensions to WPI Weka. The analysis frame loads with two sections, the **Rules** area and the **Commands** area (Figure 4.1). We explain the Rules area below and we explain each of the options in the Commands area in subsequent subsections. The Rules area is used to display base association rules along with the corresponding values for certain



Analyze Rules

Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.42857143
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.10714286
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.071428575
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.4642857
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.10714286
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.32142857
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.10714286
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.035714287
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.64285713
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.41666666
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.39285713
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.20833333
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.035714287
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.32142857

Figure 4.1: Sample Analysis Frame.

Analysis Frame can either be invoked from the mining interface of the WPI-Weka system or could be invoked as a standalone application using an exported set of mined association rules, the associated MAST results (HTML format), and a list of gene names alongside the known expression patterns. A sample of the gene expression information file is presented in Figure 4.2.

Sample Gene Expression Information File

```
nlp-27, ASI
gpc-1 , ASH^ASI^ADL
ocr-1 , ADL
gcy-12, PHA
cat-1 , HSN
tba-1 , ALM
```

Figure 4.2: A sample gene expression information file.

The file is in CSV (Comma Separated Value) format. First value in each row is the gene name (e.g., nlp-27). The next value contains a set of cell-types the gene is known to be expressed in. Elements of the set are separated using the ˆ symbol.

measures of interestingness. The design is extensible; that is, new measures of interest could be added in the future with minimal code changes. In the current state, a Rule tuple consists of the following items:

- **Id** - this is a unique id assigned to the rule. The usability of this field increases once the user starts to generate specializations from the rule, as the Id column helps us trace the history or the specialization path of new rules.

- **Antecedent** - The left-hand side of the rule. It contains the motifs present in the rule.

- **Consequent** - The right-hand side of the rule. It contains the cell-types predicted by the rule.

- **Support** - As discussed in Section 1.2, the support of a rule is the relative frequency with which the antecedent and consequent appear together in the data. That is, P(Antecedent & Consequent).

- **Confidence** - As discussed in Section 1.2, the confidence is the likelihood that the consequent appears in a data instance that contains the antecedent. That is, P(Consequent|Antecedent).

- **Lift** - The lift value of an association rule is another measure to try to quantify the interestingness of the rule. It is defined as the ratio of the confidence of the rule and the support of the consequent of the rule [BMS97]. In other words

$$lift\left(rule\right) = p\left(consequent|antecedent\right)/p\left(consequent\right)$$

- **p-Value** - The p-value of the rule is the probability that the antecedent and the consequent would be as highly correlated as they are, just by chance according to a Chi square test of independence. We calculate the p-value of an association rule using the approach in [Alv03].

- **Within Cell-Type Support** - Provides the support of the rule among only those instances of the data that contain the consequent of the rule. This metric is very important in the context of this work because we expect to see different motifs and rules for different cell types, so we are primarily interested in the support of the rule within each cell type.

The Commands area of the Analysis frame provides buttons to perform a range of functions. The following subsections describe each of these functions provided by the visualization extensions via the analysis frame. It is important to note that most of these functions are invoked in the context of a specific rule and so it is necessary to select a rule in the rules area of the analysis frame before invoking a command.

4.2.2 Inter-Motif Distance Plot

Selecting a Rule in the rules area and then invoking the inter-motif distance plot via the button with the same label lets a user visualize the data in the context of the rule from an inter-motif distance perspective (Figure 4.3). This action enables a user to perform exploratory analysis in the context of the hypothesis - "Inter-motif distance influences gene expression".

On invoking this command a new frame with the pairwise inter-motif distance plot(s) is displayed. It displays one graph for each pair of motifs in the rule (selected in the Analysis Frame). For the sake of simplicity we start with a rule with only two motifs. We revisit plots originating from rules consisting of more than two motifs later in the section.

Notice that an inter-motif distance plot (Figure 4.3) is sliced into 2 parts by a

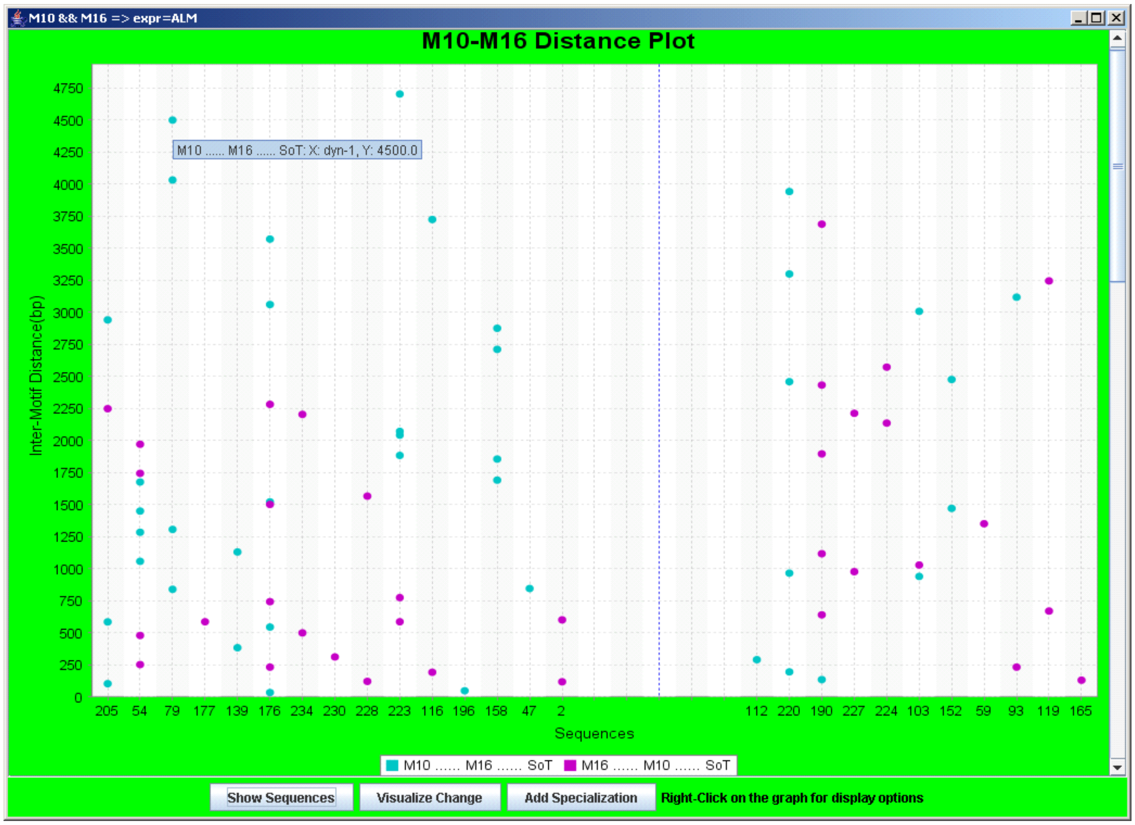


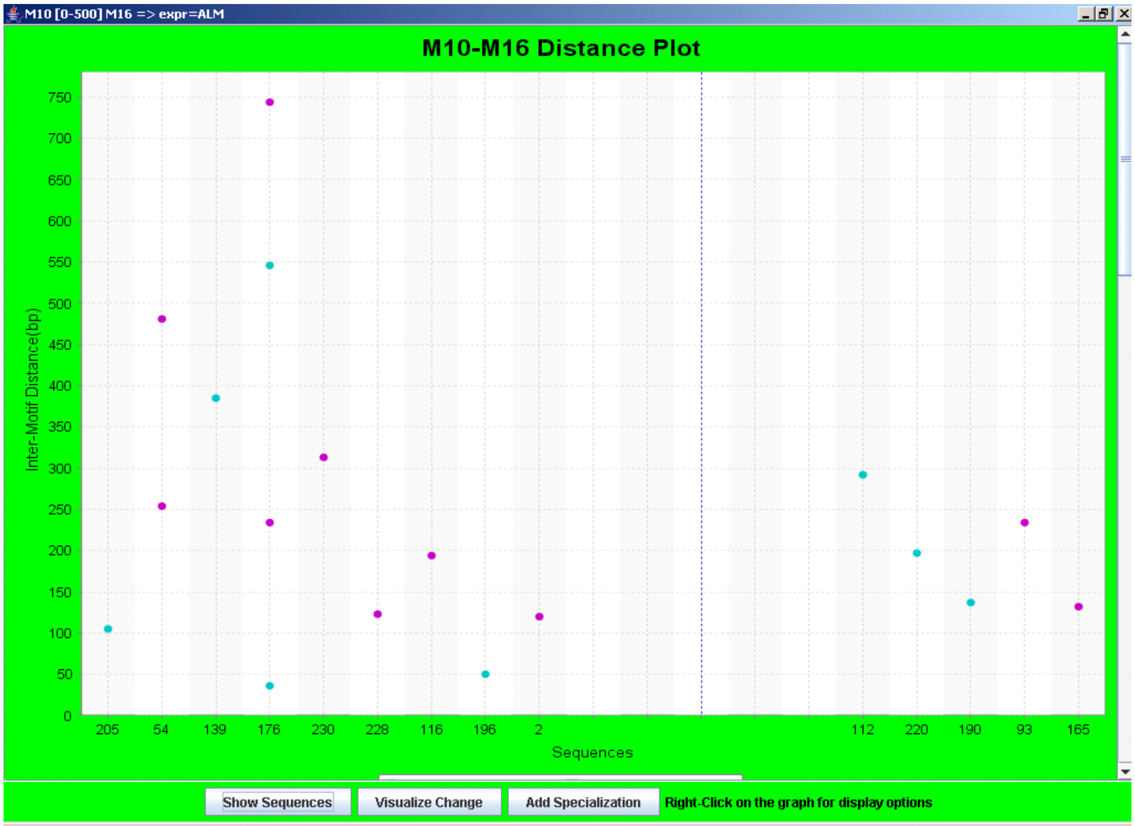
Figure 4.3: Sample Inter-Motif Distance Frame.

Each graph is displayed with the rule used to establish the context as the title of the frame. Each graph displays the pairwise inter-motif distance plots(M10 && M16 ⇒ expr=ALM, in this case). Along the x-axis of the plot are the id's of the genes in question and along the y-axis are the inter-motif distances between the pair of motifs. For each pair (a,b) of motifs, inter motif distances of all occurrences of motif a from all occurrences of motif b are plotted. The color of each point is indicative of the order of occurrence of motifs a and b relative to the SoT. In this graph, aqua (light) denotes points in which the occurrence of M16 appears in between the occurrence of M10 and SoT; and magenta (dark) denotes points in which the occurrence of M10 appears in between the occurrence of M16 and SoT. Each graph lists only those genes on the x-axis, that support the antecedent of the rule. That is, genes whose promoter regions contain at least one occurrence of each motif. Genes on the left of the dotted line also support the consequent of the rule.

dotted line. The genes on the left of this dividing line are the ones that support the antecedent and the consequent of the rule and hence support the rule. The ones on the right are the genes that support only the antecedents of the rule. This provides the user with an easy mechanism to discover inter-motif distance based patterns on the left-hand side of the plot that are not as frequent on the right hand side, as this would let us explore specializations with improved classification accuracy(and/or confidence). Once the user has utilized the dotted separation and inter-motif distance plots to define a range of interest, for instance a range of (0-500) between motifs M10 and M16, the user can invoke the “Visualize Change” command to visualize the data in the context of the specialization as depicted in Figure 4.4. Subsequently the specialized rule could be added to the Analysis Frame using the “Add Specialization” command on the inter-motif distance plot. This causes a new entry to be inserted in the Analysis Frame (Figure 4.5) with the following specialization

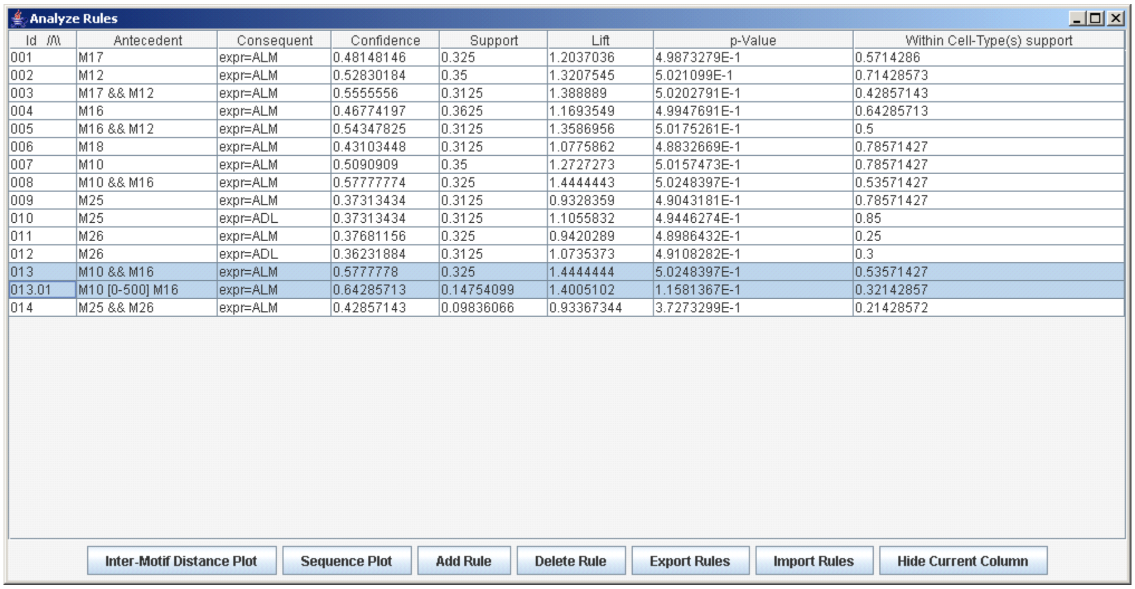
$$M10(0 - 500)M16 \Rightarrow expr = ALM \tag{4.1}$$

Note that the Id field is auto-generated in a fashion that always lets a user trace back the steps in case we want to later recall which rule was used to derive the specialization.



Right-Click on the graph for display options

Figure 4.4: Visualize change command from the inter-motif distance plot. The Visualize Change command from the inter-motif distance plot enables the user to visualize the data in the context of the specialization. This plot depicts the specialization *M10 [0-500] M16* ⇒ *expr=ALM* of the original rule *M10 && M16* ⇒ *expr=ALM* from Figure 4.3.



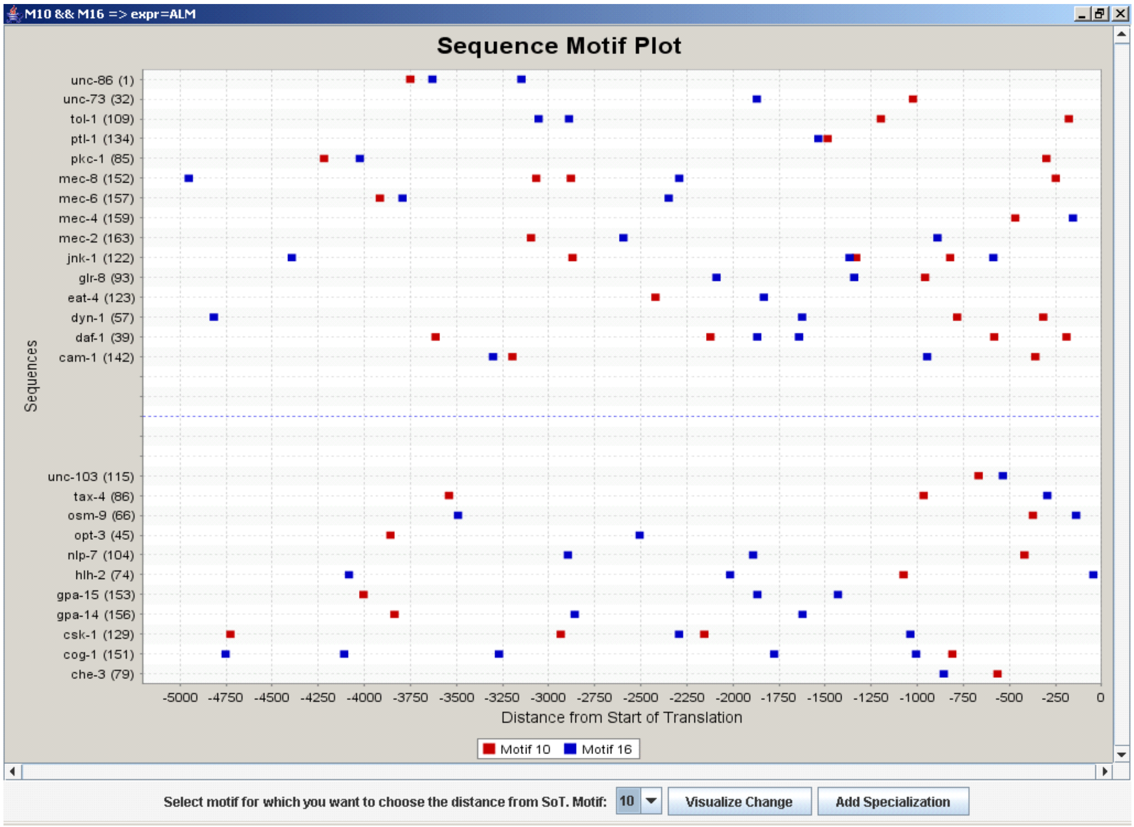
Id /\	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.78571427
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.53571427
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.78571427
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.85
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.25
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.3
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.53571427
013.01	M10 [0-500] M16	expr=ALM	0.64285713	0.14754099	1.4005102	1.1581367E-1	0.32142857
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572

Figure 4.5: A row representing the addition of a specialization to the Analysis Frame.

4.2.3 Sequence Plot

Selecting a Rule in the rules area and then clicking the sequence plot button displays a visualization all the qualifying gene sequences in the context of the rule. This action enables a user to perform exploratory analysis in the context of the hypothesis - "Distance of motifs from the SoT influences gene expression". A qualifying gene sequence is one that has at least one occurrence of each motif that appears in the rule (selected in the Analysis Frame). Invoking this command causes a new frame with the sequence plot overlaid with the motif information to be displayed.

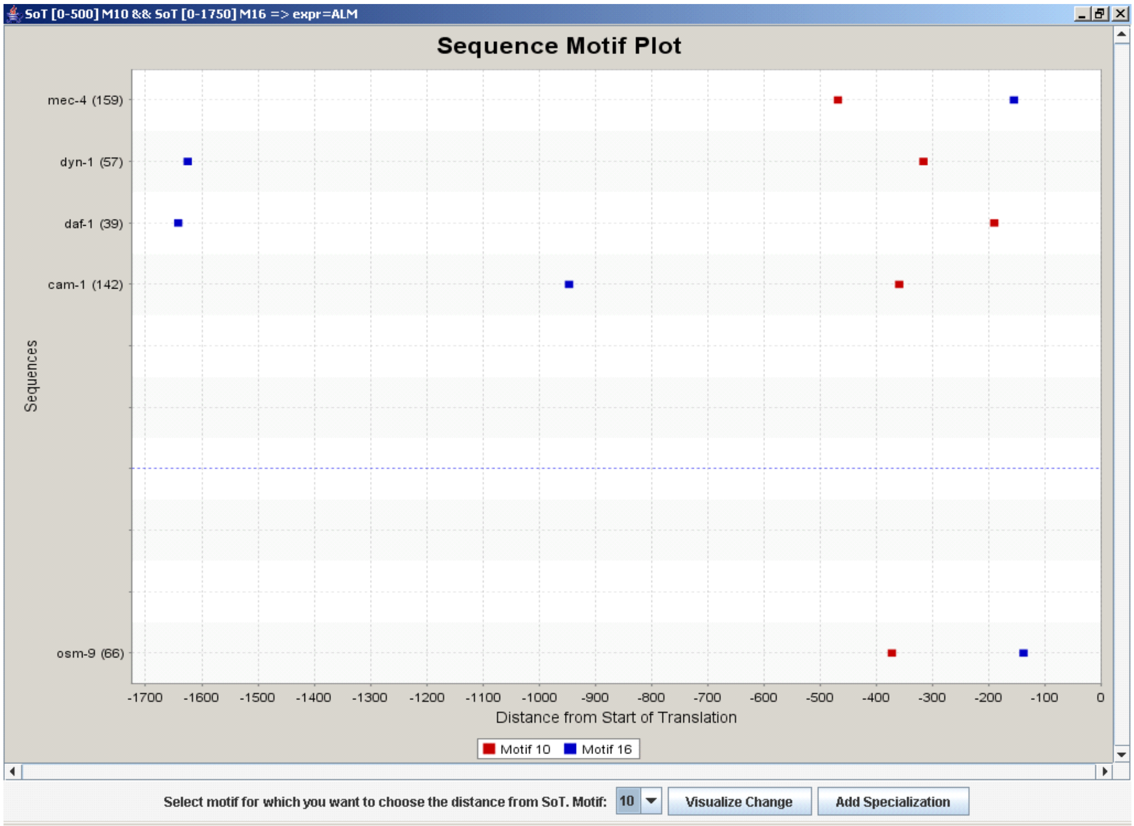
This sequence plot graph provides the user with an easy mechanism to discover "Distance from SoT" based patterns in the upper part of the plot that are not as frequent in the lower part as this would let us explore specializations with improved classification accuracy(and/or confidence). Once the user has utilized the dotted separation of the plot (into rule supporting and antecedent supporting) and the



Select motif for which you want to choose the distance from SoT. Motif: 10

Figure 4.6: Sample Sequence Plot Frame.

Each graph is displayed with the rule used to establish the context as the title of the frame(M10 && M16 ⇒ expr=ALM, in this case). This sequence plot displays all gene sequences that contain occurrences of the participating motifs (i.e., motifs M10 and M16). Along the y-axis is the list of gene promoters, that support the antecedent of the rule. That is, the gene promoters that contain at least one occurrence of each of the rule motifs. The x-coordinate of each point in the plot is the distance of the motif from the SoT, which is the far right end of the plot. The color of the point is used to identify the motif. The graph is sliced into two parts by a dotted line. The genes in the upper part of this dividing line are the ones that support the consequent of the rule and hence support the rule. The ones in the lower part are the genes that support only the antecedents of the rule.

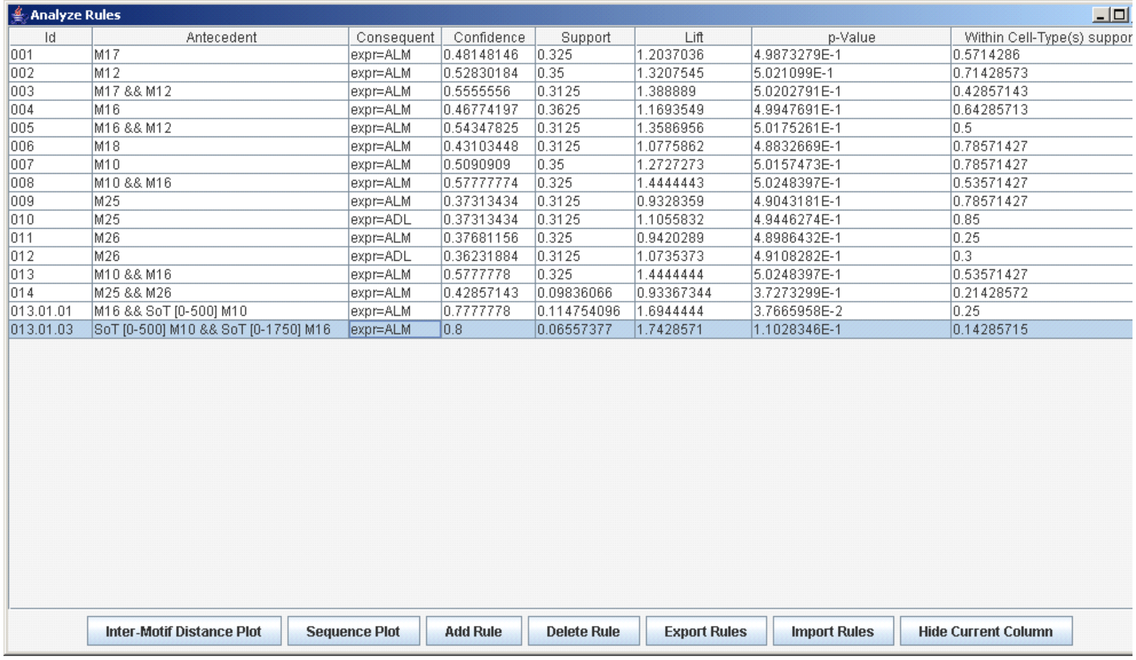


Select motif for which you want to choose the distance from SoT. Motif: 10

Figure 4.7: Visualize change command from the sequence plot.
The Visualize Change command from the sequence plot enables the user to visualize the data in the context of the specialization. This plot depicts the specialization *SoT [0-500] M10 && SoT [0-1750] M16* ⇒ *expr=ALM* of the original rule *M10 && M16* ⇒ *expr=ALM* from Figure 4.6.

rule specific sequence plots to identify a "distance from SoT" clause of interest, the "Visualize Change" command could be invoked to visualize the data in the context of the specialization rather than the original rule as depicted in Figure 4.7. Again the title of the new window is indicating the context setting rule/specialization. If the user finds the specialization of interest, it can be added to the Analysis Frame using the "Add Specialization" command on the new sequence plot. Again this causes the specialization to appear as a new entry in the Analysis Frame with an auto-generated Id that again lets a user trace back the steps in case the user wants to later recall which rule was used to derive a specialization (Figure 4.8). In case the base rule consisted of more than 1 motif and multiple "distances from SoT" relationships are defined (one for each motif) each such relationship is represented as a term and a collection of independent terms constitutes the specialized rule. For instance see Figure 4.8 for the following specialization, which is interpreted as: "An occurrence of Motif 10 within 500 bp from the SoT and the presence of an occurrence of Motif 16 within 1750 bp from the SoT implies that the gene is expressed in cells of type ALM".

$$SoT\,[0-500]\,M10 \quad \&\& \quad SoT\,[0-1750]\,M16 \Rightarrow expr = ALM \tag{4.2}$$



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) suppor
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.78571427
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.53571427
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.78571427
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.85
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.25
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.3
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.53571427
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572
013.01.01	M16 && SoT [0-500] M10	expr=ALM	0.7777778	0.114754096	1.6944444	3.7665958E-2	0.25
013.01.03	SoT [0-500] M10 && SoT [0-1750] M16	expr=ALM	0.8	0.06557377	1.7428571	1.1028346E-1	0.14285715

Figure 4.8: Analysis Frame with two distance from SoT based specializations.

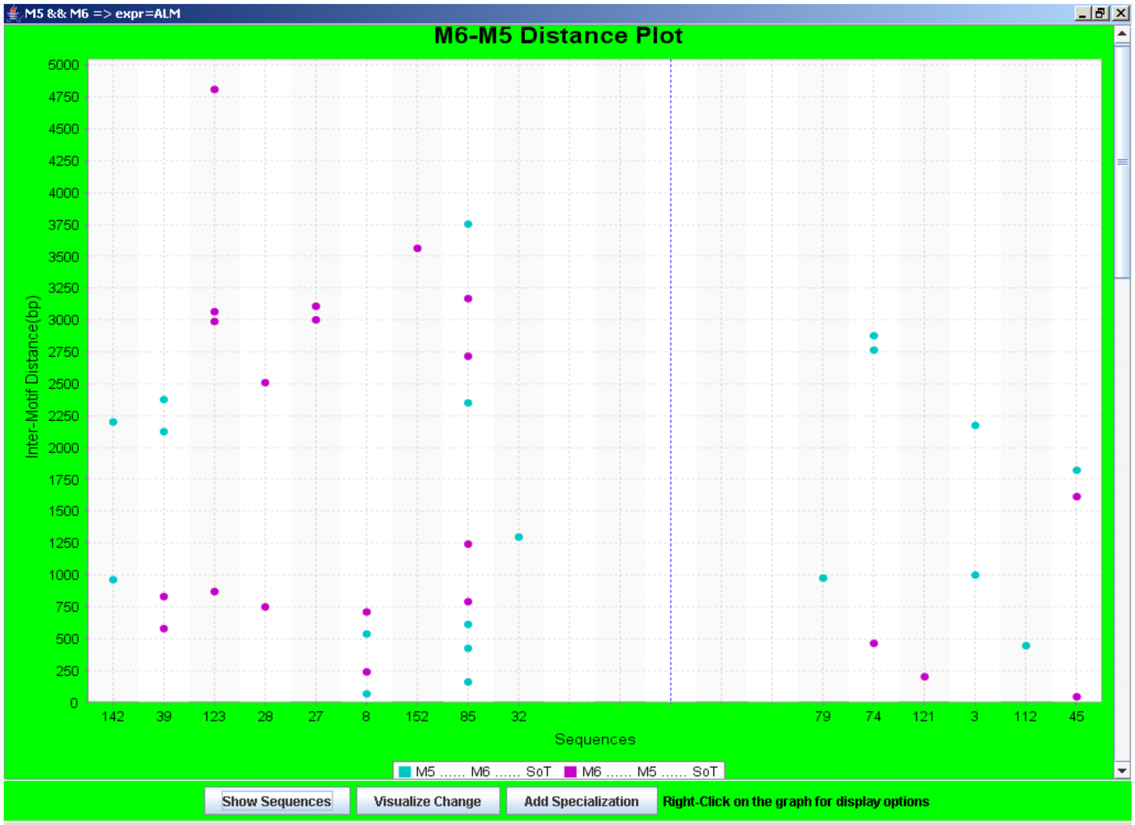
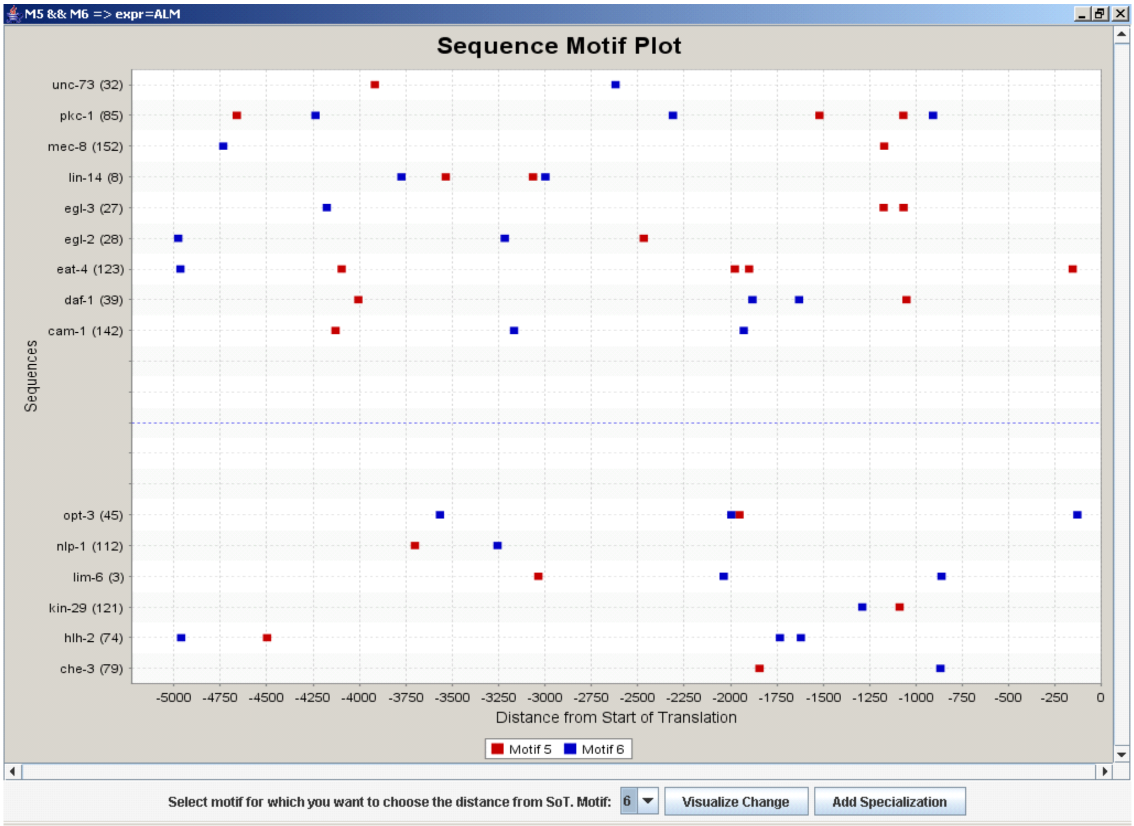


Figure 4.9: Inter-Motif Distance Plot for Motifs M5 and M6. Observe the lack of magenta (dark) dots in the right part of the frame.

4.2.4 Order of occurrence of motifs

We wanted the visual extensions to also accommodate exploratory analysis based on the hypothesis "The order of occurrence of motifs influences gene expression." But during the system design and the system use by the team(including the domain expert) it was observed that we already had a few ways to visualize gene sequence data in the context of the "order of the occurrence" of motifs and hence a new plot was not created. If order of occurrence of motifs was important it could be easily identified by one of three ways: the color of the points in inter-motif distance plot, a repeating sequence of color in the sequence plot, or through the operation of the ASAS mining algorithm itself.



Select motif for which you want to choose the distance from SoT. Motif: 6

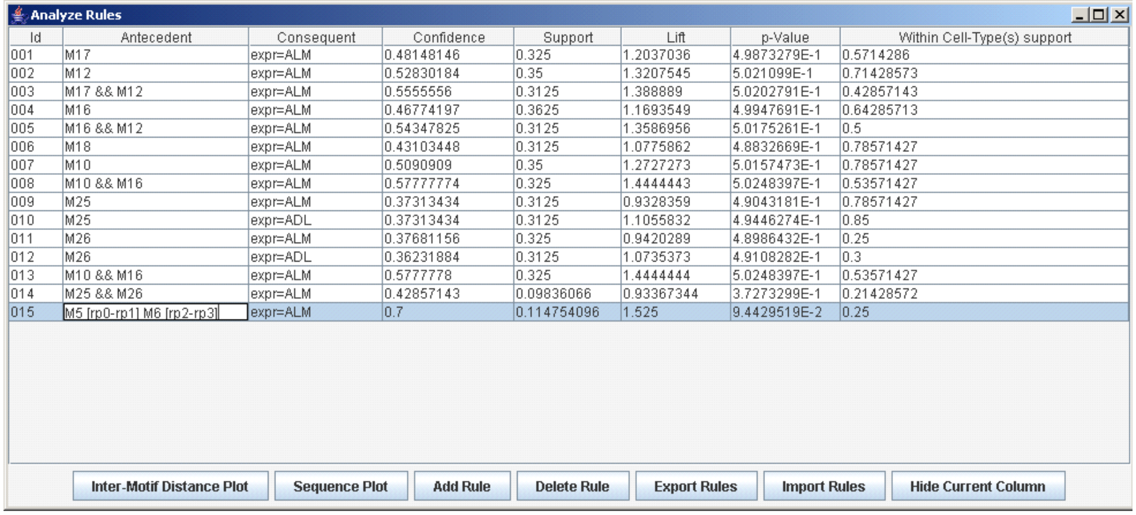
Figure 4.10: Sequence Plot for Motifs M5 and M6. Observe that in the rule supporting sequences (upper part) a red-dot is usually followed by a blue dot scanning the gene sequence from right end (SoT) to left.

Color of the points in an inter-motif distance plot. The order of the motifs in the inter-motif distance plot is represented by color. For instance in Figure 4.3, M10 to the right of M16 (i.e., M16..M10..SoT) is represented by a magenta (dark) dot, while M16 to the right of M10 is represented by a aqua (light) dot. Thus, color provides a quick visual clue whether the order of occurrence of motifs affects gene expression. If that is the case, the left part of the plot should have more points of one color than the other part.

Repeated sequence of color in the sequence plot. As mentioned in Section 4.2.3, the sequence plot displays all instances of participating motifs for qualifying sequences as they occur on the gene relative to the SoT. Since each motif appears in its own color and the data is being visualized in the context of a single rule, one can often see a repetitive pattern of color in the upper part and a lack of the same in the lower part of the plot. Such a display could also indicate an influence of order of occurrence on gene expression.

ASAS mining algorithm. The association rule mining module from the WPI-Weka System [PR05] is capable of mining association rules with order/position based information and hence it is possible to have some of these rules with order of occurrence of motifs available already at the beginning of the exploratory analysis. Either of the two means mentioned above could be used to visually confirm/observe the order of occurrence relationship.

4.2.5 Adding Rules Manually



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.78571427
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.53571427
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.78571427
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.85
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.25
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.3
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.53571427
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572
015	M5 [rp0-rp1] M6 [rp2-rp3]	expr=ALM	0.7	0.114754096	1.525	9.4429519E-2	0.25

Figure 4.11: Add Rule option in the Analysis Frame provides for free text option to add rules.

We wanted to allow users to type in specialized rules manually, particularly in the cases of rules involving order, as well as certain more complex 'hybrid' rules discussed in detail below. To allow this option, we needed to write a grammar (Figure 4.12) to parse the rules. JavaCC [JCC], Java equivalent of LEX and YACC, was used to code the grammar and auto-generate the rule parser. The user can then simply type in the Antecedents and the Consequent of the rule to calculate the different statistics indicating the interestingness metrics of the rule as shown in Figure 4.11. Simply typing a complete rule computes the statistics indicating the interestingness of the rule. The user could also visualize the new rule using either the sequence plot or the inter-motif distance plot.

A rule keyed in by the user which does not adhere to this grammar results in an error as shown in Figure 4.13

4.2.6 Hybrid Rule

As described in the grammar governing rule definitions, each rule consists of an antecedent and a consequent. Antecedents in turn consist of terms. A rule could also include specialized term, extra hypothesis-based information(constraints) that the instances of the participating motifs must satisfy in order for a gene sequence to support the rule.

The system also supports hybrid rules, rules that consist of specialized terms based on different hypotheses and a gene sequence must satisfy all constraints in order to support the rule. In Figure 4.10 note that there exists an occurrence of M5 (the red dot) usually within the first 1600 bp from the SoT (far right end of the plot). Also note that there is an occurrence of Motif M5 between an occurrence of Motif M6 (blue dot) and the SoT. It is interesting to combine the two observations into a rule as follows and visualize it or calculate its interestingness. As we see in Figure 4.14 that this hybrid specialization:

Grammar

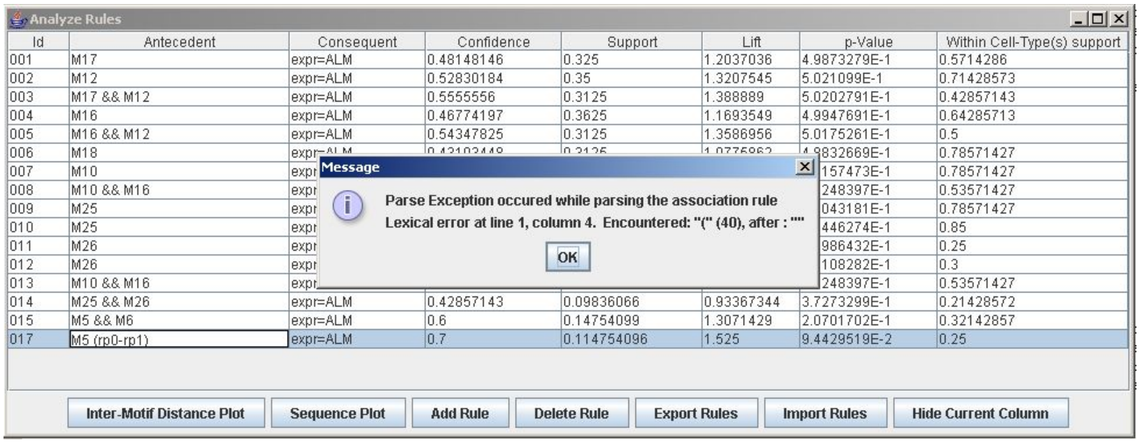
```
The grammar is defined as a 4-tuple: (SIGMA, N, P, S):
    SIGMA is an alphabet of terminal symbols
    N is an alphabet of non-terminal symbols
    P is a set of production rules
    S in N is the start symbol

Sigma   :=  {SoT, Mn, Mn(rp i - rp i+1), \&\&, (x-y)}
N       :=  {S, L1, L2, C1, C2, CL1, CL2, CL3, CL4, S}
P :=  {
        S = Term | (Term C1 Term) | (T C1 Term)
        C1 = &&                        # And
        Term = CL1 | CL2 | CL3 | CL4
        CL1 = L1 C1 L1 | CL1 C1 L1     # Covers rules based on presence.
        CL2 = L1 C2 T                  # Covers literals of the form distance from SoT.
        CL3 = L1 C2 L1 | CL3 C2 L1     # Covers literals of the form Mi at a distance of
                                       #  x-y from Mj
        CL4 = L2 C1 L2 | CL4 C1 L2     # Covers literals of the form Mi occurs before Mj
        L1 = Mn                        # Motif n
        T = SoT                        # Start of Translation
        C2 = (x-y)                     # Where x and y are integers such that y > x
        L2 = Mn (rp i - rp i+1)        # Motif n exists from Relative position i to i+1
                                       # on the gene sequence.
    }
```

A sample derivation of an antecedent from the grammar is shown below:

```
S :=  Term
      CL1
      CL1 C1 L1
      CL1 && M3
      L1 C1 L1 && M3
      M1 && M2 && M3
This antecedent is interpreted as:
If Motif 1 is present and Motif 2 is present and Motif 3 is present.
```

Figure 4.12: Grammar to parse rules.


Analyze Rules

Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.9832669E-1	0.78571427
007	M10	expr				157473E-1	0.78571427
008	M10 && M16	expr				248397E-1	0.53571427
009	M25	expr				043181E-1	0.78571427
010	M25	expr				446274E-1	0.85
011	M26	expr				986432E-1	0.25
012	M26	expr				108282E-1	0.3
013	M10 && M16	expr				248397E-1	0.53571427
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572
015	M5 && M6	expr=ALM	0.6	0.14754099	1.3071429	2.0701702E-1	0.32142857
017	M5 (rp0-rp1)	expr=ALM	0.7	0.114754096	1.525	9.4429519E-2	0.25

Message

Parse Exception occured while parsing the association rule
Lexical error at line 1, column 4. Encountered: "(" (40), after : ""

Figure 4.13: Grammar based parsing helps identify user errors in typing the rule.

$$\text{M5 [rp0-rp1] M6 [rp2-rp3] \&\& SoT [0-1600] M5} \;\Rightarrow\; \text{expr=ALM}$$
$$[Confidence = 0.8333333, Support = 0.08196721] \quad (4.3)$$

has a higher confidence as compared to the following simpler "order of occurrence" specialization

$$\text{M5 [rp0-rp1] M6 [rp2-rp3]} \;\Rightarrow\; \text{expr=ALM}$$
$$[Confidence = 0.8333333, Support = 0.08196721] \quad (4.4)$$

As seen above hybrid specialization could have multiple specialized terms that relate to a single motif. A hybrid specialization could post multiple constraints on the same motif like Distance from SoT and Order of occurrence relative to another motif. It is important to note that although the rule may have multiple constraints for the same motif, it is not required that the same instance of the motif satisfies each of them. In the context of the 4.3 above, it is not required that the instance

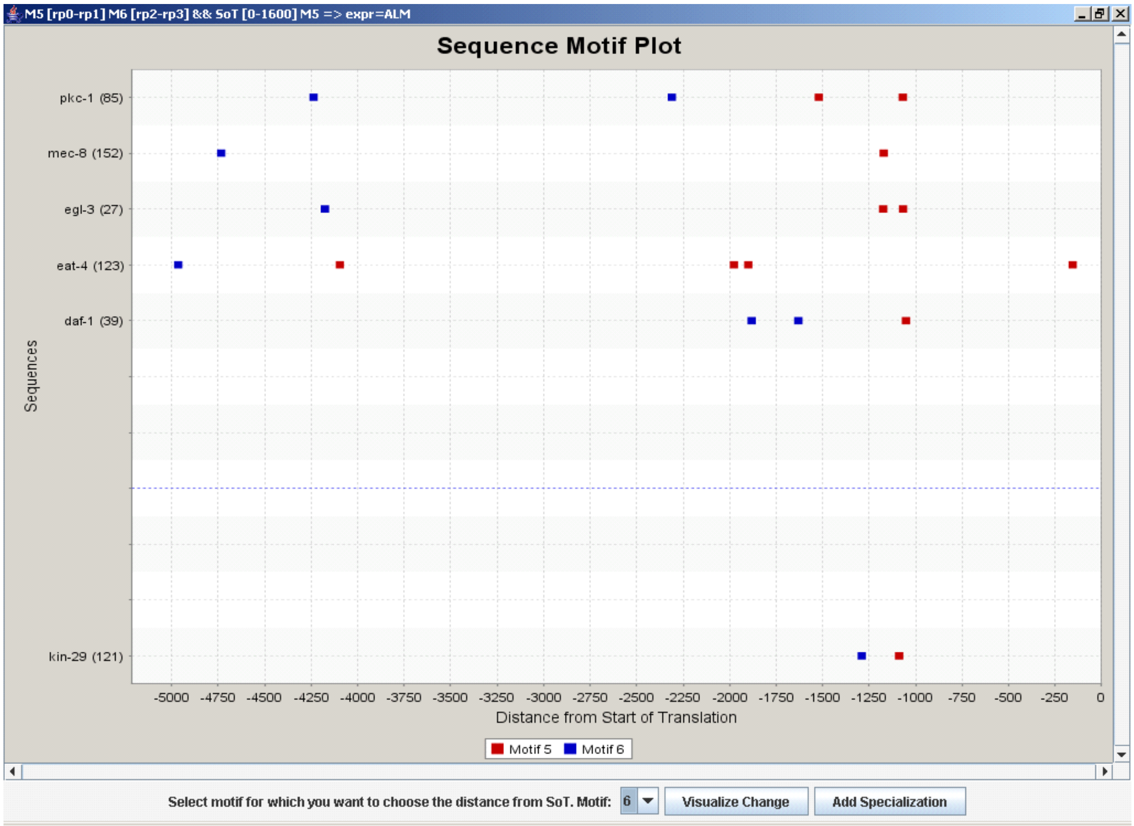
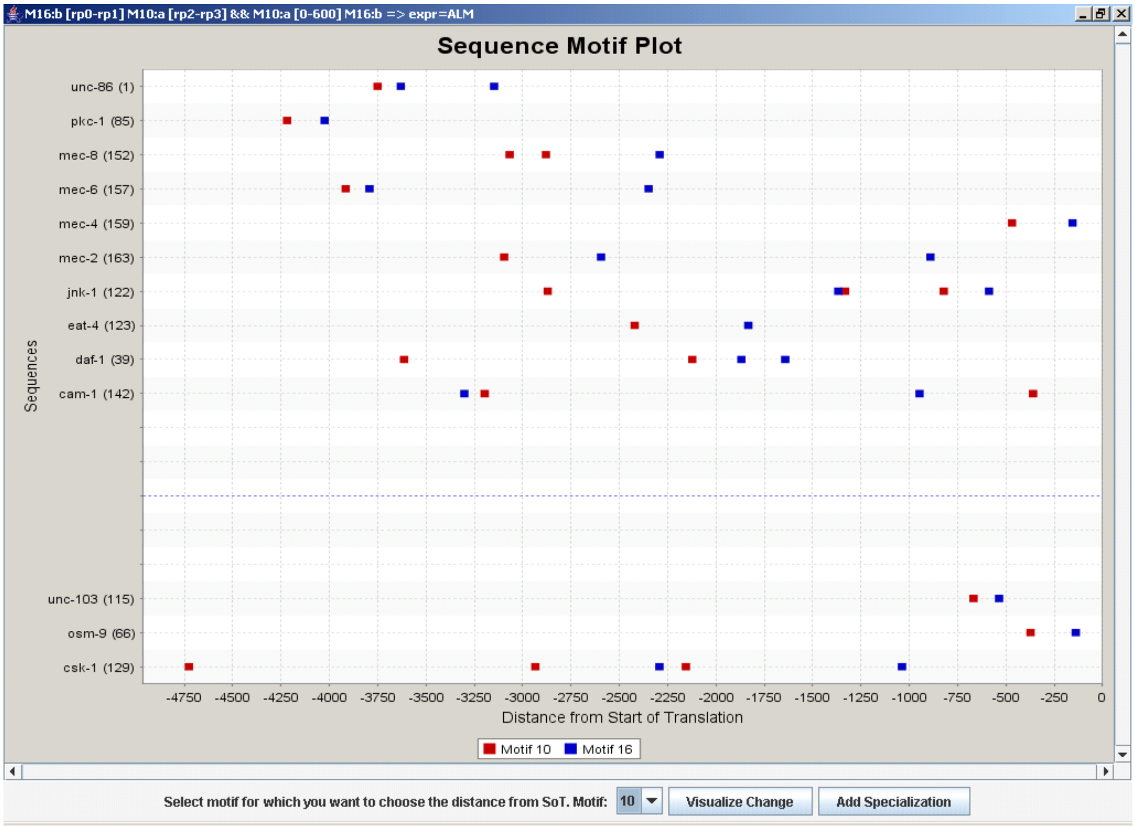


Figure 4.14: Hybrid rules help specify multiple constraints (based on different hypotheses) within a single specialization.

of Motif 5 that satisfies the order of occurrence condition is the same M5(instance) that lies within 1600 base pairs of the SoT. However, there might be a need for the user to actually specify constraints which are inter-related, aliases are supported by our rule grammar for exactly this reason.

4.2.7 Aliases

Aliases were included in the grammar to provide the user with an option of defining inter-related constraints or specialization terms. Consider the following Inter-Motif



M16:b [rp0-rp1] M10:a [rp2-rp3] && M10:a [0-600] M16:b => expr=ALM

Select motif for which you want to choose the distance from SoT. Motif: 10

Figure 4.15: Aliases let user define specializations with inter-related constraints.

Distance based specialization from Figure 4.5:

$$M10[0-500]M16 \Rightarrow expr = ALM \tag{4.5}$$

Visualizing this specialization using a sequence plot (Figure 4.6), one can see distinctly that not only do motifs M10 (red dot) and M16(blue dot) occur close together but they also occur in a pattern such that the same instances of M10 and M16 that are involved in the distance-based relationship also occur in the same order relative to the SoT. Aliases enable the user to specify such complex relationships in the rule as follows (Figure 4.15):

$$\text{M10:a [0-500] M16:b \&\& M16:b [rp0-rp1] M10:a [rp2-rp3]} \;\Rightarrow\; \text{expr=ALM} \tag{4.6}$$

For details of system operation, see User Guide.

Chapter 5

System Architecture

This tool was conceptualized as an extension to the WPI-Weka system and hence is also referred to as the *"Visualization and Specialization Modules(VSM)"*.This chapter describes the interaction of this tool with the WPI-Weka system and a high level design overview.

5.1 Component Interaction

Figure 5.1 illustrates how VSM interacts with the WPI-Weka system, to let the user perform hypothesis-driven exploratory analysis of genetic data in order to create specialized association rules that predict gene expression. The enumerated arrows with italicized text denote the different steps that constitute the process of exploratory analysis to discover specializations predicting gene expression. Each of the steps along with the inputs and outputs to the process are listed below:

1. Mine association rules. One of the contributions of [Rudss] was to integrate into a single classifier (AssociateClassifier) within the WPI-Weka system([WPI]), contributions of previous work at WPI in the field of association rule mining.

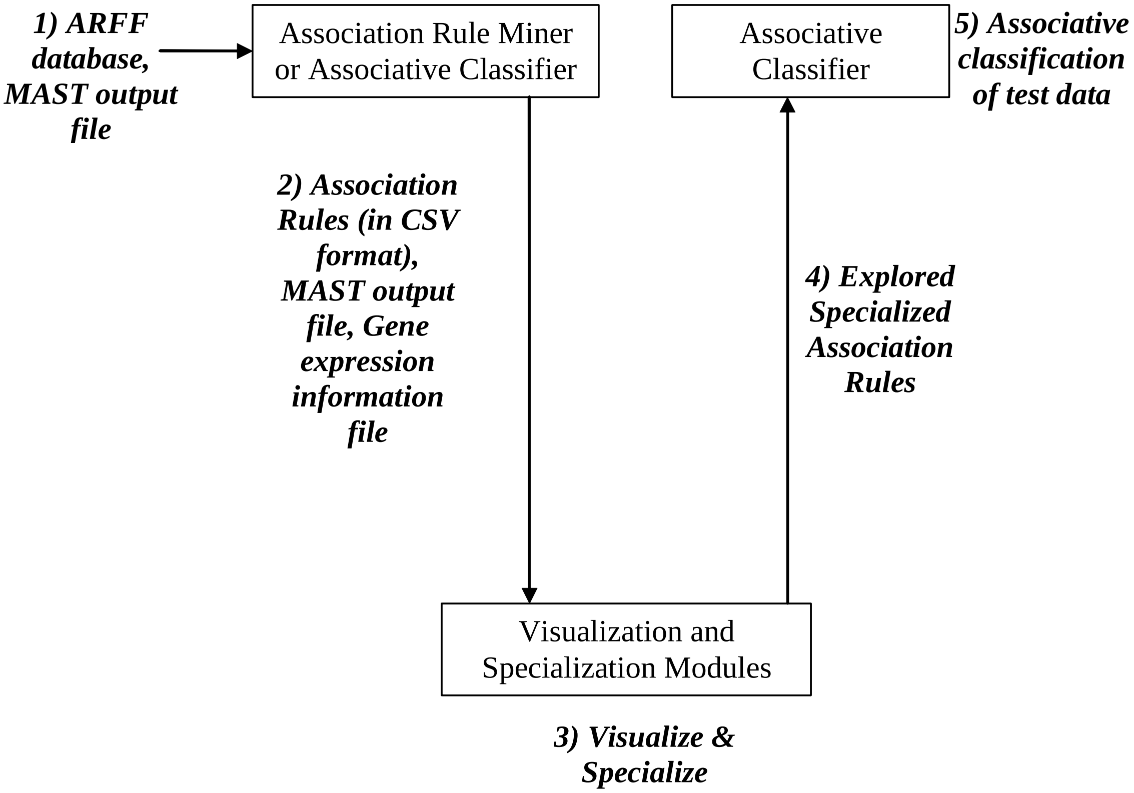

1) ARFF database, MAST output file
Association Rule Miner or Associative Classifier
Associative Classifier
5) Associative classification of test data
2) Association Rules (in CSV format), MAST output file, Gene expression information file
4) Explored Specialized Association Rules
Visualization and Specialization Modules
3) Visualize & Specialize

Figure 5.1: Process depicting the flow of data (interactions) between the WPI-Weka system and the VSM.
The data exchange between the WPI-Weka system and the VSM is in a comma separated value format which is explained in detail in Figure 5.2.

This classifier or the association rule miner (ARMiner) module of the WPI-Weka system, takes as input an ARFF file consisting of genetic data and a MAST output file (Section 3.3) and mines for association rules predicting gene expression.

2. Transfer mined association rules to VSM. As a part of the integration plan it was decided that both [Rudss] and this work would support import/export of rules in a predefined comma-separated values (CSV) format. So as a first step to analyzing gene expression association rules, [Rudss] could be used to invoke the VSM with the set of mined rules (Figure 5.2), the MAST output file and

CSV interface to transfer rule sets between WPI-Weka and VSM

```
HEADER
Id,Antecedent,Consequent,Confidence,Support,Lift,p-Value,Event Wt.,Within Cell-Type(s) support

DATA
001, M17, expr=ALM, 0.48148146, 0.325, 1.2037036, 4.9873279E-1, 0, 0.42857143
002, M12, expr=ALM, 0.52830184, 0.35, 1.3207545, 5.021099E-1, 0, 0.10714286
```

Figure 5.2: CSV interface to transfer rule sets between WPI-Weka system and VSM. The header section defines a comma separated list of the headers describing the attributes contained in each data instance. The Data section consists of the actual set of rules being exported or imported.

the gene expression information.

3. Visualize and Specialize. The implementation of this work, that is the VSM, would let the user perform hypothesis-driven visualization of data that helps specialize a set of association rules. See Section 4.2.

4. Transfer the set of specialized rules to the AssociativeClassifier. Again the VSM as well as [Rudss] support the transfer of set of specialized rules explored using the VSM to the AssociativeClassifier using the predefined CSV format.

5. Use the classifier to predict gene expression for a novel set of genes. The classifier can be used to calculate the classification accuracy of the imported specializations over a test set consisting of novel genes (i.e., genes that have not been used during mining or specialization) in order to determine the predictive power of the relationships identified.

5.2 System Design

This section describes the high level design of the VSM tool from a functional perspective that is also illustrated in Figure 5.3. We describe the different modules within the VSM subsystem implementation:

1. **MAST Parser**. The VSM could be invoked from within the WPI-Weka System or as a standalone application. Irrespective of the invocation mechanism (i.e., as a standalone application or from within WPI-Weka) it requires as input the gene expression information (Figure 4.2) and a MAST output file (Figure 3.5). The information from the MAST output file and the gene expression information is parsed using this module to populate a multi-level internal data structure (Figure 5.4) and subsequently the analysis frame (Figure 4.1) is displayed.

2. **Rule Parser**. Each base rule, either belonging to the set of rules mined using the WPI-Weka system or by using the "Add Rule" (Section 4.2.5) command from the analysis frame is parsed by using this module, that is, an implementation of the grammar defined in Figure 4.12.

3. Charting Extensions. The parsed information from rule parser and the internal data structure are used as input to the charting infrastructure to generate hypothesis-specific plots for data visualization. The charting infrastructure is created by extending the open source charting library JFreeChart. JFreeChart separates its data from its presentation layers which means the system could be extended to add new features to plots with minimal changes. The user's exploratory analysis using these plots (data visualization) subsequently produces specialized rules which in combination to the internal data structure can be

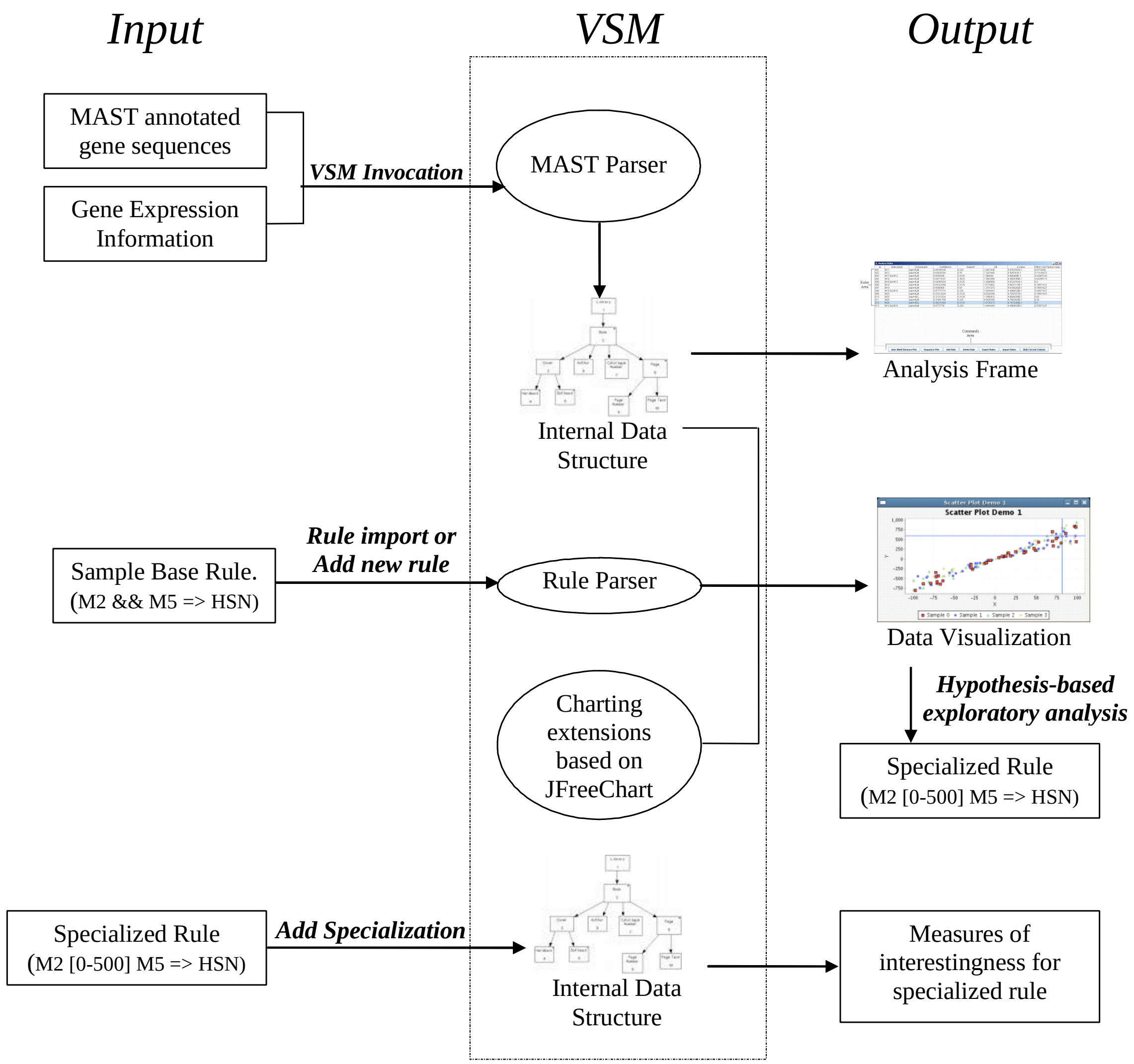


Figure 5.3: VSM System Design.
Functional design of the VSM. Each step in the exploratory analysis process is described in terms of the input(s) to the VSM subsystem, corresponding modules within the VSM invoked, and the output. The vertical rectangle represents the VSM subsystem with each oval representing a module within the VSM subsystem implementation. As delineated, the entities to the left of the VSM subsystem are the input(s) provided to each module and the entities to the right are the output(s) from the VSM at each step.

Key (Gene Name)	Value			
ceh-43	Expression Pattern	Neural		
	Gene Name	ceh-43		
	Length	993		
	Motif-based location info	M1	250, 523	
		M2	944	
nlp-5	Expression Pattern	Muscle		
	Gene Name	nlp-5		
	Length	5000		
	Motif-based location info	M1	3555	
		M2	45, 856	

Figure 5.4: Hierarchical internal data structure - Hash of hashes.
The top level hash has the gene name as the key attribute. The value element corresponding to each gene is also a hash consisting of sequence information which, besides gene specific information like length and expression pattern, contains location information organized on a per motif basis. The location information for each motif (key) consists of all occurrences of the motif on the gene sequence relative to the SoT. The sample here illustrates the organization of information for two sample sequences, ceh-43 and nlp-5, each of which contains one or more occurrences of motif M1 and M2.

used to calculate the different measures of interestingness for each specialized rule.

5.3 Implementation Details

Since the WPI-Weka system, which is an extensive collaborative effort within the KDD Research Group at WPI, is Java based, Java was chosen to implement VSM.

It was during the course of this project and [Rudss] that the WPI-Weka system was hosted in the WPI sourceforge server. As a part of this project an ant-script (Java equivalent of a makefile) was also developed that enables users to build the system from the source with minimal instructions. Eclipse was our choice of the IDE used as it provides a easy to use interface for both Java based development as well as CVS based version control.

Elements of good design were extensively applied throughout the development process. For instance, the support for import/export of CSV file containing the rules is implemented as a Java interface called the Analyzable interface which makes it easy for a new implementation to support a certain functionality without restricting a specific type of implementation. For instance if a new mining algorithm that mines gene expression rules is added to the WPI-Weka System, it could invoke VSM as long as it implements the Analyzable interface.

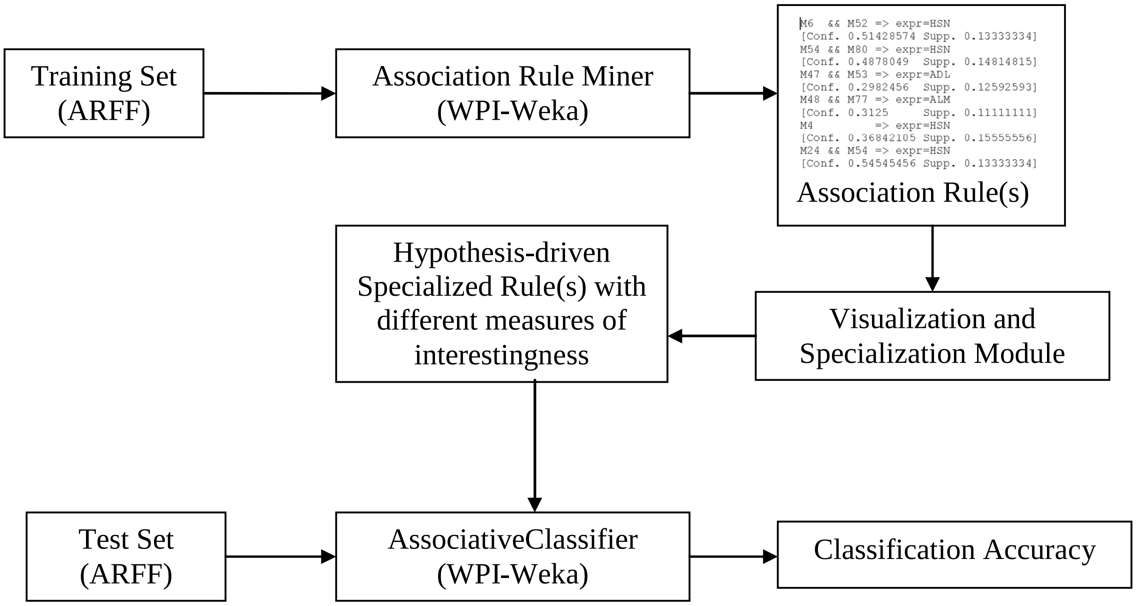
Chapter 6

Experimental Evaluation

Chapter 3 (Figure 3.2) describes in detail the process to create a dataset in a mining compatible format (ARFF) starting with data collection. As mentioned in Section 3.2 and again in Section 3.4, to avoid mining and exploring relationships that over fit the data, the sequences were divided into a training set (90% of the sequences) and a test set (10% of the sequences). Figure 6.1 illustrates the experimental setup.

6.1 Experimental Protocol and Parameters to be measured

Mining the training dataset using the WPI-Weka Association Rule Miner produces gene expression association rules. Different measures of interestingness, including the traditional measures of support and confidence, are calculated to estimate the statistical significance of the rules. Once the domain user identifies an interesting base rule (e.g., based on the values of different measures of interestingness), VSM could be used to perform hypothesis-driven exploratory analysis of the selected rule



```
M6  && M52 => expr=HSN
[Conf. 0.51428574 Supp. 0.13333334]
M54 && M80 => expr=HSN
[Conf. 0.4878049  Supp. 0.14814815]
M47 && M53 => expr=ADL
[Conf. 0.2982456  Supp. 0.12592593]
M48 && M77 => expr=ALM
[Conf. 0.3125     Supp. 0.11111111]
M4         => expr=HSN
[Conf. 0.36842105 Supp. 0.15555556]
M24 && M54 => expr=HSN
[Conf. 0.54545456 Supp. 0.13333334]
```

Figure 6.1: Experimental Setup.
Starting with mining of rules over the training data to the classification of the test dataset using the explored specialized rules.

to derive a specialized association rule such that the antecedent of the rule consists of additional constraints based on positional information of the motif(s). This process is repeated a few times with different base rules to derive a set of specialized rules that, at least statistically, seem to provide a more accurate representation of the underlying regulatory mechanism governing gene expression. This set of specialized rules is then tested for accuracy over a dataset consisting of novel genes (i.e., genes which were not used to elicit motifs and were also not involved in the mining or the specialization processes). The classification accuracy is a measure of predictive power. An improved classification accuracy as compared to the set of base rules translates to an increase in the confidence of the specialized rule with a reasonable decrease in the support of the rule and is used to estimate the potential biological validity of the relationship. Other measures of interestingness like the p-value and the lift also help to estimate the effectiveness of the rule. The system is extensible from the measures of interestingness perspective; that is, it would require minimal changes for another measure of interestingness to be added to the system.

6.2 Experimental Results and Analysis

The target audience for the tool developed are domain experts trying to identify biologically interesting relationships. We present a walk through of a small scale experiment using real genetic data that helps establish the work flow for the process of performing hypothesis-driven specializations.

As discussed in Section 3.1, the training set consists of 151 gene sequences in ARFF format. In the experiment reported here, these sequences were mined for presence based association rule mining with a minimum support of 0.1 and a minimum confidence of 0.3. It is important to understand the reason for choosing the relatively low values for support and confidence. The choice of the low value for the minimum support is based on the way the dataset is put together. Since we tried to identify gene sequences for 9 different cell types, each cell type has approximately 11% representation in the dataset, then the motif elicitation and annotation process were performed with an intent to find cell-type specific motifs. So if a biologically valid cell-type specific regulatory mechanism was found during the mining process it is reasonable to expect that it should have low support(i.e., 11% or less). Since we intend to explore specializations that represent the biological relationship more effectively as compared to the base rules from which they are derived, we can select base rules with low confidence measure. The mining process resulted in a total of 269 rules, out of which the following 6 rules were chosen to illustrate effectively all forms of hypothesis-driven specializations. So we intend to visualize and specialize 2 rules per hypothesis type to illustrate the work flow.

Motif	Consensus sequence
M6	GGAAGAAGAG
M47	GAGAAGAG
M48	TGAGAAAA
M52	GAAGAAGAAGAA
M53	GAAGAAGAAGGA
M54	GAGTGAGAGGGG
M69	GGGGGGGAGG
M77	GAGACGAAGA
M80	GAGAAGAAGAAG

Figure 6.2: List of motifs from the base rules under consideration along with their consensus sequences.

$$\text{M6 \&\& M52} \Rightarrow \text{expr=HSN [Conf. = 0.51428574, Supp. = 0.13333334]} \tag{6.1}$$

$$\text{M54 \&\& M80} \Rightarrow \text{expr=HSN [Conf. = 0.4878049, Supp. = 0.14814815]} \tag{6.2}$$

$$\text{M47 \&\& M53} \Rightarrow \text{expr=ADL [Conf. = 0.2982456, Supp. = 0.12592593]} \tag{6.3}$$

$$\text{M48 \&\& M77} \Rightarrow \text{expr=ALM [Conf. = 0.3125, Supp. = 0.11111111]} \tag{6.4}$$

$$\text{M69} \Rightarrow \text{expr=HSN [Conf. = 0.46341464, Supp. = 0.14074074]} \tag{6.5}$$

$$\text{M6} \Rightarrow \text{expr=HSN [Conf. = 0.13636364, Supp. = 0.044444446]} \tag{6.6}$$

6.2.1 Order of motif occurrence specialization of Rules (6.1) and (6.2)

In this section we specialize Rules (6.1) and (6.2) based on the hypothesis "The order of occurrence of motifs influences gene expression". Figures 6.3 and 6.5 depict the process followed to specialize the base rules from an order of occurrence of motifs perspective.

An inter-motif distance plot, for instance Figure 6.3, is split into two parts: the left part containing the sequences that support the rule and the right part containing the sequences that support the antecedent only. This provides the user with an easy mechanism to discover order of occurrence based patterns in the left part that are not as frequent in the right part, as this would let us explore specializations with improved classification accuracy(and/or confidence). For instance, in Figure 6.3, many genes on the right-hand side lack magenta (dark) dots. This implies that the occurrence of Motif 6 in between Motif 52 and the SoT (i.e., M52-M6-SoT) may positively influence a gene to be expressed in cells of type HSN. Once a user has identified such an order based relationship that is of interest, the following corresponding specialized rule can be added to the analysis frame:

$$\text{M6 [rp0-rp1] M52 [rp2-rp3]} \Rightarrow \text{expr=HSN [Conf.} = 0.6363636\text{, Supp.} = 0.1037037\text{]} \tag{6.7}$$

The analysis frame enables the user to estimate the significance of the rule using the different measures of interestingness computed for the specialized rule. The user can also compare the interestingness of the specialized rule with that of the base rule (Figure 6.4).

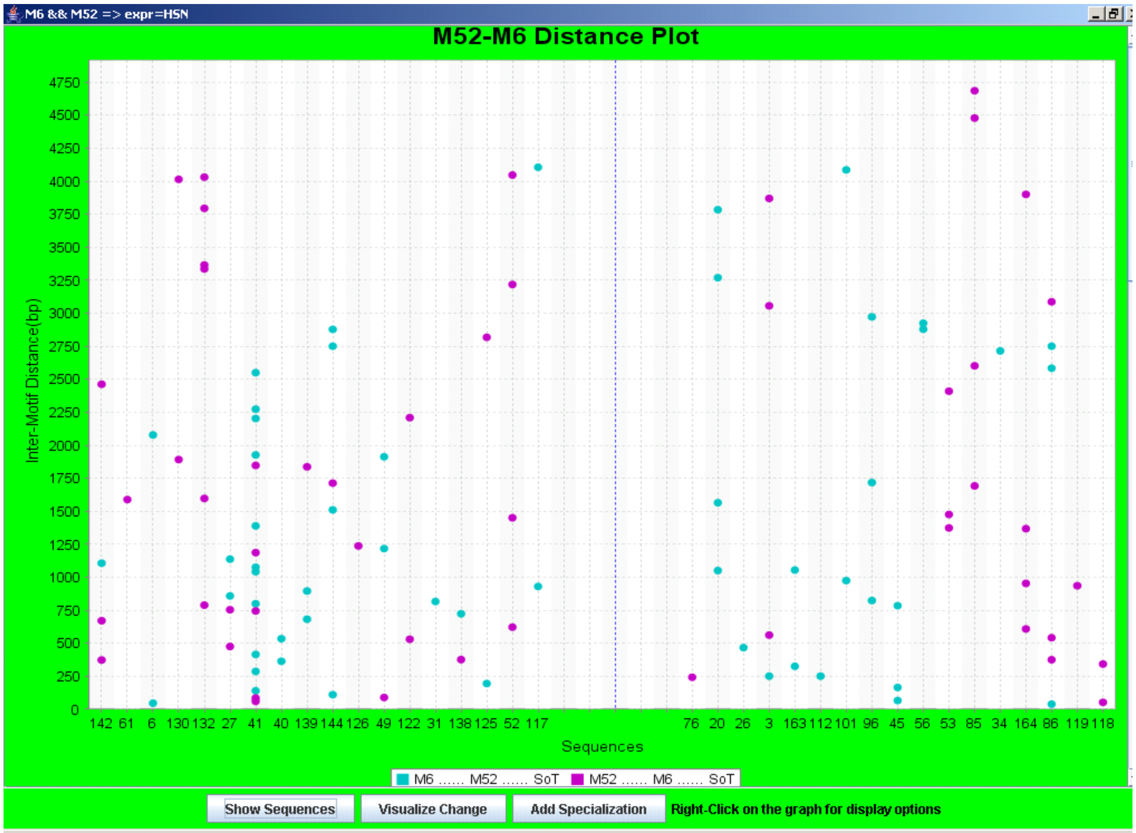
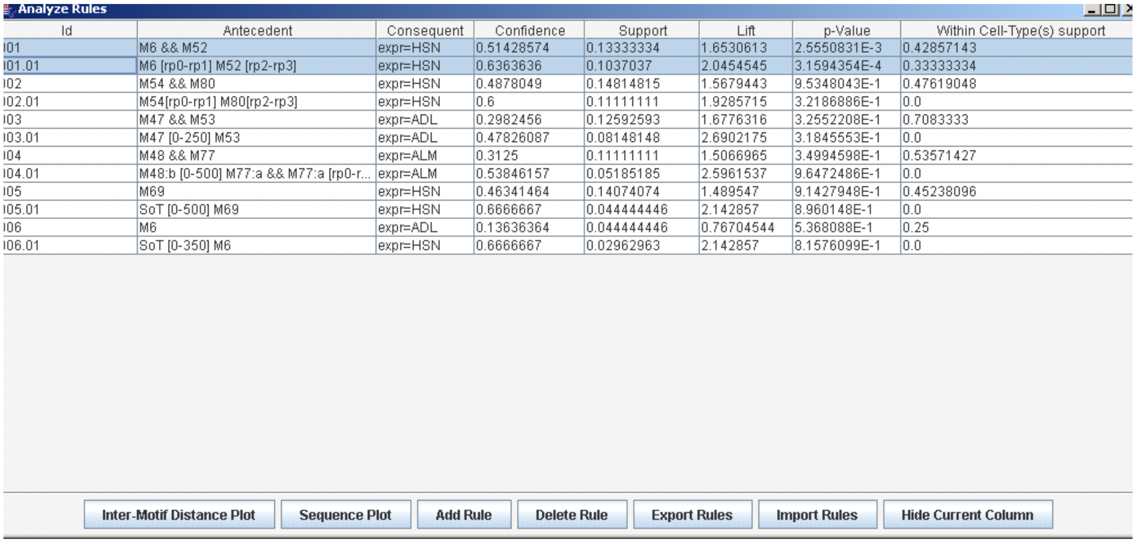


Figure 6.3: Inter-motif distance plot for the base rule *M6 && M52 ⇒ expr=HSN*. Along the x-axis are the Id's corresponding to the genes. Along the y-axis is the inter-motif distance. Each point in the plot, irrespective of the color, represents the distance between an occurrence of Motif 6 from an occurence of Motif 52. Also notice the division of the graph into two parts. Each colored dot is representative of the relative ordering of the occurrence of the motifs relative to the Start of Transcription. A magenta (dark) dot represents the inter-motif distance between an occurrence of Motif 6 and Motif 52 such that the occurrence of Motif 6 lies between the occurrence of Motif 52 and the SoT (i.e., M52-M6-SoT). A aqua (light) dot represents an inter-motif distance such that the order of occurrence of motifs is M6-M52-SoT. The left-side part consists of gene sequences which support the rule and the right-side part consists of the gene sequences that support the antecedent of the rule only.



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
01	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
01.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
02	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	9.5348043E-1	0.47619048
02.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	3.2186886E-1	0.0
03	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	3.2552208E-1	0.7083333
03.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.1845553E-1	0.0
04	M48 && M77	expr=ALM	0.3125	0.11111111	1.5066965	3.4994598E-1	0.53571427
04.01	M48:b [0-500] M77:a && M77:a [rp0-r...	expr=ALM	0.53846157	0.05185185	2.5961537	9.6472486E-1	0.0
05	M69	expr=HSN	0.46341464	0.14074074	1.489547	9.1427948E-1	0.45238096
05.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	8.960148E-1	0.0
06	M6	expr=ADL	0.13636364	0.044444446	0.76704544	5.368088E-1	0.25
06.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	8.1576099E-1	0.0

Figure 6.4: Analysis frame providing a comparison of the interestingness of the specialized rule, Rule (6.7), with the base rule, Rule (6.1).

Similarly in Figure 6.5, observe that many genes in the right part of the plot lack magenta (dark) points. This again indicates that the occurrence of motif 54 between an occurrence of Motif 80 and the SoT (i.e., SoT-M54-M80) may positively influence gene expression. Again the user could add the following corresponding specialization to the analysis frame and compare its interestingness to that of the base rule.

$$\text{M54[rp0-rp1] M80[rp2-rp3]} \Rightarrow \text{ expr=HSN [Conf. } = 0.6\text{, Supp. } = 0.11111] \quad (6.8)$$

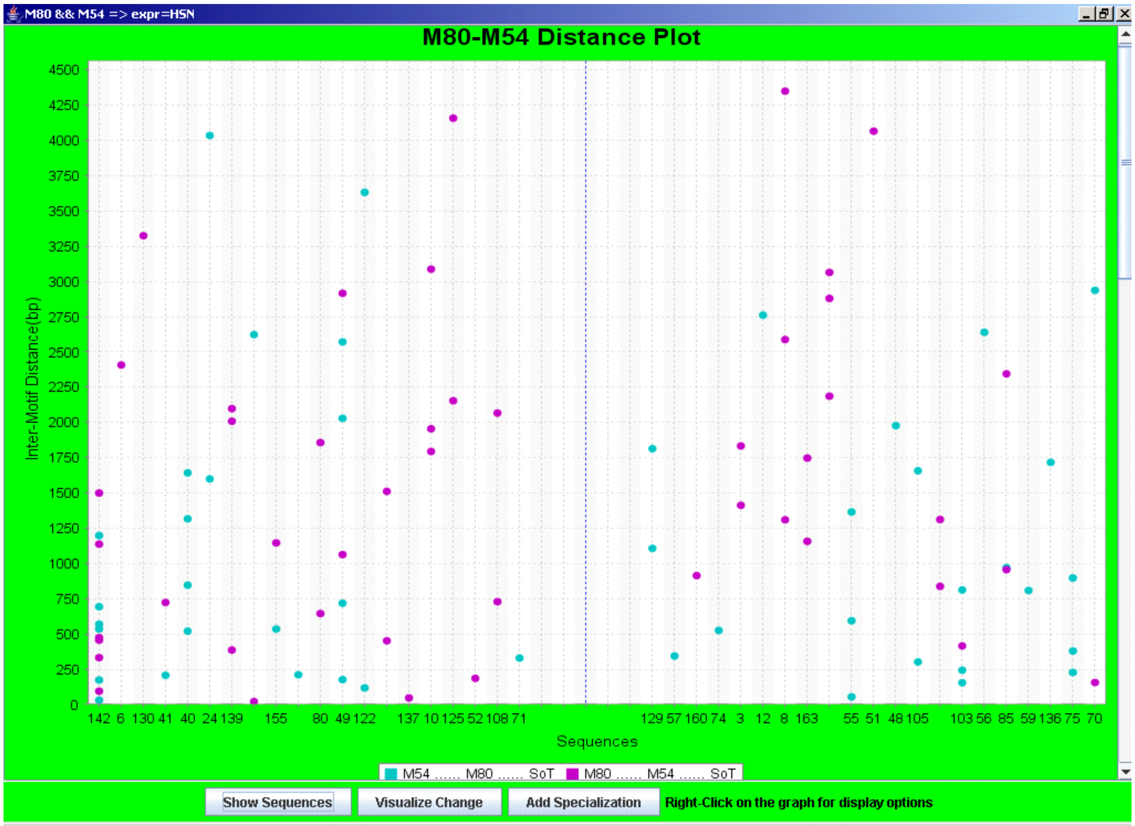
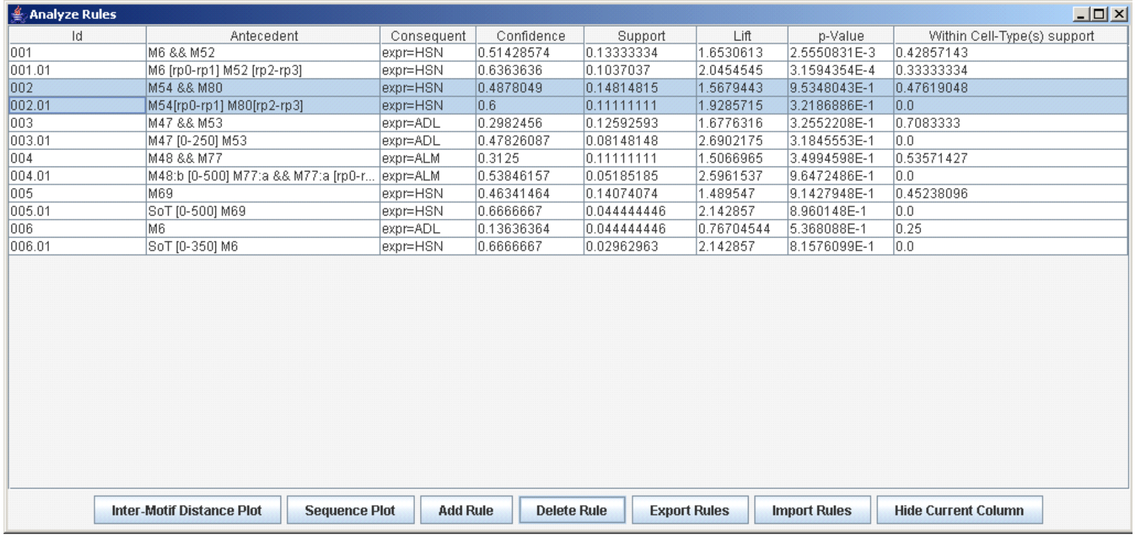


Figure 6.5: Inter-motif distance plot for the base rule *M54 && M80* ⇒ *expr=HSN*.



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
002	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	9.5348043E-1	0.47619048
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	3.2186886E-1	0.0
003	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	3.2552208E-1	0.7083333
003.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.1845553E-1	0.0
004	M48 && M77	expr=ALM	0.3125	0.11111111	1.5066965	3.4994598E-1	0.53571427
004.01	M48:b [0-500] M77:a && M77:a [rp0-r...	expr=ALM	0.53846157	0.05185185	2.5961537	9.6472486E-1	0.0
005	M69	expr=HSN	0.46341464	0.14074074	1.489547	9.1427948E-1	0.45238096
005.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	8.960148E-1	0.0
006	M6	expr=ADL	0.13636364	0.044444446	0.76704544	5.368088E-1	0.25
006.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	8.1576099E-1	0.0

Figure 6.6: Analysis frame providing a comparison of the interestingness of the specialized rule, Rule (6.8), with the base rule, Rule (6.2).

6.2.2 Inter-motif distance specialization of Rules (6.3) and (6.4)

In this section we specialize Rules (6.3) and (6.4) based on the hypothesis "Inter-motif distance influences gene expression". Figures 6.7 and 6.9 depict the process followed to specialize the base rules from an inter-motif distance perspective.

An inter-motif distance plot, for instance Figure 6.7, displays for qualifying gene the inter-motif distances between each occurrence of Motif 47 from each occurrence of Motif 53. A qualifying gene is one that contains at least one occurrence of each motif in the base rule. The splitting of the plot into two parts (as described in Section 6.2.1), provides the user with an easy mechanism to discover inter-motif distance pattern in the left part that are not as frequent in the right part, as this would let us explore specializations with improved classification accuracy(and/or confidence). For instance, in Figure 6.7, many genes on the right-side part lack dots in the shaded area of the plot, that is, an inter-motif distance range of 0-250 bp which implies that the occurrence of motifs 47 and 53 within 250 bp from each other may be positively related to the gene being expressed in cell type ADL. Once a user has defined an inter-motif distance relationship that is of interest the relevant region of the plot gets shaded as illustrated in Figure 6.7. Subsequently the following corresponding specialization can be added to the analysis frame.

$$\text{M47 [0-250] M53} \Rightarrow \text{ expr=ADL [Conf. = 0.47826087, Supp. = 0.08148148]} \quad (6.9)$$

Again the user can estimate the significance of the rule by using the different measures of interestingness computed for the specialized rule. The user can also compare the interestingness of the specialized rule as compared with that of the base rule (Figure 6.8).

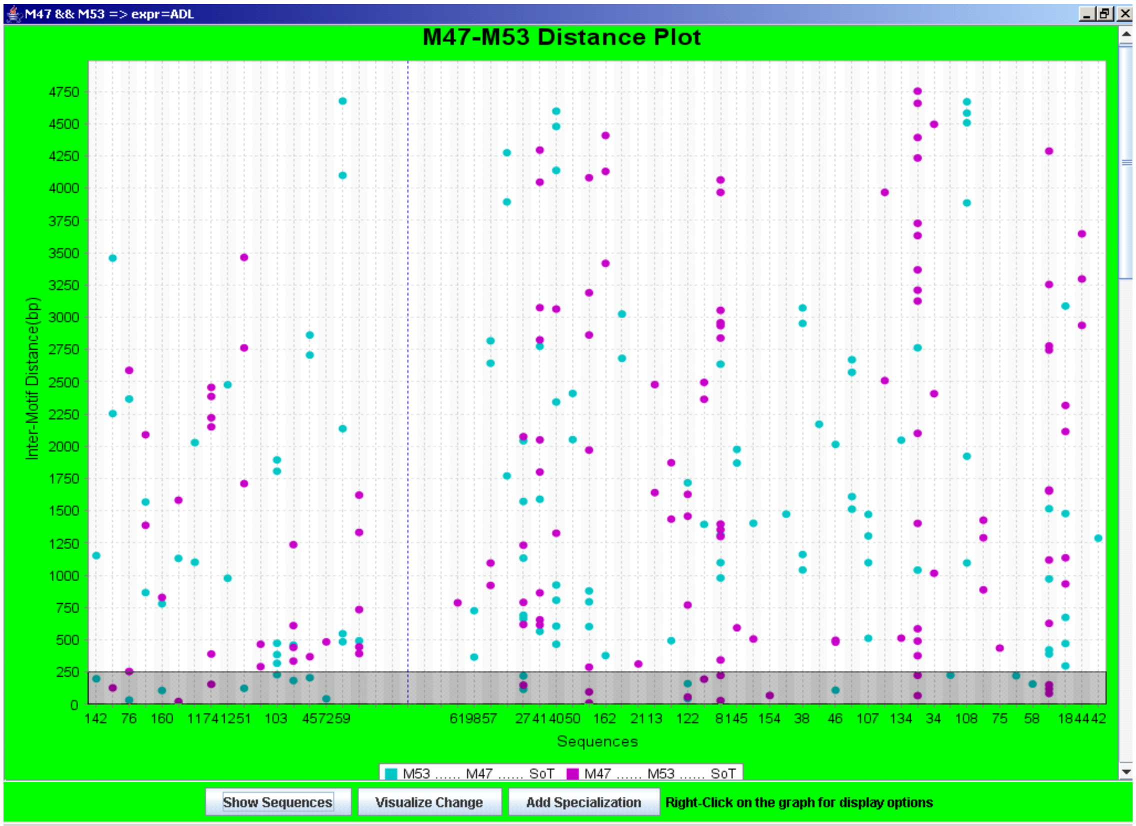
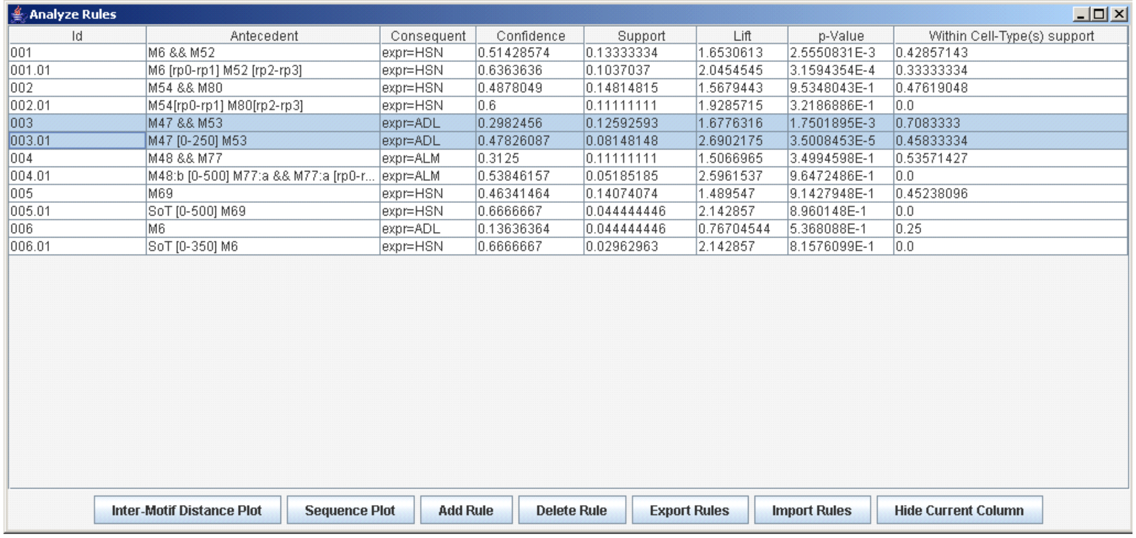


Figure 6.7: Inter-motif distance plot for the base rule *M47 && M53* ⇒ *expr=ADL*.



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
002	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	9.5348043E-1	0.47619048
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	3.2186886E-1	0.0
003	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	1.7501895E-3	0.7083333
003.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.5008453E-5	0.45833334
004	M48 && M77	expr=ALM	0.3125	0.11111111	1.5066965	3.4994598E-1	0.53571427
004.01	M48:b [0-500] M77:a && M77:a [rp0-r...	expr=ALM	0.53846157	0.05185185	2.5961537	9.6472486E-1	0.0
005	M69	expr=HSN	0.46341464	0.14074074	1.489547	9.1427948E-1	0.45238096
005.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	8.960148E-1	0.0
006	M6	expr=ADL	0.13636364	0.044444446	0.76704544	5.368088E-1	0.25
006.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	8.1576099E-1	0.0

Figure 6.8: Analysis frame providing a comparison of the interestingness of the specialized rule, Rule (6.9), with the base rule, Rule (6.3).

Similarly in Figure 6.9, observe that many genes lack magenta (dark) points in the right-side part of the shaded area of the plot, that is, an inter-motif distance range of 0-500 bp. This implies that the occurrence of motif 77 and 48 within 250 bp from each other and in the order SoT-M77-M48 may positively influence gene expression in cell type ALM. Subsequent addition of the following corresponding specialization to the analysis frame enables the user to compare the interestingness of the specialized rule to that of the base rule (Figure 6.10).

$$\text{M48:b [0-500] M77:a \&\& M77:a [rp0-rp1] M48:b[rp2-rp3]} \Rightarrow \text{ expr=ALM} \\ \text{[Conf. = 0.53846157, Supp. = 0.05185185]} \quad (6.10)$$

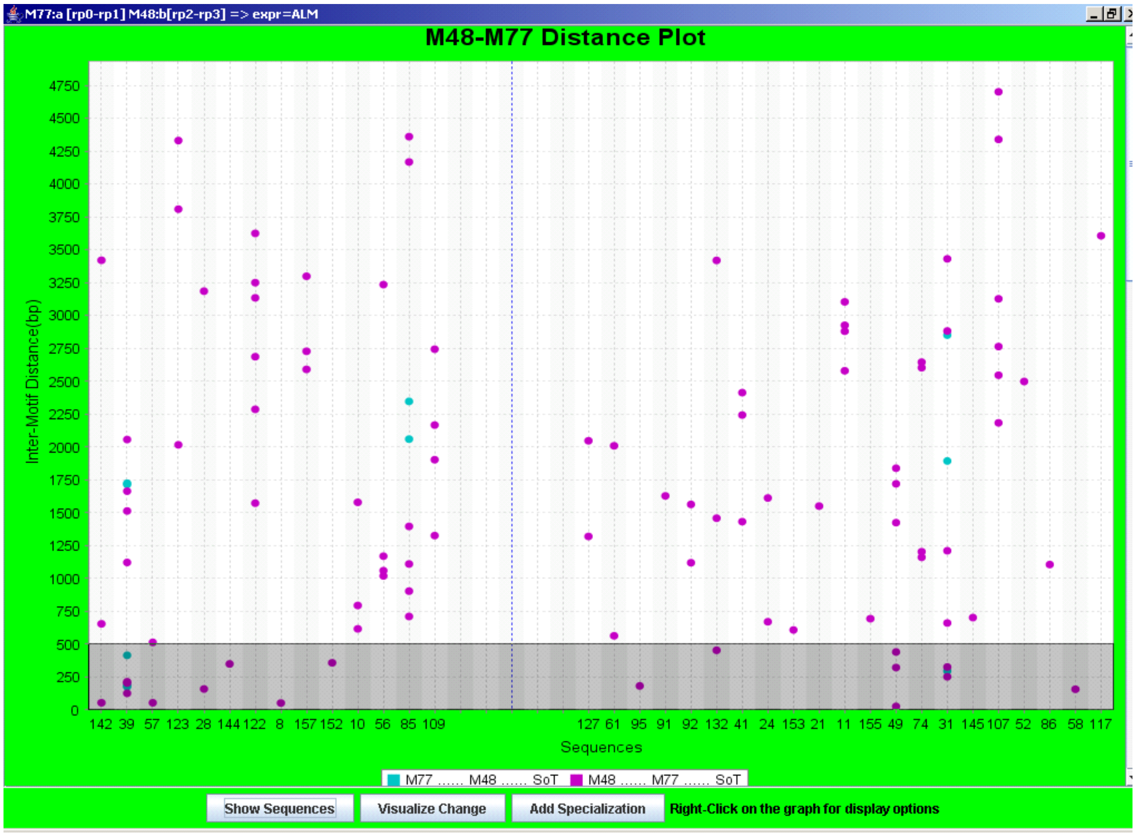
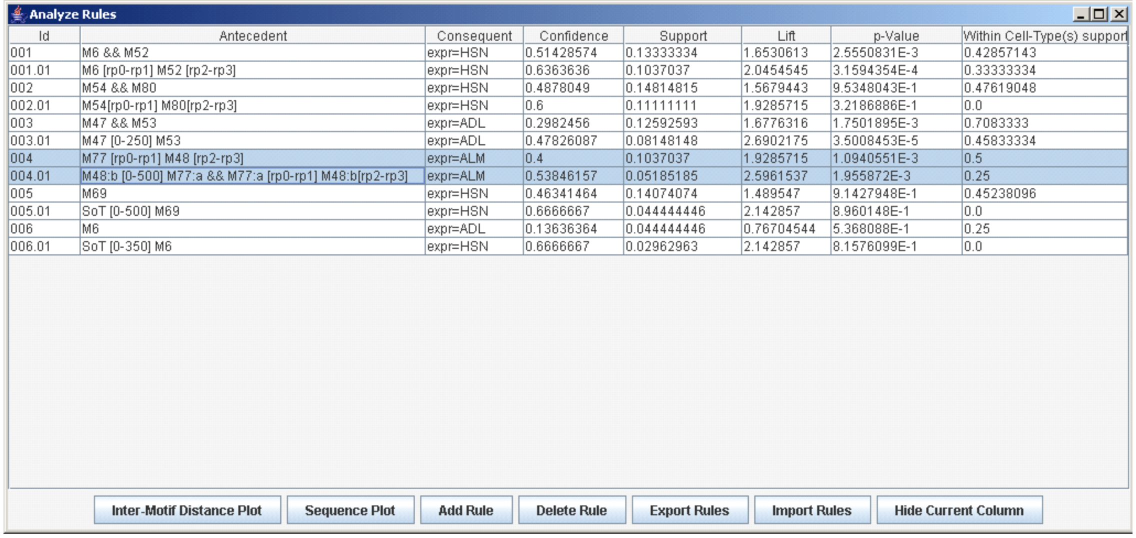


Figure 6.9: Inter-motif distance plot for the base rule *M77 [rp0-rp1] M48[rp2-rp3]* ⇒ *expr=ALM.*
Notice that the base rule selected is already an *"order of occurrence of motifs"* specialization. This was one of the rules generated by the ASAS algorithm as mentioned in Section 4.2.4.



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
002	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	9.5348043E-1	0.47619048
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	3.2186886E-1	0.0
003	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	1.7501895E-3	0.7083333
003.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.5008453E-5	0.45833334
004	M77 [rp0-rp1] M48 [rp2-rp3]	expr=ALM	0.4	0.1037037	1.9285715	1.0940551E-3	0.5
004.01	M48:b [0-500] M77:a && M77:a [rp0-rp1] M48:b[rp2-rp3]	expr=ALM	0.53846157	0.05185185	2.5961537	1.955872E-3	0.25
005	M69	expr=HSN	0.46341464	0.14074074	1.489547	9.1427948E-1	0.45238096
005.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	8.960148E-1	0.0
006	M6	expr=ADL	0.13636364	0.044444446	0.76704544	5.368088E-1	0.25
006.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	8.1576099E-1	0.0

Figure 6.10: Analysis frame providing a comparison of the interestingness of the specialized rule, Rule (6.10), with the base rule, Rule (6.4).

6.2.3 Distance from SoT specialization of Rules (6.5) and (6.6)

In this section we specialize Rules (6.5) and (6.6) based on the hypothesis "Distance of motif occurrence from SoT influences gene expression". Figures 6.11 and 6.13 depict the process followed to specialize the base rules from a distance from SoT perspective.

A sequence plot, for instance Figure 6.11, displays qualifying genes with the motifs overlaid such that a point on the plot represents the distance of that occurrence of the motif from the SoT. A qualifying gene is one that contains at least one occurrence of each motif in the base rule. The color of the point in the plot is indicative of the motif in question. The splitting of the plot into two parts (as described in Section 6.2.1), provides the user with an easy mechanism to discover distance from SoT based pattern in the top part that are not as frequent in the bottom part, as this would let us explore specializations with improved classification accuracy(and/or confidence). For instance, in Figure 6.11, many genes in the bottom part lack dots in the area delineated by the SoT (far right end of the plot) and the vertical line representing the 500 bp from SoT cutoff. This indicates that an occurrence of Motif 69 within a distance of 500 bp from the SoT may positively influence gene expression in cell type ADL. Once a user has defined such a "Distance from SoT" relationship it is indicated in the plot as the vertical line, in the same color as the one reserved for the motif, as illustrated in Figure 6.11 is added to the plot. Subsequently the following corresponding specialization can be added to the analysis frame.

$$\text{SoT [0-500] M69} \Rightarrow \text{ expr=HSN [Conf. = 0.6666667, Supp. = 0.044444446]} \tag{6.11}$$

Again the analysis frame with the specialized rule enables the user to compare the interestingness of the specialized rule with that of the base rule (Figure 6.12).

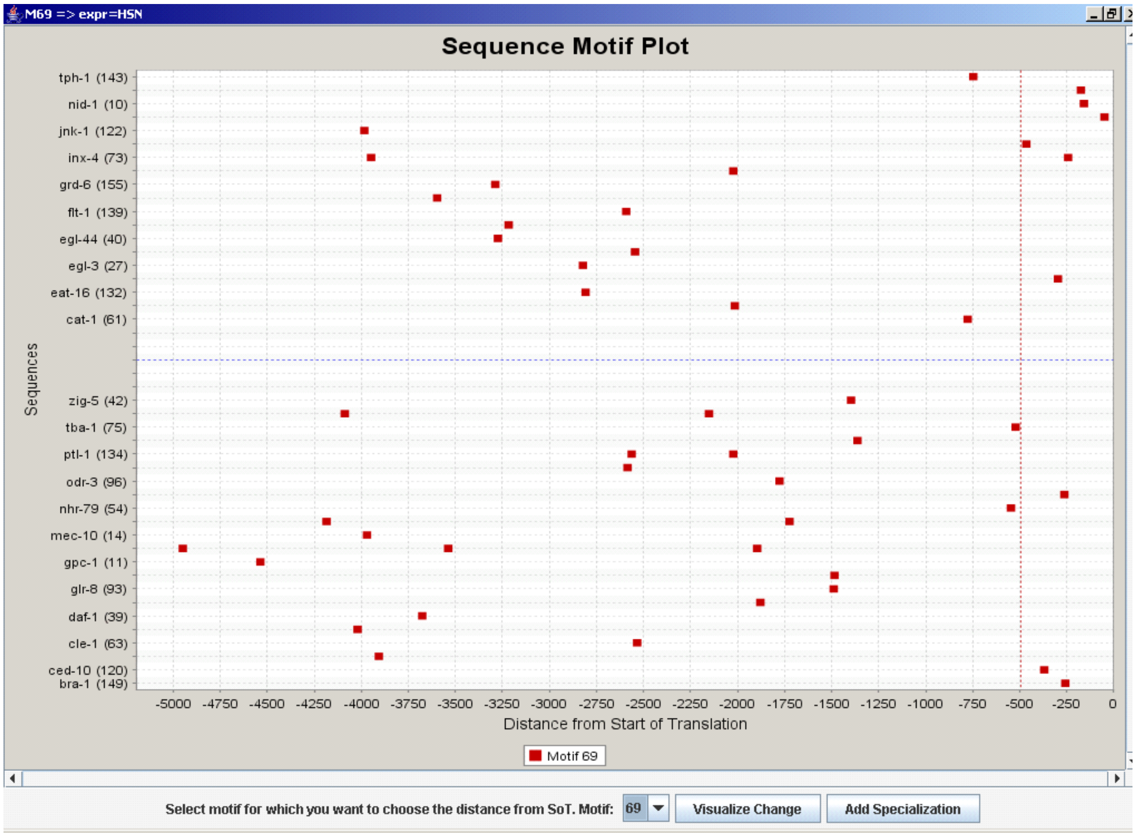
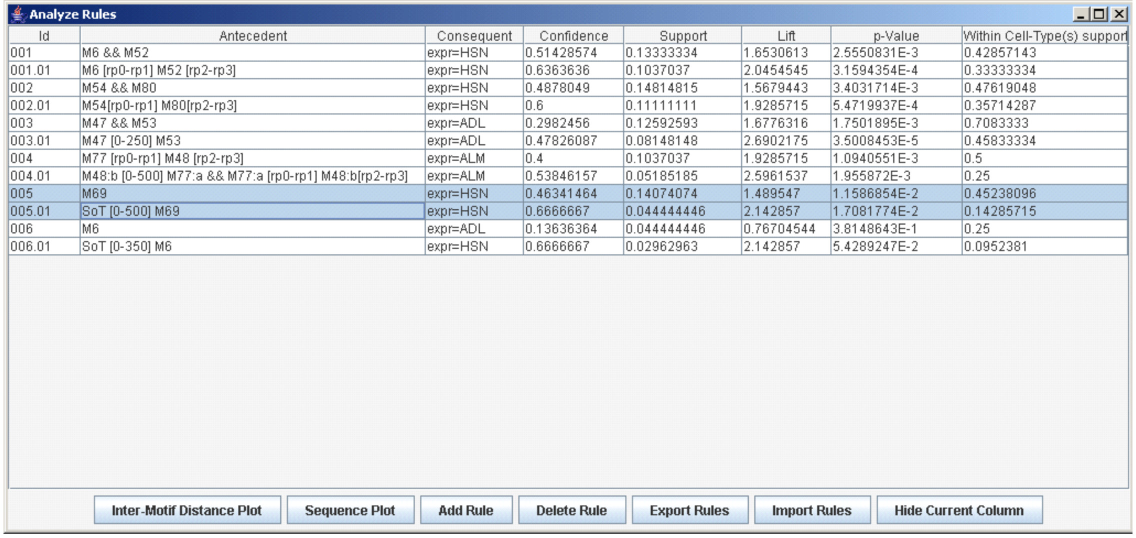


Figure 6.11: Sequence plot for the base rule *M69* ⇒ *expr=HSN*.
The sequence plot lists the genes along the y-axis and the distance from SoT along the x-axis. Each colored dot represents an occurrence of a specific motif. Notice the division of the graph into two parts using a horizontal line through the plot. The top part consists of gene sequences which support the rule and the bottom part consists of the gene sequences that support the antecedent of the rule. A user introduced vertical line marks the 500 bp distance from SoT.



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
002	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	3.4031714E-3	0.47619048
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	5.4719937E-4	0.35714287
003	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	1.7501895E-3	0.7083333
003.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.5008453E-5	0.45833334
004	M77 [rp0-rp1] M48 [rp2-rp3]	expr=ALM	0.4	0.1037037	1.9285715	1.0940551E-3	0.5
004.01	M48:b [0-500] M77:a && M77:a [rp0-rp1] M48:b[rp2-rp3]	expr=ALM	0.53846157	0.05185185	2.5961537	1.955872E-3	0.25
005	M69	expr=HSN	0.46341464	0.14074074	1.489547	1.1586854E-2	0.45238096
005.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	1.7081774E-2	0.14285715
006	M6	expr=ADL	0.13636364	0.044444446	0.76704544	3.8148643E-1	0.25
006.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	5.4289247E-2	0.0952381

Figure 6.12: Analysis frame providing a comparison of the interestingness of the specialized rule, Rule (6.11), with the base rule, Rule (6.5).

Similarly in Figure 6.13, observe that many genes lack points in the top part in the 0-350 bp region of the plot, that is, a distance from SoT of 0-350 bp. This implies that the occurrence of motif 6 within 350 bp from the SoT may positively influence gene expression in cell HSN. Subsequent addition of the following corresponding specialization to the analysis frame enables the user to compare the interestingness of the rule to that of the base rule (Figure 6.14).

$$\text{SoT [0-350] M6} \Rightarrow \text{ expr=HSN[Conf. } = 0.6666667\text{, Supp. } = 0.02962963] \quad (6.12)$$

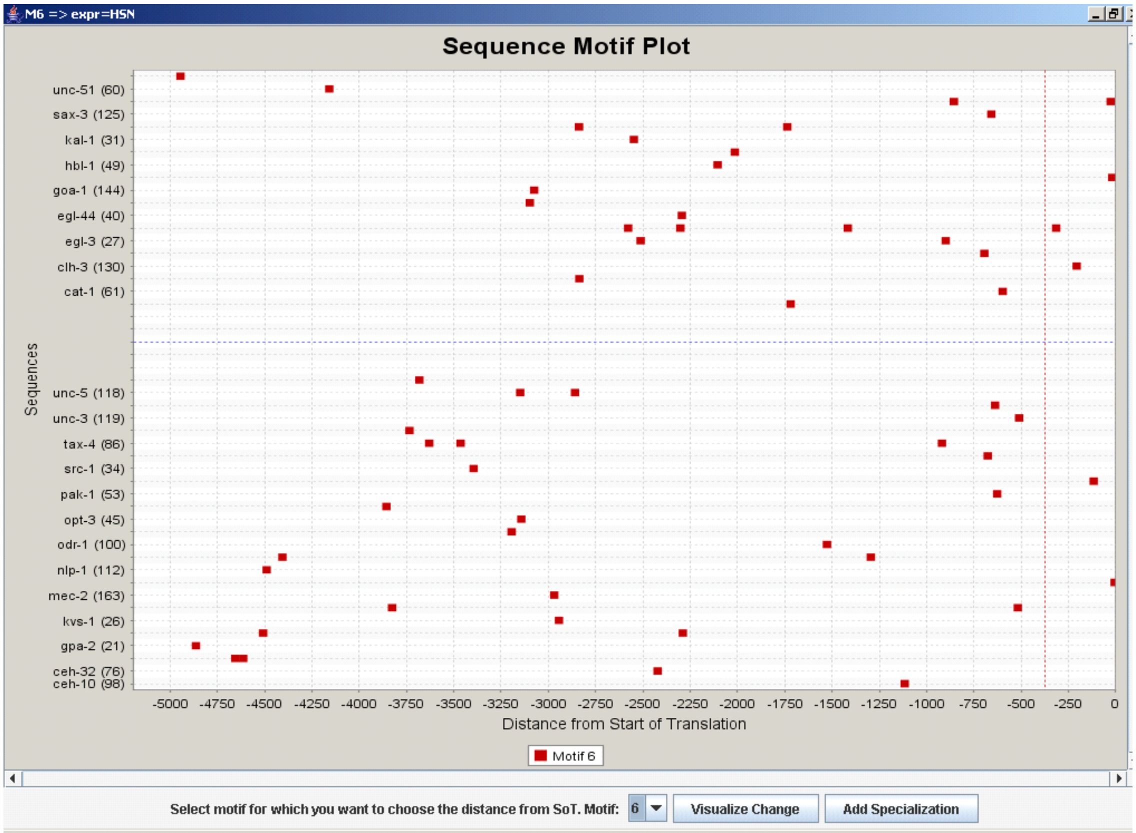
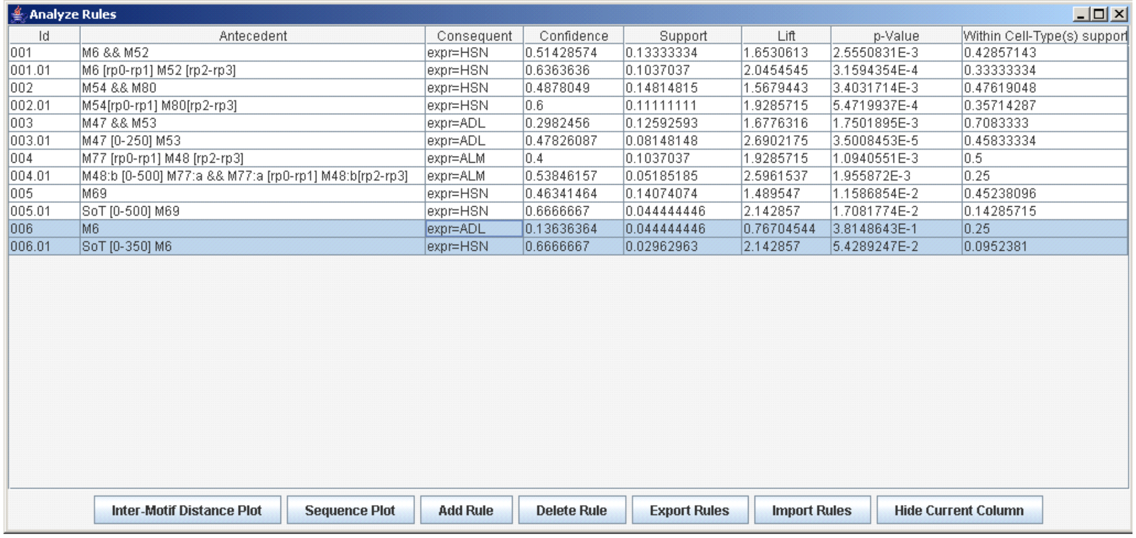


Figure 6.13: Sequence plot for the base rule *M6* ⇒ *expr=HSN*.



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
002	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	3.4031714E-3	0.47619048
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	5.4719937E-4	0.35714287
003	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	1.7501895E-3	0.7083333
003.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.5008453E-5	0.45833334
004	M77 [rp0-rp1] M48 [rp2-rp3]	expr=ALM	0.4	0.1037037	1.9285715	1.0940551E-3	0.5
004.01	M48:b [0-500] M77:a && M77:a [rp0-rp1] M48:b[rp2-rp3]	expr=ALM	0.53846157	0.05185185	2.5961537	1.955872E-3	0.25
005	M69	expr=HSN	0.46341464	0.14074074	1.489547	1.1586854E-2	0.45238096
005.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	1.7081774E-2	0.14285715
006	M6	expr=ADL	0.13636364	0.044444446	0.76704544	3.8148643E-1	0.25
006.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	5.4289247E-2	0.0952381

Figure 6.14: Analysis frame providing a comparison of the interestingness of the specialized rule, Rule (6.12), with the base rule, Rule (6.6).

6.2.4 Experimental results

As mentioned all along the analysis phase, an important feature which aids the exploratory analysis is that as soon as a specialization is added to the analysis frame using any of the visualization plots the different measures of interestingness of the newly added specialization are calculated. Thus, the domain user can instantly observe whether the specialized rule is better or worse on a specific metric and by how much. For instance, Figure 6.15 lists all the base rules and their corresponding specializations as they would appear in the analysis frame with the different measures of interestingness computed over the **training data**. At this point the user could save the rule set using the 'Export Rules' option from the analysis frame. This saved rule set can later be imported into the 'AssociativeClassifier' within the WPI-Weka system to test the classification accuracy over a set of novel gene sequences (test dataset) to estimate the predictive power of the specializations discovered. It is worth noting that at this point we have established the process which can be used to identify relationships which are interesting at least over the training data. Subsequently the user might want to evaluate the quality of the relationships discovered by using a test set either via the associative classifier which would provide the classification accuracy of the specialized rules or evaluate the different measures of interestingness using the VSM. Since the functionality to support associative classification is currently work under progress as a part of another effort here at WPI [Rudss], we decided to use the VSM to evaluate the strength of the relationship. Figure 6.16 lists the different measures of interestingness over the **test dataset** by using the VSM. This helps to determine if the specializations are really interesting, or if they are a case of overfit to the training data. Observe specifically the various measures for rules 1.01, both of which have a positive lift and has all the measures of interestingness as good as the base rule. Given that no extensive process was

Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.51428574	0.13333334	1.6530613	2.5550831E-3	0.42857143
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6363636	0.1037037	2.0454545	3.1594354E-4	0.33333334
002	M54 && M80	expr=HSN	0.4878049	0.14814815	1.5679443	3.4031714E-3	0.47619048
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.6	0.11111111	1.9285715	5.4719937E-4	0.35714287
003	M47 && M53	expr=ADL	0.2982456	0.12592593	1.6776316	1.7501895E-3	0.7083333
003.01	M47 [0-250] M53	expr=ADL	0.47826087	0.08148148	2.6902175	3.5008453E-5	0.45833334
004	M77 [rp0-rp1] M48 [rp2-rp3]	expr=ALM	0.4	0.1037037	1.9285715	1.0940551E-3	0.5
004.01	M48:b [0-500] M77:a && M77:a [rp0-rp1] M48:b[rp2-rp3]	expr=ALM	0.53846157	0.05185185	2.5961537	1.955872E-3	0.25
005	M69	expr=HSN	0.46341464	0.14074074	1.489547	1.1586854E-2	0.45238096
005.01	SoT [0-500] M69	expr=HSN	0.6666667	0.044444446	2.142857	1.7081774E-2	0.14285715
006	M6	expr=ADL	0.13636364	0.044444446	0.76704544	3.8148643E-1	0.25
006.01	SoT [0-350] M6	expr=HSN	0.6666667	0.02962963	2.142857	5.4289247E-2	0.0952381

Figure 6.15: Statistical measures over training data.
Base rules alongside their corresponding specializations and various measures of interestingness calculated over the training data, which help identify which specializations are interesting, and how much more interesting.

followed to divide the instances between the test set and the training set we believe this rule still might be an interesting discovery. Also we have hereby established the process that domain users might follow to verify the strength of a relationship which seems to be interesting based on the training data. It is important to understand at this point that this implementation equips the domain user with a tool and a work flow to identify potential relationships. The actual quality of the specializations is again contingent on many other things. For instance, the quality of the motifs elicited as well as the mining process both of which are outside the scope of this project. It is worth noting that both these areas are being researched in separate efforts here at WPI.

Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M6 && M52	expr=HSN	0.6666667	0.16666667	2.0	1.5729921E-1	0.5
001.01	M6 [rp0-rp1] M52 [rp2-rp3]	expr=HSN	0.6666667	0.16666667	2.0	1.5729921E-1	0.5
002	M54 && M80	expr=HSN	0.2	0.083333336	0.6	4.0762594E-1	0.25
002.01	M54[rp0-rp1] M80[rp2-rp3]	expr=HSN	0.33333334	0.083333336	1.0	1E0	0.25
003	M47 && M53	expr=ADL	0.5	0.25	1.2	5.5818464E-1	0.6
003.01	M47 [0-250] M53	expr=ADL	0.33333334	0.083333336	0.8	7.3531669E-1	0.2
004	M77 [rp0-rp1] M48 [rp2-rp3]	expr=ALM	0.6666667	0.16666667	2.0	1.5729921E-1	0.5
004.01	M48:b [0-500] M77:a && M77:a [rp0-rp1] M48:b[rp2-rp3]	expr=ALM	0.0	0.0	0.0	1E0	0.0
005	M69	expr=HSN	1.0	0.16666667	3.0	2.8459734E-2	0.5
005.01	SoT [0-500] M69	expr=HSN	1.0	0.083333336	3.0	1.3964939E-1	0.25
006	M6	expr=ADL	0.0	0.0	0.0	1E0	0.0
006.01	SoT [0-350] M6	expr=HSN	0.0	0.0	0.0	1E0	0.0

Figure 6.16: Specializations and various measures of interestingness evaluated over the **test data**.

Chapter 7

Conclusions and Future Work

The goal of this thesis was to computationally enable the discovery of gene expression association rules based on several biological hypotheses. We designed and implemented a tool that helps domain experts visualize genetic data in the context of various biological hypotheses and perform exploratory analysis of data to discover specialized gene expression association rules. This process of exploratory analysis allows for post mining specialization of association rules, which alleviates some of the shortcomings of incorporating hypothesis-driven information into the mining process.

This work sketched out a process work flow for exploratory analysis of genetic data to discover interesting association rules. This work facilitates the process of identifying interesting rules beyond the conventional support-confidence framework by adding other measures of interestingness to the analysis process. We established via the experimental evaluation (Section 6) that the data visualization capabilities provided by our tool helps human experts in identifying hypothesis-driven specialized rules that score better than their generic counterparts in terms of different measures of interestingness. In addition to the visual mining tool, this work provides an updated genetic dataset which is an important resource for future research.

Future work would involve verifying the scalability of the tool to ensure that the tool performs well with a substantially larger dataset (e.g., so as to support a genetic database from micro array experiments). Another potential area worth investigating is to provide the functionality wherein the tool suggests patterns to the user based on the data being visualized. For instance, the approach proposed in [Ice03] could be used in an inter-motif distance plot to suggest specialization tips (visual or textual) to the domain user based only on the section of the data relevant in the context of the rule being visualized.

Bibliography

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207–216, Washington, D.C., May 1993. ACM.

[Alv03] Sergio A. Alvarez. Chi-squared computation for association rules: Preliminary results. Technical Report BC-CS-2003-01, Computer Science Department, Boston College, July 2003.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th International Conference on Very Large Databases*, Sep 1994.

[BE94] Timothy Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press, August 1994.

[BFG$^+$03] John Baird, Jay Farmer, Rebecca Gougian, Ken Monterio, and Paul Young. Motif analysis of gene expression. Undergraduate Graduation Project (MQP), 2003.

[BLA] Wormbase sequence search. BLAST search Homepage: `http://www.wormbase.org/db/searches/blat`.

[BLT02] Kristin Blitsch, Ben Lucas, and Sarah Towey. Computational analysis of gene expression. Undergraduate Graduation Project (MQP), 2002.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.

[Ice03] Aleksandar Icev. DARM: Distance-based association rule mining. Master's thesis, Worcester Polytechnic Institute, May 2003.

[IRR03] A. Icev, C. Ruiz, and E. Ryder. Distance-enhanced association rules for gene expression. In *Proc. 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD2003). Held in conjunction with the 9th Intl. Conf. on Knowledge Discovery and Data Mining (KDD2003)*, pages 34–40, Washington DC, USA, Aug. 2003.

[JAC95] Liu JS, Neuwald AF, and Lawrence CE. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association*, 90(432):1156–1169, Dec 1995.

[JCC] Javacc. JavaCC Homepage: `https://javacc.dev.java.net/`.

[JFr] Jfreechart. JFreeChart Homepage: `http://www.jfree.org/jfreechart/`.

[MAS] MAST online. MAST Homepage: `http://meme.sdsc.edu/meme/website/mast-intro.html`.

[MEM] MEME online. MEME Homepage: http://meme.sdsc.edu/meme/website/intro.html.

[Mor98] S. Morishita. On classification and regression. In *Proc. of the First Intl Conf on Discovery Science – Lecture Notes in Artificial Intelligence*, pages 40–57. Discovery Science, 1998.

[Mot] MotifSampler online. MotifSampler Homepage: http://homes.esat.kuleuven.be/∼thijs/Work/MotifSampler.html.

[MPPT01] B. Murphy, D. Phu, I. Pushee, and F. Tan. Motif-and expression-based classification of DNA. Undergraduate Graduation Project (MQP), 2001.

[PR05] K.A. Pray and C. Ruiz. Mining expressive temporal associations from complex data. In *Proc. of the International Conference on Machine Learning and Data Mining*, pages 384–394, Leipzig, Germany, July 2005. MLDM.

[RSA] RSA: Regulatory sequence analysis tools. RSA Homepage: http://rsat.ulb.ac.be/rsat/.

[Rudss] Jonathon Rudolph. Gene expression rule modeling. Undergraduate Graduation Project (MQP), In progress.

[Sho01] Christopher Shoemaker. Mining association rules over set-valued data. Master's thesis, Worcester Polytechnic Institute, May 2001.

[Wor] Wormbase release ws119. WormBase Homepage: http://www.wormbase.org.

[WPI] WPI-Weka system http://sourceforge.wpi.edu. Project@Repository: Bioinformatics@wekacode.

Appendix A

Sample ARFF File

```
@relation test_ASAS
@attribute gene string
@attribute M1  string
%Width = 8
%Consensus = CCGGCAAT
%Log-Odds Matrix:
%-689   217     -684    -91
%-689   246     -684    -690
%-689   -685    264     -690
%-689   -685    264     -690
%-21    200     -684    -690
%165    -685    -684    -690
%165    -685    -684    -690
%-689   -685    -684    153

@attribute M2  string
%Width = 8
%Consensus = GGAAAACG
%Log-Odds Matrix:
%-736   -731    264     -737
%-736   -731    264     -737
%151    -89     -729    -737
%166    -731    -729    -737
%166    -731    -729    -737
%166    -731    -729    -737
```

```
%-736   232     -729    -181

%-736   -731    221     -41


@attribute M3  string

%Width = 8

%Consensus = GAGAGAGA

%Log-Odds Matrix:

%-689   -685    264     -690

%128    -685    50      -690

%-21    -685    218     -690

%165    -685    -684    -690

%-689   -685    264     -690

%128    -685    50      -690

%-689   -685    264     -690

%165    -685    -684    -690


@attribute M4  string

%Width = 10

%Consensus = GAGArAGAGA

%Log-Odds Matrix:

%-683   -679    263     -683

%143    -679    -16     -683

%-683   -679    263     -683

%158    -679    -172    -683

%44     -91     157     -683

%135    -679    -678    -84

%-683   -679    263     -683

%107    -679    -678    -4

%-683   -679    263     -683

%126    -679    57      -683


@attribute M5  string

%Width = 10

%Consensus = GAGAGrsAGn

%Log-Odds Matrix:

%-727   -722    264     -728

%150    -722    -62     -728

%-727   -235    259     -728

%166    -722    -720    -728
```

```
%-216   42      210     -728
%89     -722    137     -728
%-727   131     178     -728
%112    -722    59      -228
%-727   -722    264     -728
%80     -39     95      -728

@attribute M6  string
%Width = 10
%Consensus = rGAAGAAGAn
%Log-Odds Matrix:
%85     -679    125     -278
%-683   -679    263     -683
%135    -679    -16     -278
%165    -679    -678    -683
%-683   -679    263     -683
%165    -679    -678    -683
%143    -679    -678    -125
%-683   -679    263     -683
%151    -91     -678    -683
%85     -679    105     -182

@attribute M7  string
%Width = 12
%Consensus = AAwTTGCCGGAA
%Log-Odds Matrix:
%152    -685    -80     -690
%152    -685    -80     -690
%52     -685    -684    66
%-689   -685    -684    153
%-689   -685    -23     132
%-120   1       208     -690
%-689   246     -684    -690
%-120   200     -23     -690
%-689   -685    257     -284
%-689   -685    264     -690
%165    -685    -684    -690
%152    -98     -684    -690
```

```
@attribute M8  string

%Width = 12

%Consensus = rAGAAGArGAAr

%Log-Odds Matrix:

%46     -691    181     -696

%130    -691    44      -696

%-695   -691    264     -696

%166    -691    -690    -696

%159    -691    -185    -696

%-126   -691    243     -696

%166    -691    -690    -696

%83     -691    129     -291

%-5     -691    211     -696

%122    -691    70      -696

%166    -691    -690    -696

%60     -691    170     -696


@attribute M9  string

%Width = 12

%Consensus = GrGAGAGwGAGm

%Log-Odds Matrix:

%-91    -658    237     -662

%49     -658    179     -662

%-91    -658    237     -662

%139    -658    6       -662

%-18    -658    216     -662

%148    -658    -657    -160

%-50    -658    227     -662

%81     -658    -150    18

%-662   -167    255     -662

%139    -658    -657    -103

%-662   -658    263     -662

%49     161     -657    -662


@attribute M10  string

%Width = 8

%Consensus = GGGnGGnG

%Log-Odds Matrix:

%-748   -742    264     -749
```

```
%-313   -742    259     -749
%-748   -742    264     -749
%-19    151     37      -749
%-748   -742    264     -749
%-748   -742    264     -749
%105    -138    78      -749
%-748   -742    264     -749

@attribute M11  string
%Width = 8
%Consensus = GGGAGrAG
%Log-Odds Matrix:
%-716   -711    264     -717
%-716   -711    264     -717
%-716   -711    264     -717
%122    -711    70      -717
%-716   -711    264     -717
%31     -711    192     -717
%113    -711    92      -717
%-716   -711    264     -717

@attribute M12  string
%Width = 8
%Consensus = GGGnGGAG
%Log-Odds Matrix:
%0      -777    209     -785
%-784   -777    264     -785
%-784   -777    264     -785
%0      147     -774    -90
%-784   -777    264     -785
%-784   -777    264     -785
%141    -19     -774    -785
%-784   -777    264     -785

@attribute M13  string
%Width = 10
%Consensus = rnGGGnGGAG
%Log-Odds Matrix:
%94     -711    129     -717
```

```
%-27    -711    157     -40
%-27    -711    220     -717
%-716   -711    264     -717
%-716   -711    264     -717
%-27    140     -710    -40
%-716   -711    264     -717
%-716   -711    264     -717
%138    -5      -710    -717
%-716   -711    264     -717

@attribute M14  string
%Width = 10
%Consensus = GGGnGGnGnn
%Log-Odds Matrix:
%-781   -773    264     -781
%-348   -773    260     -781
%-781   -773    264     -781
%-22    134     -24     -134
%-781   -773    264     -781
%-781   -773    264     -781
%104    -42     43      -781
%-781   -773    260     -360
%-781   72      76      33
%-22    43      60      -50

@attribute M15  string
%Width = 10
%Consensus = AnnGGmGGAG
%Log-Odds Matrix:
%115    -773    89      -781
%-37    -773    152     -21
%4      -773    202     -360
%-73    -773    234     -781
%-781   -773    264     -781
%15     143     2       -781
%-781   -773    264     -781
%-781   -773    264     -781
%121    -773    76      -781
%-781   -773    264     -781
```

```
@attribute M16  string
%Width = 12
%Consensus = GGnGGrnGrGGr
%Log-Odds Matrix:
%-614   -612    263     -615
%-614   -612    263     -615
%-78    -612    198     -90
%-175   -612    217     -90
%-614   -612    234     -90
%20     -612    198     -615
%-21    -612    198     -186
%-614   59      217     -615
%78     -612    149     -615
%-614   -612    217     -33
%-614   -612    249     -186
%20     -612    198     -615

@attribute M17  string
%Width = 12
%Consensus = AwrkGGGmGGAG
%Log-Odds Matrix:
%112    -742    -62     -73
%26     -742    11      38
%97     -235    111     -749
%-217   -39     148     0
%-748   -742    264     -749
%-748   -742    264     -749
%-748   -742    264     -749
%89     119     -740    -749
%-217   -742    254     -749
%-61    -742    230     -749
%144    -742    -21     -749
%-748   -742    264     -749

@attribute M18  string
%Width = 12
%Consensus = kGrGkrrGkGkG
%Log-Odds Matrix:
```

```
%-657   -141    190     -4
%-657   -654    219     -36
%75     -654    154     -658
%-657   -654    263     -658
%-657   -654    190     22
%56     -654    131     -134
%56     -654    173     -658
%-657   -43     243     -658
%-657   -654    173     44
%-657   -654    263     -658
%-657   -654    131     80
%-657   -654    263     -658

@attribute M19  string
%Width = 8
%Consensus = GGGsGGrG
%Log-Odds Matrix:
%-836   -825    264     -837
%-836   -825    264     -837
%-63    -825    228     -414
%-836   110     193     -837
%-836   -825    264     -837
%-92    -825    238     -837
%43     -825    184     -837
%-836   -825    264     -837

@attribute M20  string
%Width = 8
%Consensus = GGGGGnGG
%Log-Odds Matrix:
%-834   -823    264     -835
%-36    -823    224     -835
%-834   -823    264     -835
%-74    -823    234     -835
%-106   -823    241     -835
%-60    74      162     -835
%-834   -823    264     -835
%-834   -823    264     -835
```

```
@attribute M21  string

%Width = 8

%Consensus = GGGnGGrG

%Log-Odds Matrix:

%-52    -836    228     -849

%-848   -836    264     -849

%-848   -836    264     -849

%-22    111     101     -427

%-848   -836    264     -849

%-848   -836    264     -849

%64     -836    167     -849

%-848   -836    264     -849


@attribute M22  string

%Width = 10

%Consensus = GAArAAGAAG

%Log-Odds Matrix:

%-763   -756    264     -764

%129    -756    50      -764

%166    -756    -754    -764

%41     -756    186     -764

%116    -756    86      -764

%166    -756    -754    -764

%-763   -756    264     -764

%166    -756    -754    -764

%161    -756    -754    -335

%-763   -756    264     -764


@attribute M23  string

%Width = 10

%Consensus = GAAGAAGAAn

%Log-Odds Matrix:

%-797   -789    264     -798

%166    -789    -786    -798

%166    -789    -786    -798

%-797   -789    264     -798

%166    -789    -786    -798

%151    -85     -786    -798

%-797   -789    264     -798
```

```
%134    -789    32      -798

%129    -789    49      -798

%-49    -53     202     -798


@attribute M24  string

%Width = 10

%Consensus = GnGrGsGrGG

%Log-Odds Matrix:

%-854   -842    264     -855

%-20    -842    190     -141

%-38    -842    224     -855

%23     -842    197     -855

%-4     -842    211     -855

%-854   104     197     -855

%-422   -842    262     -855

%29     -842    194     -855

%3      -842    208     -855

%-854   -842    264     -855


@attribute M25  string

%Width = 12

%Consensus = TGTGTrTGTGTG

%Log-Odds Matrix:

%-245   -685    -684    145

%-689   -685    264     -689

%-50    -685    -684    117

%-689   -685    264     -689

%-689   30      -684    117

%30     -685    193     -689

%-50    -685    -684    117

%-689   -685    264     -689

%8      -685    -684    95

%-245   -685    255     -689

%-148   -685    -684    136

%-689   -685    264     -689


@attribute M26  string

%Width = 12

%Consensus = AAGAAGAAGAAG
```

```
%Log-Odds Matrix:
%138    -768    14      -776
%153    -103    -766    -776
%-775   -768    264     -776
%166    -768    -766    -776
%148    -768    -44     -776
%4      -768    208     -776
%116    -61     14      -776
%133    -103    -44     -776
%-775   -768    264     -776
%166    -768    -766    -776
%138    -768    14      -776
%-775   -768    264     -776

@attribute M27  string
%Width = 12
%Consensus = TGTGTGTGTGTG
%Log-Odds Matrix:
%-50    -685    -684    117
%-689   -685    264     -689
%-148   -11     -684    106
%-18    -685    217     -689
%-689   -685    -684    153
%-689   -685    264     -689
%8      -685    -684    95
%-50    -685    227     -689
%-91    -685    -684    127
%-689   -685    264     -689
%-148   -685    -684    136
%-689   -685    264     -689

@attribute M28  string
%Width = 8
%Consensus = nGGmGGAG
%Log-Odds Matrix:
%-2     20      159     -738
%-737   -732    264     -738
%-737   -732    264     -738
%50     107     -730    -99
```

```
%-737   -732    264     -738
%-737   -732    264     -738
%166    -732    -730    -738
%-737   -732    264     -738

@attribute M29  string
%Width = 8
%Consensus = GAAGAAGA
%Log-Odds Matrix:
%-664   -661    263     -664
%165    -661    -660    -664
%165    -661    -660    -664
%-664   -661    263     -664
%165    -661    -660    -664
%165    -661    -660    -664
%-664   -661    263     -664
%165    -661    -660    -664

@attribute M30  string
%Width = 8
%Consensus = GAGAGAGA
%Log-Odds Matrix:
%-679   -676    263     -680
%149    -676    -59     -680
%-679   -676    263     -680
%111    -676    -675    -12
%-679   -676    263     -680
%165    -676    -675    -680
%-679   -676    263     -680
%165    -676    -675    -680

@attribute M31  string
%Width = 10
%Consensus = GATTTACGrG
%Log-Odds Matrix:
%-723   -717    258     -309
%166    -717    -716    -723
%-723   -219    -716    148
%-297   -717    -716    148
```

```
%-723   -717    -716    153

%166    -717    -716    -723

%-723   240     -716    -309

%-201   -717    252     -723

%54     -717    175     -723

%-297   -717    258     -723


@attribute M32  string

%Width = 10

%Consensus = AGAAGAAGAw

%Log-Odds Matrix:

%166    -741    -739    -747

%-746   -741    264     -747

%166    -741    -739    -747

%166    -741    -739    -747

%-746   -741    264     -747

%166    -741    -739    -747

%110    -741    101     -747

%-746   -741    264     -747

%135    10      -739    -747

%29     -741    69      4


@attribute M33  string

%Width = 10

%Consensus = nAAGAAGAAG

%Log-Odds Matrix:

%-11    -707    186     -144

%116    -707    86      -712

%166    -707    -706    -712

%-712   -707    264     -712

%166    -707    -706    -712

%166    -707    -706    -712

%-712   -707    264     -712

%166    -707    -706    -712

%116    -707    -35     -103

%-712   -707    264     -712


@attribute M34  string

%Width = 12
```

```
%Consensus = AAGAnGAAGAAG
%Log-Odds Matrix:
%146    -653    -652    -143
%146    -150    -133    -656
%-228   -653    254     -656
%165    -653    -652    -656
%98     -653    64      -143
%-131   -653    233     -239
%165    -653    -652    -656
%165    -653    -652    -656
%-656   -653    263     -656
%135    -52     -133    -656
%146    -52     -652    -656
%-656   -653    263     -656

@attribute M35  string
%Width = 12
%Consensus = ATGATGATGATG
%Log-Odds Matrix:
%117    -131    42      -637
%-637   -634    -634    153
%-637   -634    263     -637
%165    -634    -634    -637
%-112   -634    -634    131
%-637   24      228     -637
%154    -634    -634    -220
%-637   -634    -114    142
%-209   -634    252     -637
%165    -634    -634    -637
%-637   -634    -114    142
%-637   -33     241     -637

@attribute M36  string
%Width = 12
%Consensus = GAArAAGAAGAr
%Log-Odds Matrix:
%-656   -653    263     -656
%165    -653    -652    -656
%124    -653    64      -656
```

```
%47     -653    180     -656
%165    -653    -652    -656
%146    -653    -652    -143
%-656   -653    233     -86
%146    -653    -35     -656
%165    -653    -652    -656
%-656   -653    263     -656
%111    -52     23      -656
%98     -653    122     -656

@attribute M37  string
%Width = 8
%Consensus = GGGmGGnG
%Log-Odds Matrix:
%-570   -569    262     -571
%-570   -569    262     -571
%-570   -569    262     -571
%65     145     -568    -571
%-570   -83     247     -571
%-570   -569    262     -571
%91     14      32      -571
%-570   -569    262     -571

@attribute M38  string
%Width = 8
%Consensus = AGGTAGGC
%Log-Odds Matrix:
%124    -633    64      -636
%-636   -633    263     -636
%-636   -633    263     -636
%-636   -633    -632    153
%124    -633    -632    -45
%-636   -633    263     -636
%-636   -633    263     -636
%-636   245     -632    -636

@attribute M39  string
%Width = 8
%Consensus = GAAGAAGA
```

```
%Log-Odds Matrix:
%-607   -605    263     -607
%165    -605    -604    -607
%165    -605    -604    -607
%-607   -605    263     -607
%165    -605    -604    -607
%165    -605    -604    -607
%-607   -605    263     -607
%165    -605    -604    -607

@attribute M40  string
%Width = 10
%Consensus = GAAGAAGAAG
%Log-Odds Matrix:
%-570   -569    262     -571
%132    -569    32      -571
%113    72      -568    -571
%-7     -569    211     -571
%164    -569    -568    -571
%164    -569    -568    -571
%-570   -569    262     -571
%164    -569    -568    -571
%164    -569    -568    -571
%-570   -569    262     -571

@attribute M41  string
%Width = 10
%Consensus = GrAGmAGAAG
%Log-Odds Matrix:
%-596   -594    263     -596
%8      -594    204     -596
%165    -594    -593    -596
%-596   -594    250     -197
%65     119     -92     -596
%165    -594    -593    -596
%-596   -594    250     -197
%139    -11     -593    -596
%165    -594    -593    -596
%-596   -594    263     -596
```

```
@attribute M42  string

%Width = 10

%Consensus = GnAGGCAGGC

%Log-Odds Matrix:

%-644   -641    263     -644

%-235   -158    55      90

%165    -641    -640    -644

%-644   -641    263     -644

%-644   -641    263     -644

%-644   195     -640    -22

%127    -641    -141    -95

%-644   -641    245     -152

%-235   -3      225     -644

%-644   218     -640    -95


@attribute M43  string

%Width = 12

%Consensus = AAGAArrAGAAG

%Log-Odds Matrix:

%145    -538    -34     -539

%164    -538    -537    -539

%-539   -538    262     -539

%145    -538    -34     -539

%164    -538    -537    -539

%97     -538    121     -539

%97     -538    121     -539

%164    -538    -537    -539

%-539   -538    262     -539

%164    -538    -537    -539

%164    -538    -537    -539

%-539   -538    262     -539


@attribute M44  string

%Width = 12

%Consensus = AGAAGAAGrAGA

%Log-Odds Matrix:

%164    -519    -519    -521

%-521   -519    261     -521
```

```
%164    -519    -519    -521
%164    -519    -519    -521
%-521   -519    261     -521
%164    -519    -519    -521
%116    -33     -16     -521
%-521   -519    261     -521
%84     -519    140     -521
%116    65      -519    -521
%-521   -33     239     -521
%116    65      -519    -521

@attribute M45  string
%Width = 12
%Consensus = GACGACGACGnC
%Log-Odds Matrix:
%-176   -195    242     -680
%137    -676    18      -680
%-21    191     -675    -284
%-680   -195    257     -680
%152    -676    -80     -680
%-680   217     -23     -284
%-79    -676    235     -680
%137    -195    -23     -680
%-680   246     -675    -680
%-21    -676    208     -284
%52     -676    99      -59
%-680   246     -675    -680

@attribute M46  string
%Width = 8
%Consensus = GnrGGnGG
%Log-Odds Matrix:
%-722   -717    264     -723
%-43    -717    74      54
%21     -717    198     -723
%-164   -717    249     -723
%-722   -717    264     -723
%-43    146     -716    -36
%-722   -717    264     -723
```

```
%-722   -717    264     -723

@attribute M47  string
%Width = 8
%Consensus = GAGAArAG
%Log-Odds Matrix:
%-718   -713    264     -718
%119    -713    -711    -32
%-718   -713    264     -718
%166    -713    -711    -718
%166    -713    -711    -718
%89     -713    137     -718
%166    -713    -711    -718
%-718   -713    264     -718

@attribute M48  string
%Width = 8
%Consensus = TGAGAAAA
%Log-Odds Matrix:
%-743   -738    -736    153
%-743   -738    264     -744
%148    -738    -48     -744
%-743   -738    264     -744
%153    -107    -736    -744
%166    -738    -736    -744
%166    -738    -736    -744
%166    -738    -736    -744

@attribute M49  string
%Width = 10
%Consensus = ArAGrAnGAG
%Log-Odds Matrix:
%120    -676    77      -680
%20     -676    198     -680
%165    -676    -675    -680
%-680   -676    263     -680
%78     -676    150     -680
%128    -676    50      -680
%78     -676    77      -91
```

```
%-680   -676    263     -680
%165    -676    -675    -680
%-272   -676    257     -680

@attribute M50  string
%Width = 10
%Consensus = rGAArAAAGA
%Log-Odds Matrix:
%28     -698    174     -212
%-703   -698    264     -703
%113    -698    94      -703
%142    -698    -5      -703
%54     -698    174     -703
%166    -698    -697    -703
%166    -698    -697    -703
%135    -698    27      -703
%-703   -698    264     -703
%166    -698    -697    -703

@attribute M51  string
%Width = 10
%Consensus = GAGArnAAGA
%Log-Odds Matrix:
%-123   -717    243     -723
%107    -717    106     -723
%-722   -717    264     -723
%166    -717    -716    -723
%21     -717    198     -723
%100    -717    32      -103
%114    -717    90      -723
%134    -717    32      -723
%-722   -717    264     -723
%166    -717    -716    -723

@attribute M52  string
%Width = 12
%Consensus = GAAnAnGAAGAA
%Log-Odds Matrix:
%-33    -663    212     -270
```

```
%165    -663    -662    -667
%142    -663    -9      -667
%103    -663    112     -667
%165    -663    -662    -667
%92     -663    90      -175
%-667   -26     240     -667
%142    -663    -662    -118
%165    -663    -662    -667
%-667   -663    263     -667
%158    -663    -164    -667
%165    -663    -662    -667

@attribute M53  string
%Width = 12
%Consensus = GAArAAGAAnrA
%Log-Odds Matrix:
%8      -688    205     -692
%166    -688    -687    -692
%166    -688    -687    -692
%25     -688    196     -692
%124    -688    64      -692
%159    -688    -687    -296
%-11    -688    205     -296
%166    -688    -687    -692
%166    -688    -687    -692
%-692   -208    196     -4
%98     -688    122     -692
%166    -688    -687    -692

@attribute M54  string
%Width = 12
%Consensus = GArwGAGArrnG
%Log-Odds Matrix:
%-713   -708    264     -714
%137    -708    16      -714
%44     -708    174     -318
%31     -708    -16     43
%-713   -708    264     -714
%117    -708    83      -714
```

```
%-154   -708    247     -714
%110    -34     16      -714
%102    -708    115     -714
%66     -34     115     -714
%76     -2      83      -714
%-713   -708    264     -714

@attribute M55  string
%Width = 8
%Consensus = rGGCGGnG
%Log-Odds Matrix:
%83     -687    144     -691
%-691   -687    264     -691
%-126   -687    243     -691
%-53    194     -87     -691
%-691   -687    264     -691
%-691   -687    264     -691
%104    94      -686    -691
%-691   -687    264     -691

@attribute M56  string
%Width = 8
%Consensus = GGGmGGAG
%Log-Odds Matrix:
%-708   -703    258     -308
%-103   -703    240     -708
%-708   -703    264     -708
%77     135     -702    -708
%-200   -703    252     -708
%-4     -703    211     -708
%113    -23     -5      -708
%-708   -703    264     -708

@attribute M57  string
%Width = 8
%Consensus = GGnGnsGG
%Log-Odds Matrix:
%-678   -675    263     -679
%-678   -675    241     -125
```

```
%-113   88      170     -679

%-40    -675    224     -679

%8      8       157     -679

%-678   177     125     -679

%-678   -675    263     -679

%-678   -675    263     -679


@attribute M58  string

%Width = 10

%Consensus = nnnGGnGGnG

%Log-Odds Matrix:

%14     -202    192     -691

%-85    -104    181     -97

%83     -5      70      -691

%-691   -687    264     -691

%-691   -104    250     -691

%-691   174     -87     -17

%-691   -687    264     -691

%-691   -687    257     -290

%60     53      70      -691

%-691   -687    264     -691


@attribute M59  string

%Width = 10

%Consensus = GrnnnGGCGG

%Log-Odds Matrix:

%-159   -718    203     -51

%39     -718    187     -723

%-60    -718    187     -73

%-723   42      195     -130

%-60    -39     78      14

%-159   -718    248     -723

%-723   -718    264     -723

%-723   185     110     -723

%-723   -718    264     -723

%-723   -718    264     -723


@attribute M60  string

%Width = 10
```

```
%Consensus = rGnGGGGGkG

%Log-Odds Matrix:

%25     -693    196     -697

%-33    -693    222     -697

%-285   88      -35     54

%-697   -693    264     -697

%-132   -693    222     -144

%-697   88      205     -697

%-697   -208    205     -24

%-697   -693    264     -697

%-59    -693    122     27

%-697   -693    264     -697


@attribute M61  string

%Width = 12

%Consensus = nGTGTGTGTGTG

%Log-Odds Matrix:

%88     -599    6       -45

%-601   -599    263     -601

%-90    -599    -598    126

%-90    -599    236     -601

%-601   -599    -598    153

%-601   -599    236     -102

%-186   -599    64      94

%-33    -599    221     -601

%-601   -599    -92     140

%-601   -599    263     -601

%-601   -599    -92     140

%-601   -599    263     -601


@attribute M62  string

%Width = 12

%Consensus = wGTGnGknkGTG

%Log-Odds Matrix:

%46     -687    -184    60

%-691   -5      236     -691

%-279   -687    -686    147

%-691   -202    257     -691

%14     -687    70      18
```

```
%-691   -687    243     -138
%-691   -104    128     60
%14     -687    202     -691
%-691   -687    144     71
%-126   -687    243     -691
%-691   -687    44      118
%-691   -687    264     -691

@attribute M63  string
%Width = 12
%Consensus = GArAAGArrnnG
%Log-Odds Matrix:
%-45    -64     202     -708
%154    -122    -702    -708
%86     -703    140     -708
%154    -703    -104    -708
%166    -703    -702    -708
%-708   -703    264     -708
%166    -703    -702    -708
%96     -703    126     -708
%86     -703    140     -708
%54     -703    111     -83
%86     -703    94      -155
%-708   -703    264     -708

@attribute M64  string
%Width = 8
%Consensus = GGGmGGAG
%Log-Odds Matrix:
%-644   -641    263     -644
%-644   -641    263     -644
%-644   -641    263     -644
%16     118     55      -644
%-644   -158    254     -644
%-644   -641    263     -644
%165    -641    -640    -644
%-644   -641    263     -644

@attribute M65  string
```

```
%Width = 8
%Consensus = GGGrGGrG
%Log-Odds Matrix:
%-667   -663    263     -667
%-667   -663    256     -270
%-667   -663    263     -667
%15     -663    148     -77
%-667   -663    263     -667
%-259   -663    256     -667
%103    -663    112     -667
%-667   -663    263     -667

@attribute M66  string
%Width = 8
%Consensus = GkGkGrGG
%Log-Odds Matrix:
%-739   -734    264     -740
%-739   -734    201     4
%-25    -734    220     -740
%-739   -734    187     26
%-739   -734    264     -740
%82     -734    146     -740
%-739   -734    264     -740
%-739   -734    264     -740

@attribute M67  string
%Width = 10
%Consensus = GAAGnAGAnG
%Log-Odds Matrix:
%-680   -676    263     -680
%159    -676    -178    -680
%165    -676    -675    -680
%-680   -676    235     -91
%100    59      -80     -680
%144    -676    -23     -680
%-680   -676    263     -680
%165    -676    -675    -680
%100    -195    -80     -59
%-680   -676    263     -680
```

```
@attribute M68  string
%Width = 10
%Consensus = GAAGAAGAnG
%Log-Odds Matrix:
%-718   -713    264     -718
%166    -713    -711    -718
%166    -713    -711    -718
%-718   -713    264     -718
%161    -713    -217    -718
%119    -713    59      -323
%-159   -713    242     -323
%105    -713    110     -718
%61     -137    59      -73
%-718   -235    259     -718

@attribute M69  string
%Width = 10
%Consensus = GrGnGrGAGr
%Log-Odds Matrix:
%-735   -730    264     -736
%70     -730    160     -736
%-735   -730    264     -736
%32     -730    77      -4
%-137   -730    245     -736
%79     -730    130     -246
%-57    -730    229     -736
%142    -730    -8      -736
%-735   -730    264     -736
%52     -730    176     -736

@attribute M70  string
%Width = 12
%Consensus = CGATGCACCATG
%Log-Odds Matrix:
%-686   226     -681    -138
%-126   -682    243     -686
%159    -682    -681    -290
%-686   -201    -681    147
```

```
%-686   -682    264     -686

%-686   239     -681    -290

%165    -682    -681    -686

%-686   226     -681    -138

%-686   239     -681    -290

%159    -682    -184    -686

%-686   -682    -681    153

%-53    -682    228     -686


@attribute M71  string

%Width = 12

%Consensus = AAGAAGAAGAnG

%Log-Odds Matrix:

%165    -649    -648    -652

%129    30      -648    -652

%-91    -649    227     -255

%165    -649    -648    -652

%139    -649    -52     -255

%-652   -649    263     -652

%148    -166    -149    -652

%139    -166    -52     -652

%-652   -649    263     -652

%165    -649    -648    -652

%30     62      47      -160

%-652   -649    263     -652


@attribute M72  string

%Width = 12

%Consensus = GAAGAAGAnGAA

%Log-Odds Matrix:

%-652   -649    263     -652

%165    -649    -648    -652

%129    -649    47      -652

%-50    -649    227     -652

%165    -649    -648    -652

%165    -649    -648    -652

%-50    -649    227     -652

%148    -166    -149    -652

%95     -69     -648    -30
```

```
%-652   -649    263     -652
%129    -649    47      -652
%139    -166    -52     -652

@attribute M73  string
%Width = 8
%Consensus = GrkGGGGG
%Log-Odds Matrix:
%-718   -713    264     -718
%102    -713    115     -718
%-718   -713    164     54
%-718   -713    264     -718
%-718   -713    235     -94
%-718   -34     208     -94
%-718   -713    264     -718
%-718   -713    264     -718

@attribute M74  string
%Width = 8
%Consensus = GrAGAGAG
%Log-Odds Matrix:
%-697   -693    264     -697
%77     -693    151     -697
%146    -53     -691    -697
%-697   -693    264     -697
%139    -693    -691    -103
%-697   -693    264     -697
%159    -693    -191    -697
%-697   -693    264     -697

@attribute M75  string
%Width = 8
%Consensus = kGnkGGGG
%Log-Odds Matrix:
%-198   -753    164     29
%-759   -753    264     -760
%94     -753    20      -70
%-759   -753    185     29
%-125   -753    244     -760
```

```
%-759   -176    256     -760

%-759   -753    264     -760

%-759   -753    264     -760


@attribute M76  string

%Width = 10

%Consensus = GGrGrnGGAG

%Log-Odds Matrix:

%13     -718    203     -723

%-216   -718    253     -723

%71     -718    159     -723

%-312   -718    259     -723

%39     -718    187     -723

%13     61      124     -723

%13     -718    203     -723

%-723   -718    264     -723

%155    -718    -120    -723

%-723   -718    264     -723


@attribute M77  string

%Width = 10

%Consensus = nAGAnGAAGA

%Log-Odds Matrix:

%8      -693    196     -296

%166    -693    -691    -697

%-697   -693    264     -697

%159    -208    -691    -697

%66     88      6       -697

%-697   -693    264     -697

%116    69      -691    -697

%166    -693    -691    -697

%-697   -693    264     -697

%139    -693    6       -697


@attribute M78  string

%Width = 10

%Consensus = GGnGGnGGAG

%Log-Odds Matrix:

%-38    -735    224     -741
```

```
%-740   -735    264     -741

%-57    -735    150     -4

%-740   42      224     -741

%-79    -25     205     -741

%-38    74      150     -741

%-740   -25     240     -741

%-330   -735    259     -741

%107    -735    19      -117

%-740   -735    264     -741


@attribute M79  string

%Width = 12

%Consensus = TGnsTGTGyGnG

%Log-Odds Matrix:

%-702   -698    100     98

%-291   -698    258     -703

%-702   -58     100     70

%-65    114     132     -703

%-702   -698    -196    147

%-702   -698    264     -703

%-291   -698    -697    147

%-97    -213    231     -703

%-702   99      -697    89

%-702   -698    264     -703

%-291   -698    100     89

%-702   82      208     -703


@attribute M80  string

%Width = 12

%Consensus = rAGAAGAAGAAG

%Log-Odds Matrix:

%85     -675    142     -679

%165    -675    -674    -679

%-678   -675    241     -125

%165    -675    -674    -679

%117    -33     -674    -125

%-678   -189    256     -679

%126    -675    25      -277

%143    -675    -16     -679
```

```
%-72    -675    224     -277

%158    -189    -674    -679

%135    -33     -674    -277

%-678   -675    263     -679


@attribute M81  string

%Width = 12

%Consensus = GGGnGnrGAnnG

%Log-Odds Matrix:

%-718   -713    264     -718

%-14    -713    215     -718

%-33    -713    222     -718

%-33    -75     200     -718

%-718   -713    264     -718

%3      98      100     -718

%44     -713    183     -718

%-718   -2      235     -718

%131    -713    42      -718

%3      -75     174     -318

%-81    83      129     -166

%-718   -2      235     -718


@attribute expr string


@data

osm-6, ’{}’, ’{}’, ’{28:35}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{

}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’

, ’{}’, ’{}’, ’{}’, ’{}’, ’{67:74}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’

{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{129:138

}’, ’{}’, ’{}’, ’{}’, ’{48:59}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’,

 ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’

{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’ASH^ASI^PHA^ADL^ASE^ASK’


unc-97, ’{}’, ’{2158:2165^1927:1934}’, ’{}’, ’{474:483}’, ’{}’, ’{2124:2133}’, ’

{}’, ’{2087:2098^349:360^331:342^90:101}’, ’{1149:1160}’, ’{}’, ’{}’, ’{}’, ’{}’

, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{2560:2569}’, ’{}’, ’{}’, ’{5

05:516}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{2524:2535^589:

600}’, ’{}’, ’{}’, ’{1090:1097}’, ’{}’, ’{}’, ’{1300:1309}’, ’{}’, ’{}’, ’{}’, ’

{}’, ’{}’, ’{}’, ’{1886:1893^665:672}’, ’{}’, ’{617:626}’, ’{1288:1297}’, ’{}’,
'{}', '{}', '{}', '{571:578}', '{}', '{1544:1553^416:425^141:150}', '{1106:1115}
', '{313:322}', '{}', '{160:171}', '{1022:1033}', '{}', '{}', '{991:998}', '{}',
 '{}', '{}', '{}', '{}', '{2540:2551^1060:1071}', '{74:81}', '{}', '{}', '{}', '
{}', '{}', '{}', '{}', '{2505:2516}', 'ALM'

flp-6, '{}', '{}', '{}', '{4198:4207}', '{513:522}', '{}', '{3836:3847^2013:2024
^1779:1790}', '{3555:3566}', '{3947:3958}', '{}', '{}', '{}', '{}', '{}', '{447:
456}', '{4533:4544^301:312^274:285}', '{3975:3986^469:480}', '{}', '{3231:3238^4
61:468}', '{}', '{}', '{}', '{730:739}', '{}', '{}', '{}', '{2568:2579}', '{}',
'{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{4167:4176^30
82:3091}', '{}', '{}', '{1889:1900}', '{2480:2491}', '{}', '{}', '{4724:4731^414
9:4156^713:720^320:327}', '{3881:3888^2121:2128^932:939}', '{3166:3175}', '{2377
:2386^499:508}', '{3120:3129^864:873}', '{4415:4426^2986:2997}', '{4284:4295^245
1:2462^1854:1865}', '{573:584}', '{}', '{}', '{1833:1840}', '{}', '{428:437}', '
{}', '{1166:1177}', '{2197:2208^1791:1802}', '{3694:3705}', '{}', '{3293:3300}',
 '{}', '{2100:2109}', '{}', '{}', '{1199:1210}', '{3935:3946^384:395}', '{}', '{
550:557}', '{}', '{2862:2869}', '{}', '{3901:3910}', '{}', '{}', '{526:537}', '{
}', 'ASE'

unc-32, '{}', '{}', '{155:162}', '{}', '{1255:1264}', '{}', '{18:29}', '{1061:10
72}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{1315:1326}', '{}',
 '{}', '{}', '{}', '{872:881}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}',
'{}', '{}', '{46:57}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}
', '{}', '{}', '{}', '{}', '{1243:1252^419:428^232:241}', '{1388:1397^853:862}',
 '{}', '{}', '{1483:1494^1014:1025}', '{498:509}', '{}', '{1275:1282^118:125}',
'{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{
}', '{}', '{}', '{}', '{187:194}', '{}', '{}', '{338:347}', '{598:607}', '{}', '
{386:397^165:176}', '{1423:1434}', 'ALM'

unc-86, '{4936:4943^4428:4435}', '{68:75}', '{}', '{1543:1552^1020:1029}', '{}',
 '{1824:1833^1172:1181}', '{}', '{754:765}', '{1838:1849}', '{}', '{}', '{}', '{
}', '{4694:4703}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{}', '{2663:
2672}', '{}', '{1905:1916^188:199}', '{1622:1633}', '{480:487}', '{}', '{1579:15
86^1504:1511}', '{}', '{1555:1564^1292:1301}', '{}', '{4656:4667}', '{286:297}',
 '{}', '{}', '{}', '{}', '{}', '{}', '{542:551^527:536}', '{}', '{}', '{842:853}
', '{787:794}', '{821:828}', '{2095:2102^1405:1412}', '{1969:1978}', '{4683:4692
^4512:4521^2802:2811^1871:1880^1476:1485^714:723}', '{}', '{2643:2654^901:912}',
 '{957:968^804:815^48:59}', '{4618:4629}', '{}', '{}', '{}', '{3529:3538}', '{}'
, '{}', '{2601:2612}', '{3704:3715}', '{}', '{}', '{3843:3850}', '{}', '{}', '{}
```

```
’, ’{}’, ’{}’, ’{}’, ’{4461:4472^1733:1744^730:741}’, ’{3670:3677^695:702}’, ’{}
’, ’{4439:4446}’, ’{}’, ’{3750:3759^866:875}’, ’{}’, ’{}’, ’{}’, ’{4597:4608}’,
’ALM^HSN’

unc-103, ’{2159:2166}’, ’{}’, ’{3055:3062}’, ’{1165:1174}’, ’{4626:4635^669:678}
’, ’{}’, ’{3910:3921^3694:3705^1213:1224}’, ’{575:586}’, ’{3637:3648^3093:3104}’
, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{3393:3402^1451:1460^369:378^75:84}’, ’{2045:20
56^1856:1867^904:915}’, ’{}’, ’{}’, ’{}’, ’{3489:3496}’, ’{2721:2728}’, ’{2969:2
978}’, ’{2461:2470^265:274}’, ’{}’, ’{761:772^11:22}’, ’{2772:2783^2638:2649^91:
102}’, ’{}’, ’{}’, ’{}’, ’{3252:3259^3220:3227^857:864}’, ’{2194:2203}’, ’{2979:
2988^1364:1373}’, ’{}’, ’{}’, ’{2064:2075^938:949^282:293}’, ’{2104:2115}’, ’{}’
, ’{}’, ’{}’, ’{}’, ’{4706:4715^2789:2798}’, ’{4953:4962^2218:2227^1135:1144^104
5:1054}’, ’{2923:2934^125:136}’, ’{3679:3690^3518:3529^2748:2759^2558:2569}’, ’{
4531:4542^4507:4518^4393:4404^3110:3121^2606:2617}’, ’{}’, ’{3042:3049}’, ’{3983
:3990^3624:3631^3369:3376^3327:3334^1929:1936}’, ’{3752:3761^1870:1879^1375:1384
^700:709}’, ’{2007:2016^1277:1286}’, ’{}’, ’{3063:3074^2026:2037^1073:1084^839:8
50}’, ’{3781:3792^3354:3365^3145:3156^3010:3021}’, ’{4231:4242^3764:3775^387:398
}’, ’{4378:4385}’, ’{235:242}’, ’{4315:4322^3806:3813^2096:2103}’, ’{4466:4475^4
454:4463^187:196}’, ’{4117:4126}’, ’{684:693}’, ’{145:156}’, ’{503:514}’, ’{3960
:3971^2544:2555^2495:2506^1352:1363}’, ’{}’, ’{}’, ’{}’, ’{2626:2635^829:838}’,
’{2711:2720}’, ’{419:428}’, ’{}’, ’{}’, ’{2863:2874}’, ’{4486:4493^1945:1952}’,
’{}’, ’{}’, ’{199:208}’, ’{}’, ’{}’, ’{488:499}’, ’{}’, ’{}’, ’HSN^PHA^ADL^ASK’

unc-129, ’{1022:1029^523:530^277:284}’, ’{3597:3604^1177:1184}’, ’{}’, ’{2711:27
20^1954:1963}’, ’{2758:2767}’, ’{654:663}’, ’{3216:3227^18:29}’, ’{1707:1718^565
:576}’, ’{3525:3536}’, ’{}’, ’{}’, ’{}’, ’{99:108}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}
’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{4219:4228}’, ’{}’, ’{3370:3381}’, ’{}’, ’{178
7:1794}’, ’{}’, ’{3537:3544}’, ’{3193:3202^1214:1223}’, ’{}’, ’{880:889}’, ’{}’,
 ’{4844:4855^2931:2942^1971:1982}’, ’{417:428}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{4455:
4464^3734:3743}’, ’{}’, ’{4192:4203}’, ’{2291:2302}’, ’{}’, ’{3098:3105}’, ’{}’,
 ’{300:307}’, ’{}’, ’{1765:1774}’, ’{}’, ’{4622:4633^3000:3011^690:701}’, ’{4670
:4681^853:864^824:835^197:208}’, ’{3553:3564}’, ’{}’, ’{792:799}’, ’{3481:3488}’
, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{4010:4021^2786:2797^134:145}’, ’{}’, ’{}’, ’{}
’, ’{}’, ’{}’, ’{}’, ’{}’, ’{1828:1839^1668:1679}’, ’{}’, ’{}’, ’{3859:3866}’, ’
{}’, ’{}’, ’{1857:1866}’, ’{}’, ’{2851:2862}’, ’{3884:3895^3625:3636}’, ’{6:17}’
, ’CAN’

tax-6, ’{}’, ’{749:756^337:344}’, ’{}’, ’{262:271}’, ’{419:428}’, ’{}’, ’{658:66
9}’, ’{703:714}’, ’{}’, ’{}’, ’{}’, ’{120:127}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’
```

```
{}’, ’{}’, ’{}’, ’{786:793}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{83:94}’, ’{372:383}’, ’{
}’, ’{1283:1290}’, ’{200:207}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{1041:1052}’, ’{630:641
^407:418^287:298}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{167:178}’, ’{}’, ’{771
:782}’, ’{470:477}’, ’{497:504}’, ’{}’, ’{214:223}’, ’{}’, ’{}’, ’{}’, ’{1006:10
17}’, ’{}’, ’{834:841}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{431:442}’, ’{}’,
’{}’, ’{110:117}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’
{}’, ’{805:814}’, ’{}’, ’{}’, ’{}’, ’{}’, ’ASH^ASI^PHA^ADL^ASE^ASK’

unc-76, ’{1201:1208}’, ’{910:917}’, ’{}’, ’{439:448^159:168}’, ’{}’, ’{}’, ’{167
3:1684^1159:1170^953:964^921:932}’, ’{2049:2060^1837:1848^510:521}’, ’{693:704^5
98:609^423:434}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{724:733}’, ’{}’, ’{}’, ’{}’, ’{}’, ’
{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{1055:1064}’, ’{681:692^410:421}’, ’{}’, ’{}’, ’{}
’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{1875:1886}’, ’{}’, ’{}’, ’{}’, ’{
}’, ’{}’, ’{}’, ’{1719:1730}’, ’{2134:2145^586:597}’, ’{934:945^225:236^73:84}’,
 ’{}’, ’{1646:1653}’, ’{}’, ’{}’, ’{102:111}’, ’{}’, ’{}’, ’{1983:1994}’, ’{1783
:1794^52:63^22:33}’, ’{}’, ’{}’, ’{1932:1939}’, ’{179:188}’, ’{876:885}’, ’{}’,
’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{1295:1304}’, ’{}’, ’{}’, ’{}’, ’{114
:125}’, ’{}’, ’{818:825^272:279}’, ’{493:500}’, ’{547:556}’, ’{333:342^143:152}’
, ’{480:489}’, ’{641:652}’, ’{312:323^240:251}’, ’{}’, ’HSN^CAN’
ttx-3, ’{}’, ’{1646:1653}’, ’{}’, ’{}’, ’{1803:1812^51:60}’, ’{}’, ’{1313:1324}’
, ’{2352:2363^1450:1461}’, ’{}’, ’{}’, ’{}’, ’{4419:4426}’, ’{}’, ’{3890:3899}’,
 ’{3977:3986^3128:3137^1996:2005}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{
}’, ’{4394:4403}’, ’{}’, ’{2322:2333}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’
, ’{}’, ’{}’, ’{1528:1539}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{3643:3654}’,
’{}’, ’{}’, ’{1517:1524}’, ’{4525:4532}’, ’{3287:3294}’, ’{1205:1214}’, ’{452:46
1}’, ’{}’, ’{1486:1497}’, ’{3556:3567}’, ’{1172:1183}’, ’{}’, ’{}’, ’{}’, ’{1499
:1508}’, ’{}’, ’{2525:2534^701:710}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’,
’{}’, ’{}’, ’{}’, ’{}’, ’{1558:1569}’, ’{143:150}’, ’{3268:3275}’, ’{4483:4490}’
, ’{}’, ’{1573:1582^1112:1121}’, ’{}’, ’{}’, ’{}’, ’{4428:4439}’, ’ASI^ADL’

unc-73, ’{3306:3313^922:929^459:466}’, ’{}’, ’{}’, ’{}’, ’{1003:1012}’, ’{1250:1
259}’, ’{1548:1559^1491:1502}’, ’{4088:4099}’, ’{4164:4175^3751:3762^3504:3515}’
, ’{}’, ’{}’, ’{3819:3826}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{3958:3969}’, ’{}’,
’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{4137:4148^3939:3950^3221:3232}’, ’{}’, ’{3909:39
20}’, ’{}’, ’{}’, ’{3861:3868^2236:2243}’, ’{243:252}’, ’{3871:3880}’, ’{}’, ’{}
’, ’{1072:1083}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{}’, ’{2538:2547^1165:1174}’, ’
{}’, ’{}’, ’{2388:2399^1137:1148}’, ’{3790:3797}’, ’{4045:4052}’, ’{3019:3026^16
6:173}’, ’{}’, ’{2674:2683}’, ’{}’, ’{2272:2283^842:853}’, ’{}’, ’{3802:3813^644
:655}’, ’{}’, ’{}’, ’{}’, ’{3833:3842^1094:1103^938:947}’, ’{}’, ’{1742:1751}’,
'{}', '{}', '{854:865^192:203}', '{}', '{}', '{}', '{}', '{}', '{4209:4218}', '{
2074:2085}', '{}', '{442:453}', '{4111:4118^1153:1160^907:914}', '{}', '{}', '{}
', '{1025:1034}', '{}', '{69:80}', '{}', '{4008:4019}', 'ALM^HSN^CAN'
```

```
osm-9, '{3944:3951^2699:2706^1460:1467}', '{3545:3552^2368:2375}', '{}', '{203:2
12}', '{4121:4130}', '{}', '{4836:4847^2925:2936}', '{}', '{}', '{}', '{}', '{}'
, '{}', '{}', '{}', '{}', '{}', '{2452:2463^573:584}', '{}', '{}', '{}', '{}', '
{682:691}', '{}', '{2207:2218}', '{}', '{3211:3222^2227:2238}', '{}', '{}', '{}'
, '{}', '{}', '{}', '{}', '{855:866}', '{3952:3963}', '{}', '{1379:1386}', '{}',
 '{}', '{2270:2279^109:118}', '{}', '{2319:2330}', '{2153:2164}', '{2800:2811}',
 '{1239:1246}', '{}', '{4616:4623^3526:3533^1538:1545^552:559}', '{}', '{3597:36
06}', '{}', '{3357:3368^1224:1235}', '{3831:3842}', '{693:704^406:417}', '{387:3
94}', '{}', '{}', '{2335:2344^71:80}', '{2349:2358^288:297}', '{4053:4062^3573:3
582}', '{299:310}', '{3260:3271^2569:2580^366:377}', '{}', '{}', '{2428:2435}',
'{}', '{}', '{}', '{}', '{}', '{}', '{4767:4778}', '{2625:2632}', '{2300:2307}',
 '{}', '{2387:2396}', '{}', '{3620:3629^1714:1723}', '{}', '{2597:2608}', '{}',
'ASH^ASI^PHA^ADL^ASE^ASK'
```

Appendix B

Perl script to convert probability matrix to log odds matrix.

```
#!/usr/local/bin/perl -w
use strict;
use Math::Complex;

#
#

my @args = @ARGV;

if(scalar(@args) != 3){
        print "\n The # of command line arguments supplied not correct";
        print "\nUsage: convert <Prob Matrix motif file> <background Probability File> <output file name>\n";
        exit(0);
}


my $probMatrix = shift @args;
my $bgFrequencyFile= shift @args;
my $outputFile= shift @args;

sub slurp {
        local $/ = undef;
        open my $fh, $_[0] or die "Can’t open $_[0]: $!";
        my $slurp = <$fh>;
        return \$slurp;
}

# Open Gene Names file
open(PROB, $probMatrix) || &return_error("File Error","Unable to open " . $probMatrix . ". Reason $!");

# Open Sequence File
#open(BG, $bgFrequencyFile) || &return_error("File Error","Unable to open " . $bgFrequencyFile . ". Reason $!");

# Open Gene Names file
open(FD, "> $outputFile") || &return_error("File Error","Unable to open " . $outputFile . ". Reason $!");

my @bg;
my $bgline = slurp($bgFrequencyFile);
#foreach $bgline (<BG>){
  @bg = trim($$bgline) =~ /([\d]+\.[\d]+)\s+([\d]+\.[\d]+)\s+([\d]+\.[\d]+)\s+([\d]+\.[\d]+)/;
#print $bgline;
#last;
#}

print join(@bg);

my @parts;

foreach my $line(<PROB>){
        #if($line =~ /^#/){
        #       print FD $line;
        #       next;
        #}

        if(@parts = $line =~ /([\d]+\.[\d]+)\s+([\d]+\.[\d]+)\s+([\d]+\.[\d]+)\s+([\d]+\.[\d]+)/){
                for(my $i=0; $i<4; $i++){
                        print FD round(logn( $parts[$i]/$bg[$i], 2) * 100);
                        print FD "\t";
                }
                print FD "\n";
        }else{
                print FD $line;
                next;
        }

}


sub trim {
         my @out = @_;
         for (@out) {
                 s/^\s+//;
                 s/\s+$//;
         }
         return wantarray ? @out : $out[0];
}

sub round {
    my($number) = shift;
    return int($number + .5 * ($number <=> 0));
}
```

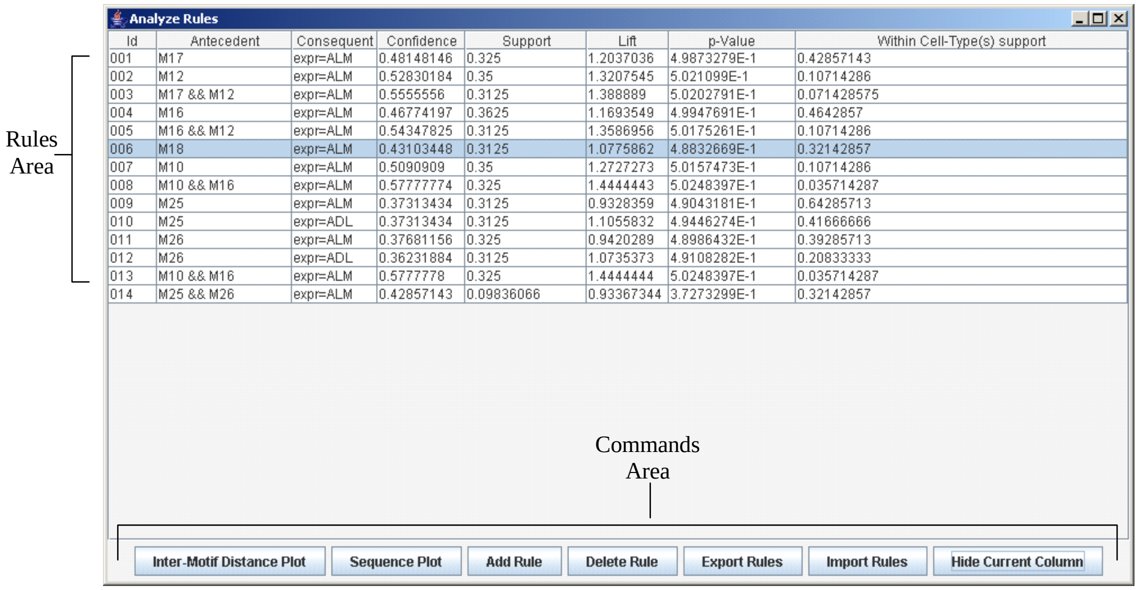
Appendix C

User Guide

C.1 Analysis Frame

The visualization and specialization modules (VSM) could either be invoked from the mining interface of WPI-Weka system or could be invoked as a standalone application using exported set of mined association rules, the associated MAST results (HTML format) and a list of gene names alongside the known expression patterns. The primary interface of the VSM is the Analysis frame, that is the first screen to be displayed when VSM is invoked. See Figure C.1. We explain the Analysis frame below.

The analysis frame loads with two sections, the **Rules** area and the **Commands** area as shown in Figure C.1. We explain the Rules area below and we explain each of the options in the Commands area in subsequent subsections. The Rules area is used to display base association rules along with the corresponding values for certain measures of interestingness. The design is extensible, that is, new measures of interestingness could be added in the future with minimal code changes. In the current state, a rule tuple consists of the following items:



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.42857143
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.10714286
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.071428575
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.4642857
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.10714286
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.32142857
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.10714286
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.035714287
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.64285713
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.41666666
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.39285713
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.20833333
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.035714287
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.32142857

Figure C.1: Sample Analysis Frame.

- **Id** - this is a unique id assigned to the rule. The usability of this field increases once the user starts to generate specializations from the rule, as the Id column helps us trace the history or the specialization path of new rules.

- **Antecedent** - The left-hand side of the rule. It contains the motifs present in the rule.

- **Consequent** - The right-hand side of the rule. It contains the cell-types predicted by the rule.

- **Support** - As discussed in Section 1.2, Rule 1.1; the support of the rule is one of the popular scales to measure the interestingness of the rule

- **Confidence** - Confidence is the other popular metric to measure the interestingness of the rule.

- **Lift** - The lift value of an association rule is another measure to try to quantify the interestingness of the rule. It is defined as the ratio of the confidence of the rule and the support of the consequent of the rule [BMS97]. In other words

$$lift\left(rule\right) = p\left(consequent|antecedent\right)/p\left(consequent\right)$$

- **p-Value** - The p-value of the rule is the probability that the correlation between the antecedent and the consequent is due to chance by using the chi square test.
- **Within Cell-Type Support** - Provides the support of the rule among only those instances of the data that contain the consequent of the rule.

The Commands area of the Analysis frame provides buttons to perform a range of functions. Each of the following subsections describe each of these functions provided by the visualization extensions via the analysis frame. It is important to note that most of these functions are invoked in the context of a specific rule and so it is necessary to select a rule in the rules area of the analysis frame before we invoke a command.

C.2 Inter-Motif Distance Plot

Selecting a Rule in the rules area and then invoking the inter-motif distance plot via the button with the same label, lets a user visualize the data in the context of the rule from a inter-motif distance perspective. This action enables a user to perform exploratory analysis in the context of the hypothesis - “Inter-motif distance influences gene expression”.

On invoking this command a new frame with the pairwise inter-motif distance plot(s) is displayed. It displays one graph for each pair of motifs in the rule (selected

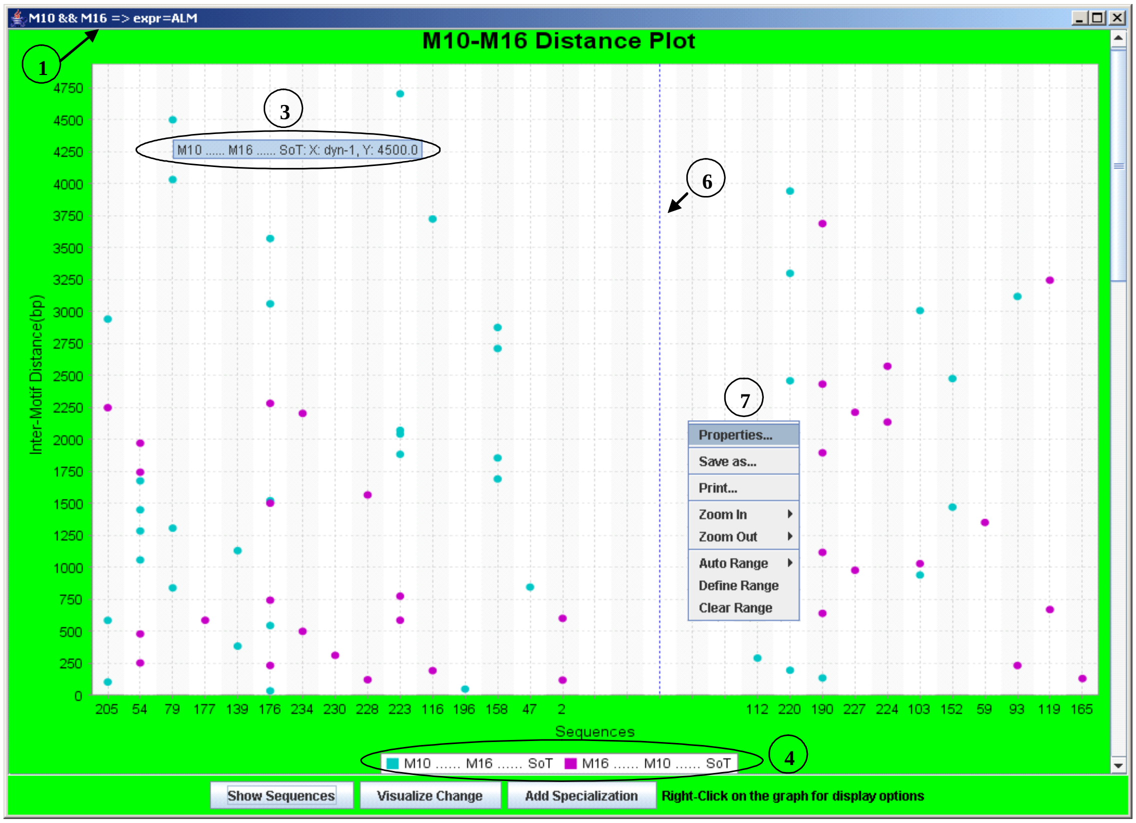


Figure C.2: Sample Inter-Motif Distance Frame - The numeric pointers refer to the enumerated text explaining Inter-Motif Distance Plot.

in the Analysis Frame). For sake of simplicity we start with a rule with only two motifs and we revisit plots originating from rules consisting of more than two motifs in item later in the section. We enlist below the highlights of the information displayed in an inter-motif distance plot and the numeric annotations in Figure C.2 are references to the information in the following enumeration.

1. Each graph is displayed with the rule used to establish the context as the title of the frame.

2. Each graph displays the pairwise inter-motif distance plots. For instance, in Figure C.2 the rule consists of motifs M10 and M16. So for each instance of M10 on a gene, a distance value from every instance of M16 is computed. Each

such value corresponds to a point on the graph with distance being plotted on the y-axis and the x-axis is an id for the gene in question.

3. Rolling your mouse over any such data point displays the relevant gene name and the distance computed.

4. Each graph could potentially contain points in two colours. The legend explains the difference. Each point is a distance of an instance of M10 from an instance of M16 but the color helps identify the order in which these motif instances occur in the gene sequence relative to the Start of Transcription(SoT).

5. Each graph lists only those genes on the X-axis, that support the antecedent of the rule. That is, genes whose promoter regions contain at least one instance of both motifs.

6. Another mechanism to aid visual exploration is that each graph is sliced into two parts by a dotted line. The genes in the left part are the ones that support the consequent of the rule and hence support the rule. The ones on the right are the genes that only support the antecedent of the rule. This provides the user with an easy mechanism to discover inter-motif distance based patterns on the left part of the plot that are not as frequent on the right part as this would let us explore specializations with improved classification accuracy(and/or confidence).

7. We used the charting library, JFreeChart [JFr], as the charting infrastructure for the visualization extensions. This was an obvious choice because it was an open source, well-documented API, supporting a wide range of chart types with a flexible design that is easy to extend. A right-click on the graph area displays a popup menu with the following options:

- **Zoom** - Lets you zoom in or out on the graph.
- **Range** - Lets the user change the scale of either axis.
- **Save** - Lets the user save an interesting graph as an image.
- **Define Range** - This is an important extension we made to the charting infrastructure that lets the user define, in a graphical fashion, the inter-motif distance to be used in the specialization. An inter-motif distance based specialization places conditions on the distance between instances of the motifs involved. For instance, it could be worth noting that there are significantly more data points with an inter-motif distance value between 0 and 500 in the left part of the Figure C.2 as compared to the same distance range in the right-hand side. The definition of this distance condition is in the form of a range, for instance, *(0-500)*. Selecting this option from the menu changes the graph to a range define mode and a subsequent click and drag can be used to define the range.
- **Clear Range** - Another extension to the charting infrastructure which lets a user clear a currently defined range providing an option to redefine a range.

8. Once the user has utilized the dotted separation and inter-motif distance plots to identify a range of interest, for instance a range of (0-500) between motifs M10 and M16, the “Define Range” option from the pop-up menu can be used to graphically define this range. Once a range has been defined for a plot, it is highlighted on the graph in a shade of gray. For instance, refer to Figure C.3 numerical annotation 8.

9. Even after a range has been defined the data being visualized is in the context of the original rule. At this point the user can invoke the “Visualize Change”

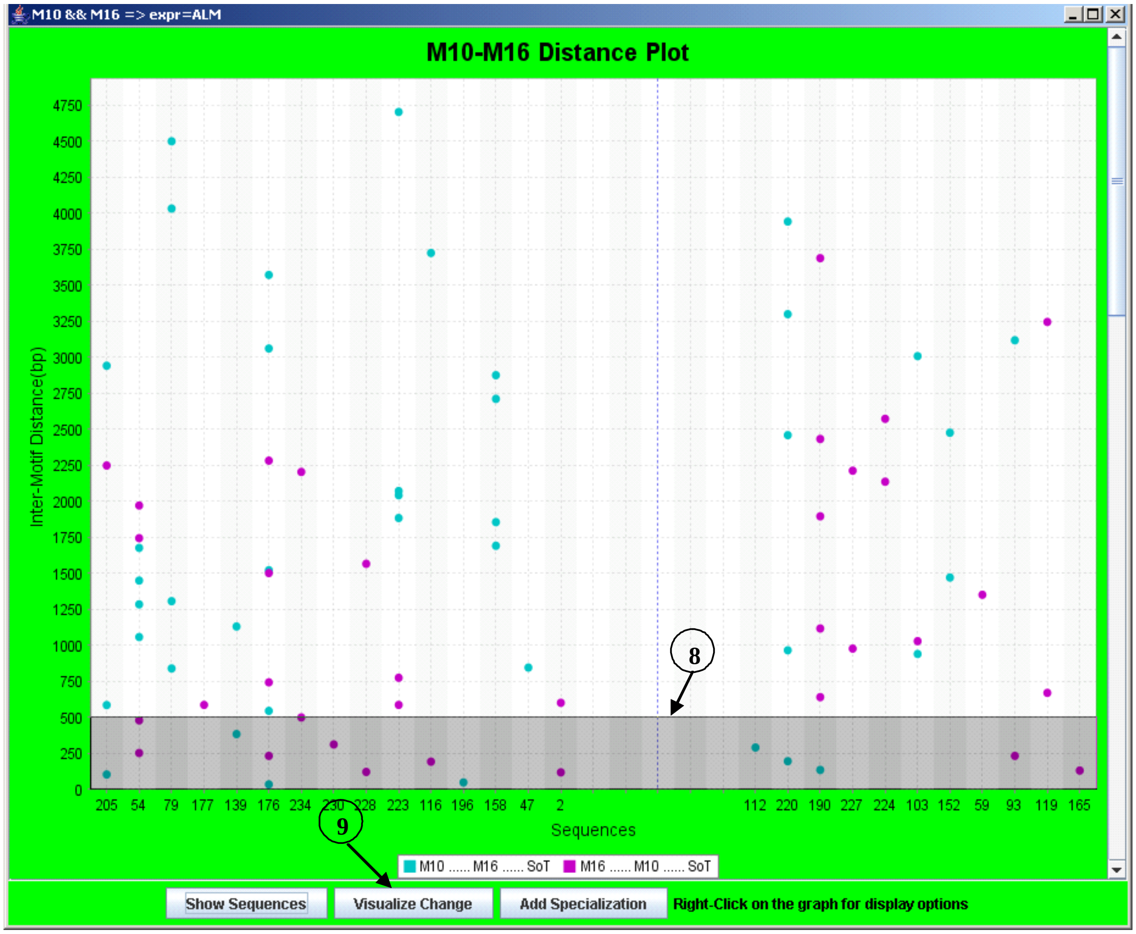


Figure C.3: Inter-Motif Distance Frame depicting an inter-motif range of 0-500 bp for Motif pair M10 and M16.

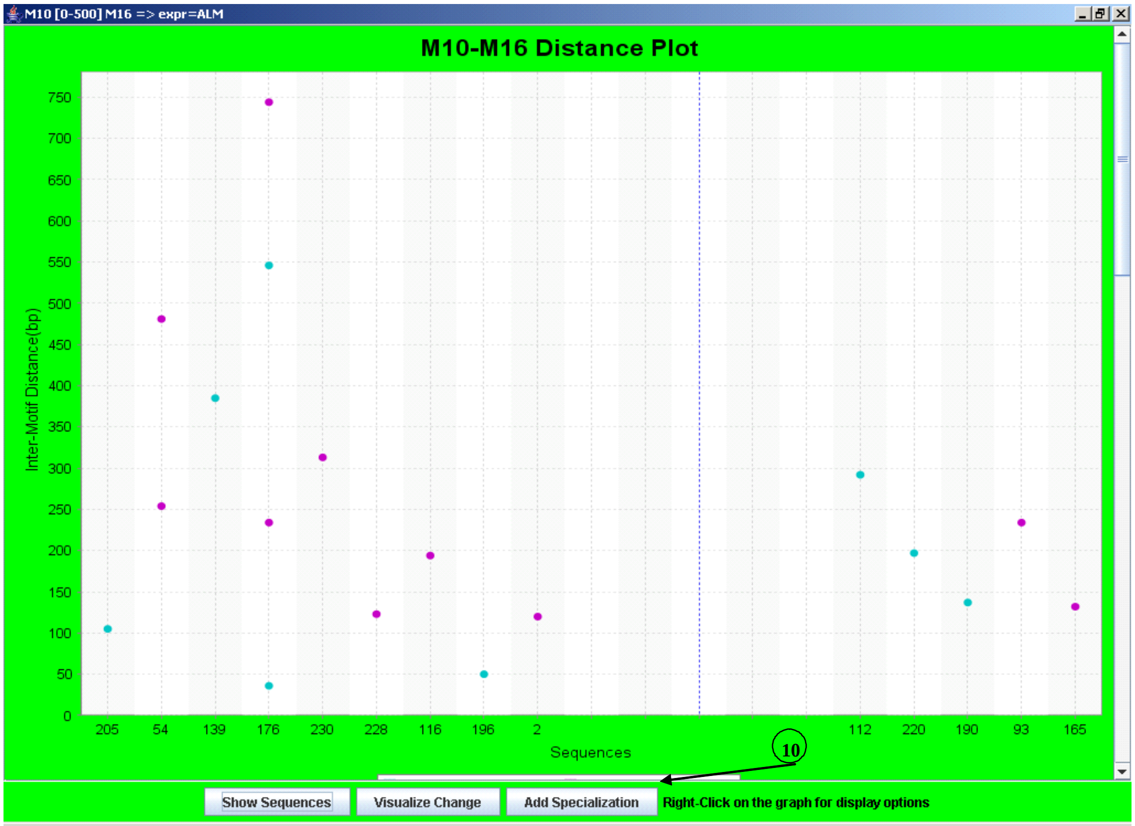
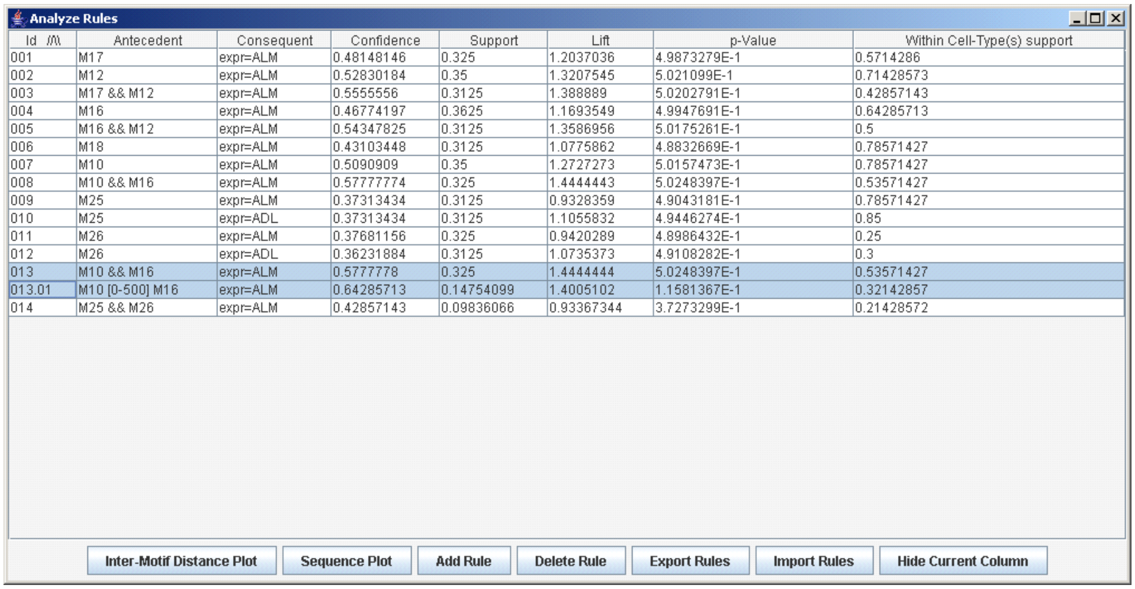


Figure C.4: Visualize command invoked inter-Motif distance plot displaying data in the context of the specialised rule M10 (0-500) M16 ⇒ expr = ALM.



Id /A\	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.78571427
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.53571427
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.78571427
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.85
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.25
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.3
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.53571427
013.01	M10 [0-500] M16	expr=ALM	0.64285713	0.14754099	1.4005102	1.1581367E-1	0.32142857
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572

Figure C.5: A row representing the addition of a specialization to the Analysis Frame.

(Figure C.3 numerical annotation 9) command to visualize the data in the context of the specialization rather than in the context of the original rule. For instance Figure C.4. It is worth noting that the title of the new window is indicating the new context.

10. If the user finds this specialization (Figure C.4)of interest, the specialized rule can be added to the Analysis Frame using the "Add Specialization" command on the Inter-Motif distance plot. This causes a new entry to be inserted in the Analysis Frame with the following specialization

$$M10(0-500)M16 \Rightarrow expr = ALM \tag{C.1}$$

as shown in Figure C.5. Note that the Id field is auto-generated in a fashion that always lets a user trace back the steps in case we want to later recall which rule was used to derive the specialization. Also note the different measures of interestingness are computed for the specialized rule.

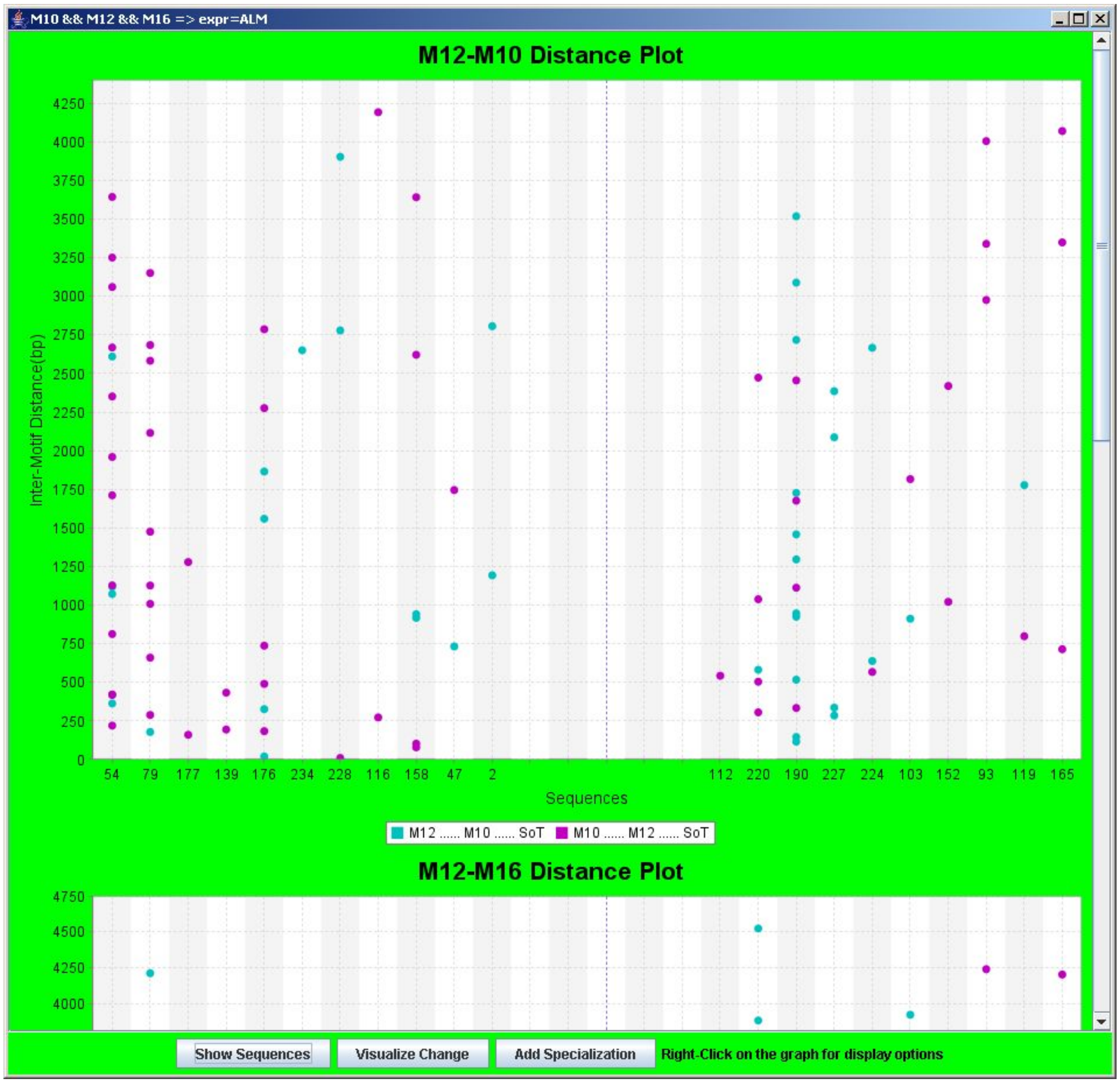
11. If the base rule consisted of more than two motifs, an inter-motif distance plot for each pair of motifs is displayed in the same frame (Figure C.6). Each chart or plot individually provides for defining relationships between a pair of motifs. In case multiple relationships are defined for more than one pair of motifs each motif is represented as a term and a collection of independent terms constitutes the specialized rule. Mechanism to define advanced relationship's between each term is also provided and we would revisit the topic later.

C.2.1 Sequence Plot

Select a Rule in the rules area and then click the sequence plot button to visualize all the qualifying gene sequences in the context of the rule. This action enables a user to perform exploratory analysis in the context of the hypothesis - "Distance of motifs from the SoT influence gene expression". A qualifying gene sequence is one that has that has at least one instance of each motif that appears in the rule (selected in the Analysis Frame). Invoking this command causes a new frame with the sequence plot overlaid with the motif information to be displayed. We enlist below the highlights of the information displayed in an inter-motif distance plot and the numeric annotations in Figure C.7 are references to the information in the following enumeration.

1. Each graph is displayed with the rule used to establish the context as the title of the frame.

2. Displays the gene sequence plots with motif instances in the context of the rule. For example, in Figure C.7 the sequence plot displays all relevant gene



Right-Click on the graph for display options

Figure C.6: Multiple pairwise inter-motif distance plots for a base rule with more than two motifs. For instance the base rule for this plot is M10 && M12 && M16 ⇒ expr = ALM.

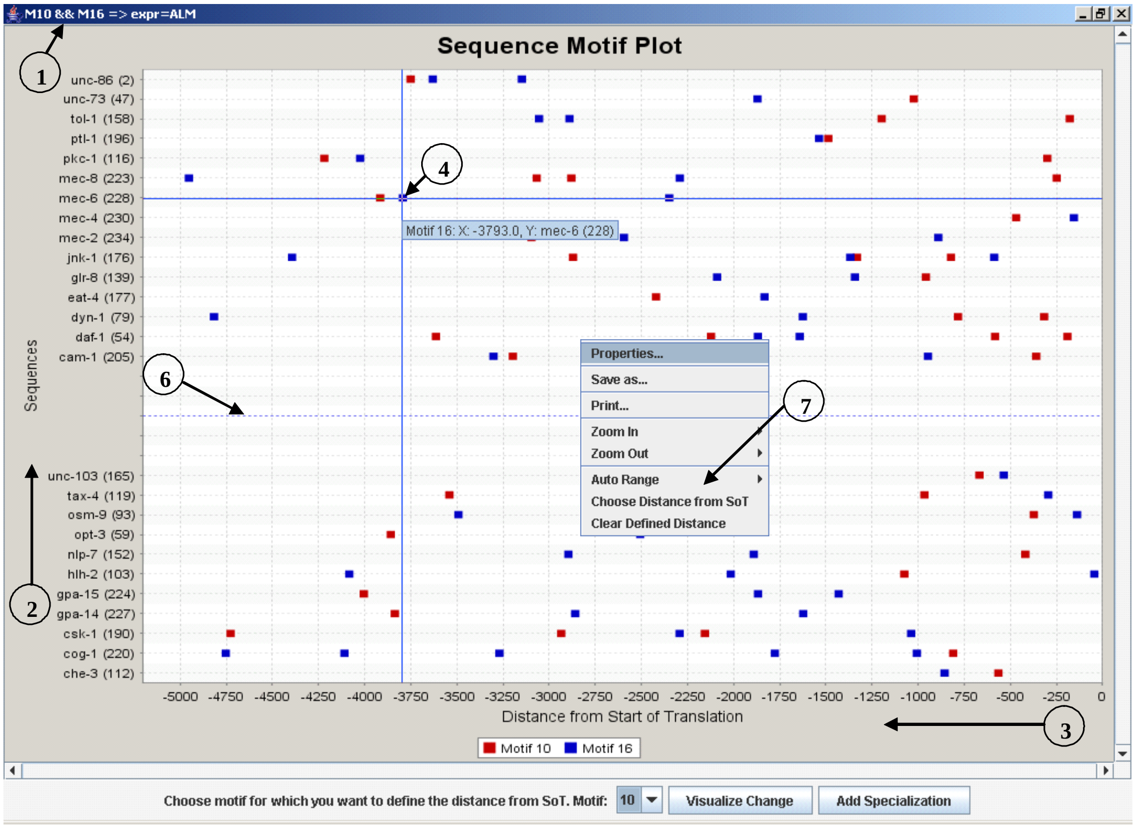


Figure C.7: Sample Sequence Plot Frame - The numeric pointers refer to the enumerated text explaining the Sequence Plot.

sequences with the instances of participating motifs (i.e., motifs M10 and M16). So along the y-axis is the list of qualifying gene promoters.

3. For each such gene sequence we plot all the instances of the participating motifs as they exist on the gene sequence relative to the Start of Transcription(SoT), which is the far right end of the plot. This makes the x-coordinate of each point in the plot the distance of the motif from the SoT and the color of the point is used to identify the motif.

4. Rolling your mouse over any such point displays the relevant gene name and the distance of the instance of the motif from the SoT.

5. Lists only qualifying genes on the Y-axis, that support the antecedent of the

rule. That is, the gene sequence has at least one instance of both motifs.

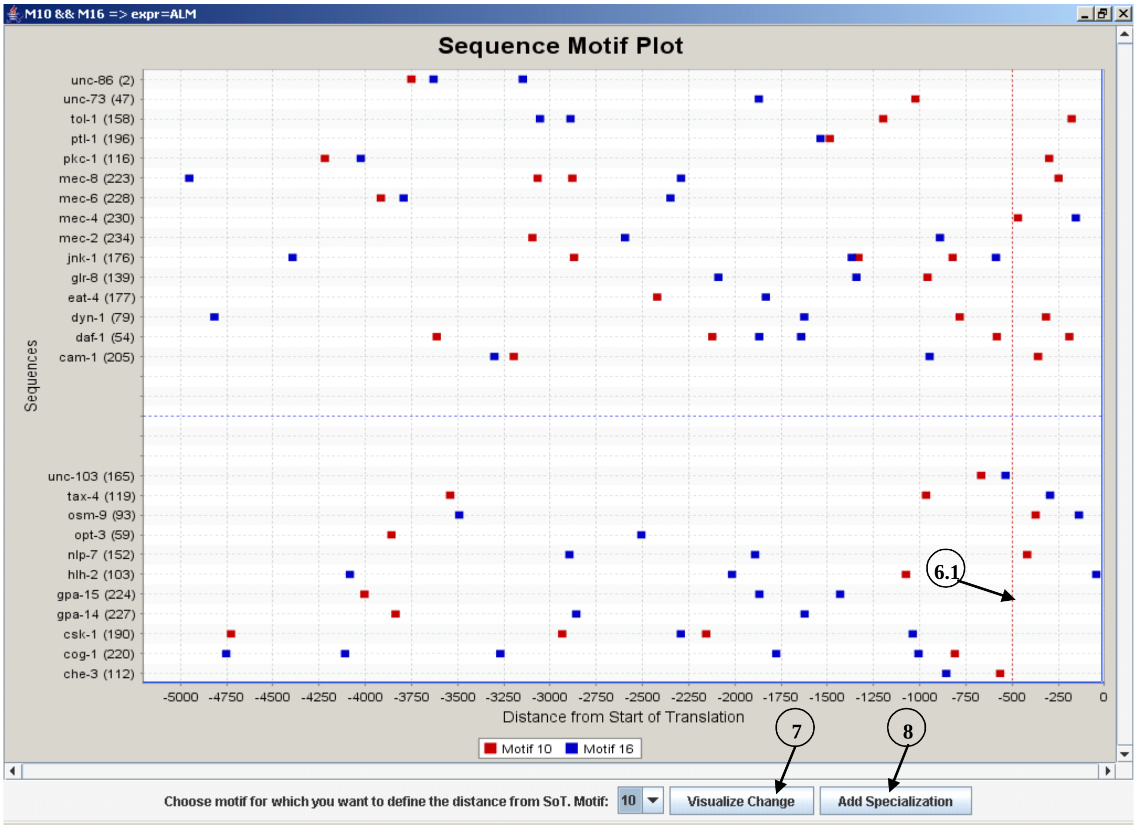
6. Another mechanism to aid visual exploration is that each graph is sliced into two parts by a horizontal dotted line. The genes in the upper part are the ones that support the consequent of the rule and hence support the rule. The ones in the lower part are the genes that support only the antecedents of the rule. This provides the user with an easy mechanism to discover "Distance from SoT" based patterns in the upper part of the plot that are not as frequent in the lower part as this would let us explore specializations with improved classification accuracy(and/or confidence).

7. The following context-specific options were added to the graph right-click popup menu in the charting infrastructure:

    - **Choose Distance from SoT** - This extension lets the user choose the "distance of a motif from the SoT" clause-based specialization in a graphical fashion. Selecting this option from the menu changes the graph to a distance selection mode. A subsequent click can be used to define the chosen value for the "distance from the SoT" and a visual confirmation of the defined distance clause is provided in the form of a vertical dotted line in the same color as the one reserved for the motif. One such distance can be defined for each participating motif. For example, in Figure C.8 a distance term of *SoT [0-500] M10* is chosen.

    - **Clear Distance from SoT** - This option lets a user clear all currently defined distances from SoT.

8. Once the user has utilized the dotted separation of the plot (into rule supporting and antecedent supporting) and the rule specific sequence plots to identify

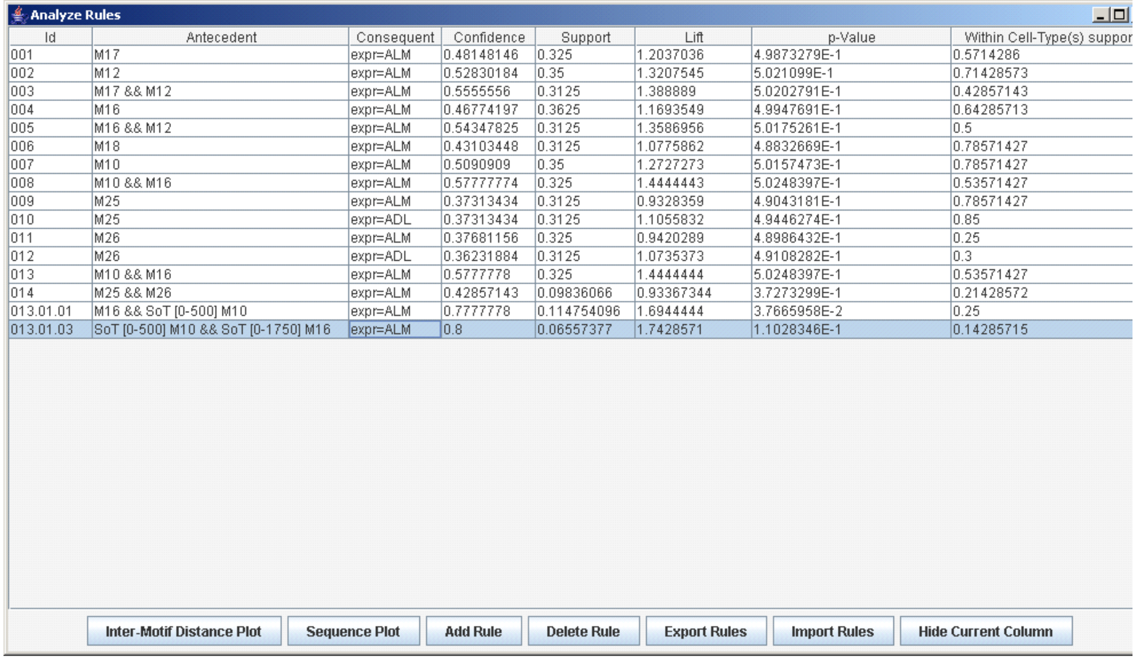


Choose motif for which you want to define the distance from SoT. Motif: 10

Figure C.8: Sequence Plot Frame depicting a distance of 500 bp from SoT for M10.

a “distance from SoT” clause of interest, the “Visualize Change” command could be invoked to visualize the data in the context of the specialization rather than the original rule. Again the title of the new window is indicating the context setting rule/specialization.

9. If the user finds the specialization of interest, it can be added to the Analysis Frame using the “Add Specialization” command on the new sequence plot. Again this causes the specialization to appear as a new entry in the Analysis Frame with an auto-generated Id that again lets a user trace back the steps in case the user wants to later recall which rule was used to derive a specialization (Figure C.9).



Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) suppor
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.78571427
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.53571427
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.78571427
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.85
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.25
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.3
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.53571427
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572
013.01.01	M16 && SoT [0-500] M10	expr=ALM	0.7777778	0.114754096	1.6944444	3.7665958E-2	0.25
013.01.03	SoT [0-500] M10 && SoT [0-1750] M16	expr=ALM	0.8	0.06557377	1.7428571	1.1028346E-1	0.14285715

Figure C.9: Analysis Frame depicting a couple of newly added distance from SoT based specializations.

10. In case the base rule consisted of more than one motif and multiple "distances from SoT" relationships are defined (one for each motif) each such relationship is represented as a term and a collection of independent terms constitutes the specialized rule. For instance see Figure C.9 for the following specialization.

$$SoT\,[0-500]\,M10 \quad \&\& \quad SoT\,[0-1750]\,M16 \Rightarrow expr = ALM \qquad \text{(C.2)}$$

C.2.2 Order of occurrence of motifs

We wanted the VSM to facilitate exploratory analysis based on the hypothesis "Order of occurrence of motifs influences gene expression". But during the system design and the system use by the team(including the domain expert) it was observed that we already had a few ways to visualize gene sequence data in the context of the "order of the occurrence" of motifs. If order of occurrence of motifs was important

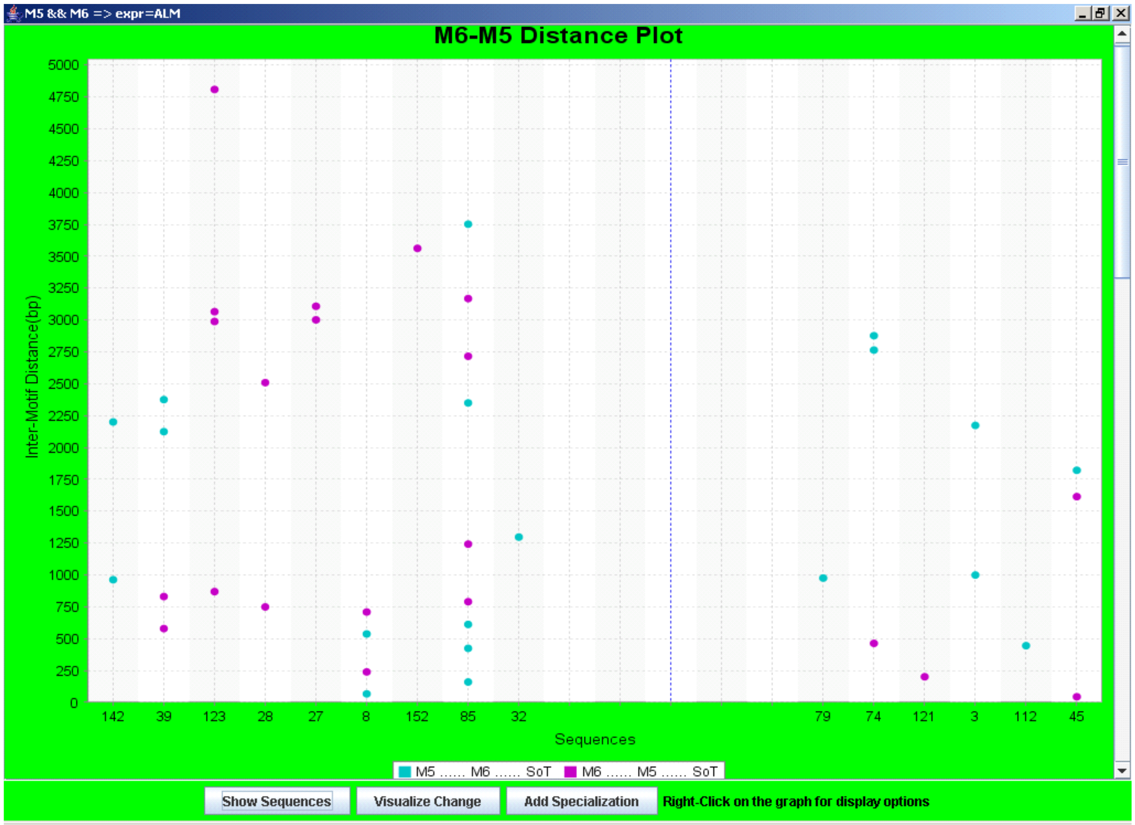
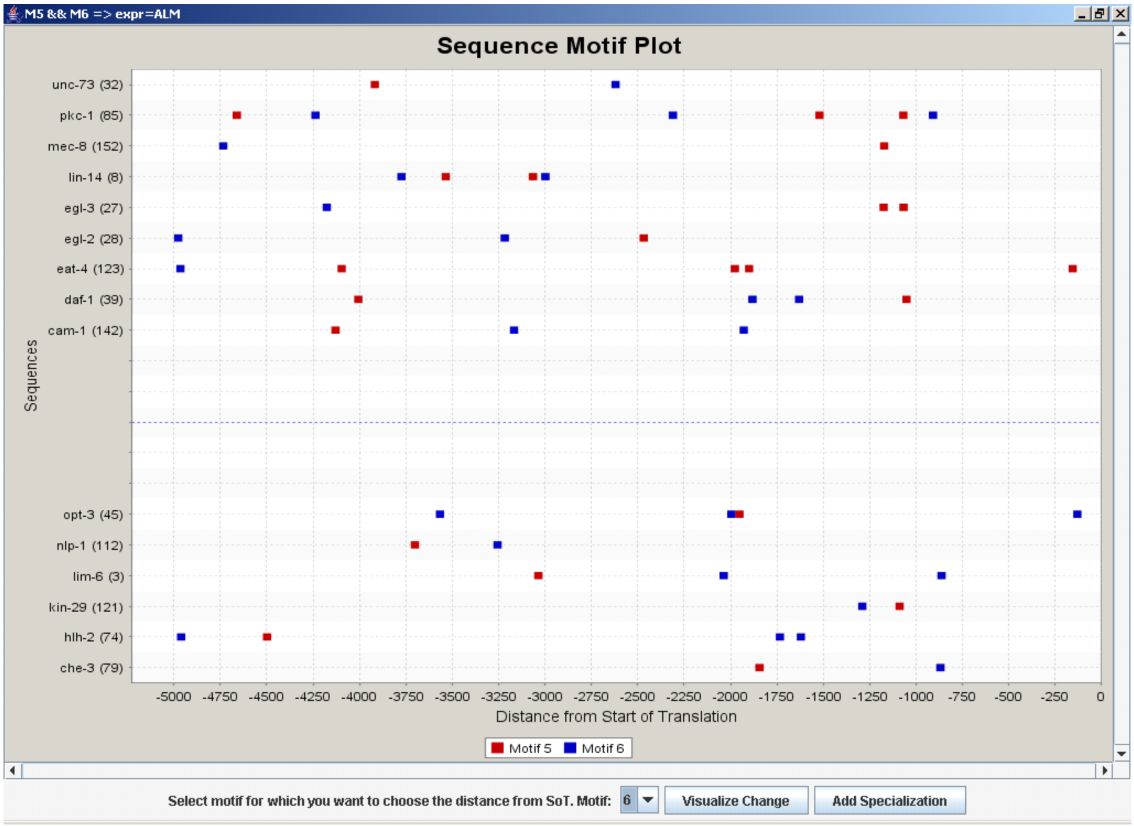


Figure C.10: Inter-Motif Distance Plot for Motifs M5 and M6. Observe the lack of magenta (dark) dots in the right half of the frame.

it could be easily identified by one of the following ways:

1. **Color of the points in inter-motif distance plot.** The order of the motifs in the inter-motif distance plot is represented by color. For instance in Figure C.2, M10 to the right of M16 (i.e., M16..M10..SoT) is represented by a magenta (dark) dot, while M16 to the right of M10 is represented by a aqua (light) dot. Thus, color provides a quick visual clue whether the order of occurrence of motifs affects gene expression; the left half of the plot should have more point of one color than the other in this case.

2. **Repeating sequence of color in the sequence plot** - As mentioned in



Select motif for which you want to choose the distance from SoT. Motif: 6

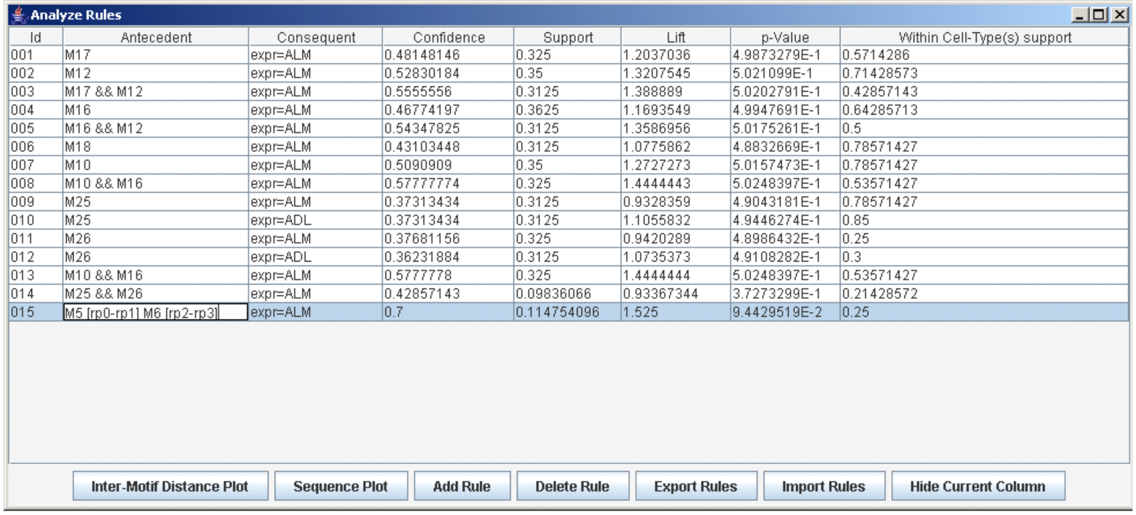
Figure C.11: Sequence Plot for Motifs M5 and M6. Observe that in the rule supporting sequences (upper part) a red dot is usually followed by a blue dot scanning the gene sequence from right end (SoT) to left.

Section C.2.1, enumeration item 2, the sequence plot displays all instances of participating motifs for qualifying sequences as they occur on the gene relative to the SoT. Since each motif appears in its own color and the data is being visualized in the context of a single rule, one can often see a repetitive pattern of color in the upper half and a lack of the same in the lower half of the plot. Such a display could also indicate an influence of order of occurrence on gene expression. In Figure C.11 one can observe that most sequences in the upper half have an occurrence of M5 (light point in the graph) closer to the SoT (the far right end of the plot) that is followed by a dark dot somewhere on the gene sequence. Also that this pattern is not so frequent in the lower part. Such pattern observation could indicate order of occurrence type relationship.

3. **ASAS mining algorithm** - The WPI implementation of association rule mining [PR05] used by us is capable of mining association rules with order based information and hence it is possible to have some of these rules being available already at the beginning of the exploratory analysis. Any of the two means mentioned above could be used to visually confirm/observe the order of occurrence relationship.

C.2.3 Add Rule

Irrespective of the method used to identify a potential order of occurrence relationship between participating motifs the following option from the Analysis Frame could be used to add order-based rules. Once a user has identified an order-based (or any other) relationship between motifs, it can use the "Add Rule" option in the Analysis Frame to add a blank row for the new rule. The user can then simply type in the Antecedents and the Consequent of the rule to calculate the different

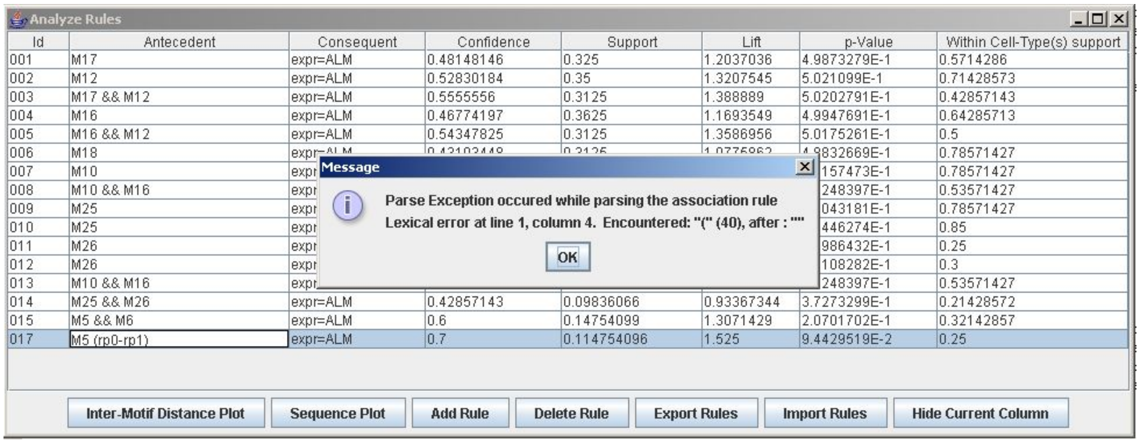


Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr=ALM	0.5090909	0.35	1.2727273	5.0157473E-1	0.78571427
008	M10 && M16	expr=ALM	0.57777774	0.325	1.4444443	5.0248397E-1	0.53571427
009	M25	expr=ALM	0.37313434	0.3125	0.9328359	4.9043181E-1	0.78571427
010	M25	expr=ADL	0.37313434	0.3125	1.1055832	4.9446274E-1	0.85
011	M26	expr=ALM	0.37681156	0.325	0.9420289	4.8986432E-1	0.25
012	M26	expr=ADL	0.36231884	0.3125	1.0735373	4.9108282E-1	0.3
013	M10 && M16	expr=ALM	0.5777778	0.325	1.4444444	5.0248397E-1	0.53571427
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572
015	M5 [rp0-rp1] M6 [rp2-rp3]	expr=ALM	0.7	0.114754096	1.525	9.4429519E-2	0.25

Figure C.12: Add Rule option in the Analysis Frame provides for free text option to add rules.

statistics indicating the interestingness metrics of the rule as shown in Figure C.12. Simply typing a complete rule computes the statistics indicating the interestingness of the rule. The user could also visualize the new rule using either the sequence plot or the inter-motif distance plot.

A rule keyed in by the user which does not have a valid syntax results in an error as shown in Figure C.13



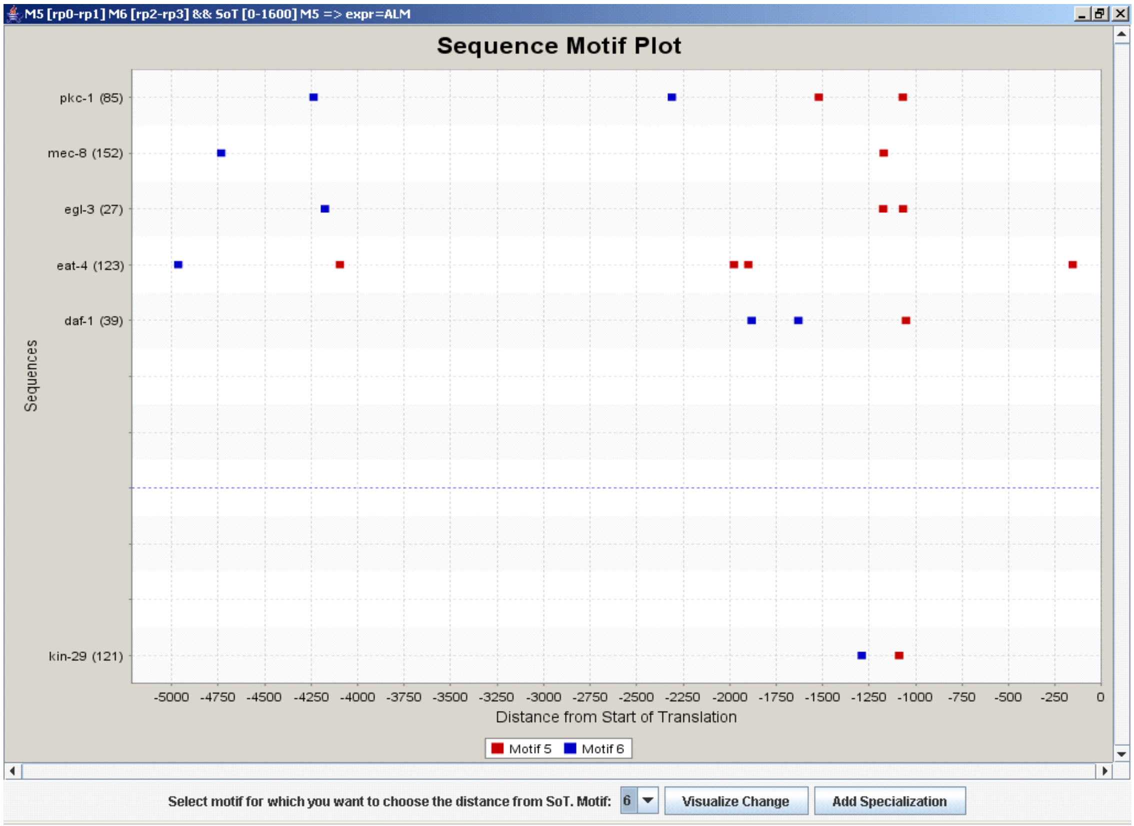
Id	Antecedent	Consequent	Confidence	Support	Lift	p-Value	Within Cell-Type(s) support
001	M17	expr=ALM	0.48148146	0.325	1.2037036	4.9873279E-1	0.5714286
002	M12	expr=ALM	0.52830184	0.35	1.3207545	5.021099E-1	0.71428573
003	M17 && M12	expr=ALM	0.5555556	0.3125	1.388889	5.0202791E-1	0.42857143
004	M16	expr=ALM	0.46774197	0.3625	1.1693549	4.9947691E-1	0.64285713
005	M16 && M12	expr=ALM	0.54347825	0.3125	1.3586956	5.0175261E-1	0.5
006	M18	expr=ALM	0.43103448	0.3125	1.0775862	4.8832669E-1	0.78571427
007	M10	expr				157473E-1	0.78571427
008	M10 && M16	expr				248397E-1	0.53571427
009	M25	expr				043181E-1	0.78571427
010	M25	expr				446274E-1	0.85
011	M26	expr				986432E-1	0.25
012	M26	expr				108282E-1	0.3
013	M10 && M16	expr				248397E-1	0.53571427
014	M25 && M26	expr=ALM	0.42857143	0.09836066	0.93367344	3.7273299E-1	0.21428572
015	M5 && M6	expr=ALM	0.6	0.14754099	1.3071429	2.0701702E-1	0.32142857
017	M5 (rp0-rp1)	expr=ALM	0.7	0.114754096	1.525	9.4429519E-2	0.25

Parse Exception occured while parsing the association rule
Lexical error at line 1, column 4. Encountered: "(" (40), after : ""

Figure C.13: Grammar based parsing helps identify user-errors in typing the rule.

C.2.4 Hybrid Rule

As described in the grammar governing rule definitions, each rule consists of an antecedent and a consequent. Antecedents in turn consists of terms. A rule could also include specialized term, extra hypothesis-based information(constraints) that the instances of the participating motifs must satisfy in order for a gene sequence to support the rule.


M5 [rp0-rp1] M6 [rp2-rp3] && SoT [0-1600] M5 => expr=ALM

Select motif for which you want to choose the distance from SoT. Motif: 6

Figure C.14: Hybrid rules help specify multiple constraints (based on different hypothesis) within a single specialization.

The system also supports hybrid rules, rules that consists of specialized terms based on different hypothesis and a gene sequence must satisfy all constraints in order to support the rule. E.g. With reference to Figure C.11 note that there exists an instance of M5 (the red dot) usually within the first 1600 bp from the SoT (Far right end of the plot). It is interesting to combine the two observations into a rule as follows and visualize it or calculate its interestingness. As we see in the figure C.14 that this hybrid specialization:

$$M5[rp0 - rp1]M6[rp2 - rp3]\&\&SoT[0 - 1600]M5 \Rightarrow expr = ALM \tag{C.3}$$

has a higher confidence as compared to the following simpler "order of occurrence" specialization

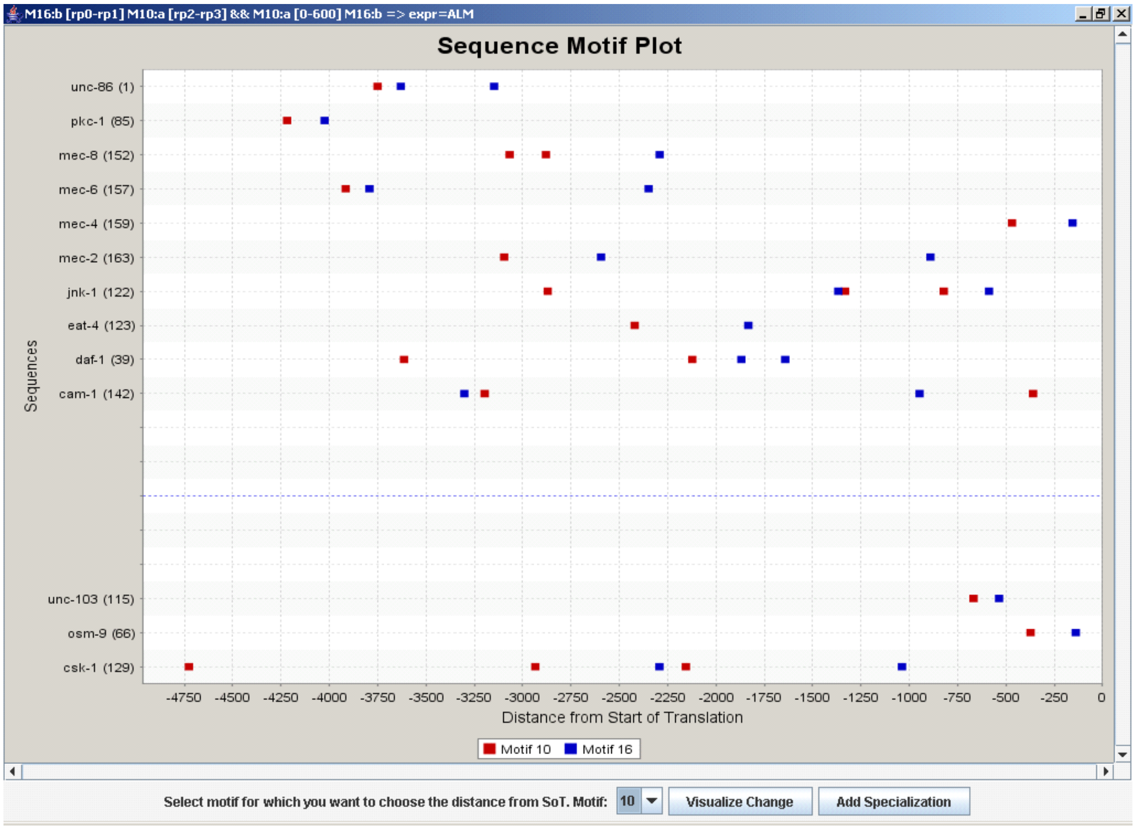
$$M5[rp0 - rp1]M6[rp2 - rp3] \Rightarrow expr = ALM \tag{C.4}$$

As seen above hybrid specialization could have multiple specialized terms that relate to a single motif. A hybrid specialization could post multiple constraints on the same motif like Distance from SoT and Order of occurrence relative to another motif. It is important to note that although the rule may have multiple constraints for the same motif, it is not required that the same instance of the motif satisfies each of them. In the context of the C.3 above, it is not required that the instance of Motif 5 that satisfies the order of occurrence condition is the same M5(instance) that lies within 1600 base pairs of the SoT. Although there might be a need for the user to actually specify constraints which are inter-related and aliases are supported by the rule grammar for exactly this reason.

C.2.5 Aliases

Aliases were included in the grammar to provide the user with an option of defining inter-related constraints or specialization terms. Consider the following Inter-Motif Distance based specialization from Figure C.5:

$$M10[0 - 500]M16 \Rightarrow expr = ALM \tag{C.5}$$


M16:b [rp0-rp1] M10:a [rp2-rp3] && M10:a [0-600] M16:b => expr=ALM

Select motif for which you want to choose the distance from SoT. Motif: 10

Figure C.15: Aliases let user define specializations with inter-related constraints.

Visualizing this specialization using a sequence plot (Figure C.7, one can see distinctly that not only do motifs M10 (red dot) and M16(blue dot) occur close together but they also occur in pattern such that the same instances of M10 and M16 that are involved in the distance-based relationship also occur in the same order relative to the SoT. Aliases enable the user to specify such complex relationships in the rule as follows (Figure C.15:

$$M10 : a[0-600]M16 : b\&\&M16 : b[rp0-rp1]M10 : a[rp2-rp3] \Rightarrow expr = ALM \tag{C.6}$$

C.2.6 Delete Rule

A user (usually a domain expert) can often identify rules which are of little biological significance and may want to delete such rules from the rule set. Simply selecting a rule from the Analysis frame and then clicking on the "Delete Rule" button could be used to accomplish exactly this.

C.2.7 Export Rules

This option enables a user to save a copy of the rule set currently in the Analysis Frame to a text file. This provides the user the facility to resume working on the rule set at a later point in time or maintain motif based rule sets. The extensions to WPI-Weka rule-miner [Rudss] ensured that the rule model can also be imported into the rule mining interface.

C.2.8 Import Rules

This option enables to import a rule model to be imported from a text-file. It could be either from a previous session of the analysis-frame or could be a rule set exported from the rule-mining interface. (Thanks Jon)

C.2.9 Hide Current Column

If a user thinks, that one of the columns in the analysis frame is not of use in the current context the user has the option of removing a column from the display. Simply select a cell in the column that the user wants to delete and click the "Hide Current Column" button.

C.2.10 Sorting in Analysis Frame

The Analysis Frame also provides the option of sorting the rule model in the analysis frame by simply clicking on the header of the column by which the rule model needs to be sorted. The order of sorting could also be reversed by simply clicking on the header column once more.