DE NOVO TRANSCRIPTION FACTOR BINDING SITE DISCOVERY: A MACHINE LEARNING AND MODEL SELECTION APPROACH

James E Scherschel

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree of
Master of Science in Bioinformatics,
Indiana University

May 2009

Accepted by the Faculty of Indiana University,
in partial fulfillment of the requirements for the degree of Master of Science
in Bioinformatics

**Master's Thesis
Committee**

Faculty Name, Degree Title, Chair

Faculty Name, Degree Title

Faculty Name, Degree Title

© 2009
James E Scherschel
ALL RIGHTS RESERVED

Acknowledgements

It is my sincere pleasure to acknowledge some of the many people without whom this thesis would not be what it is. Foremost, I must thank my advisor, Dr Narayanan Perumal, for his guidance, insights, encouragement, and patience; his input and efforts have truly been invaluable. I must also thank the other members of my thesis committee, Dr Yunlong Liu and Dr Pedro Romero, for their time and critical feedback. I would like to sincerely thank Michael Martin and the members of Dr Perumal's lab for patiently endured progress reports on this work and frequently offering useful suggestions and comments. For providing the financial support and flexible work schedule that made my studies possible, I would like to thank my employer.

Without adequate academic preparation, my studies could not have been a positive and successful experience. For properly preparing me mentally for the process of writing this thesis, I thank my undergraduate thesis advisor, Dr Michael Karls. My thanks also to the mathematics, chemistry, computer science, and physics departments at Ball State University for having provided a firm foundation on which to build.

The years of my graduate studies have been tumultuous personally, marked by the loss of several people close to me. I would like to thank those that remain for their love, support, and understanding. Lastly, my thanks to "The Like-Minded Patriots" for helping me to keep it all in perspective.

For all the others that deserve my thanks but have never expected it: Thank you, All.

Abstract

Computational methods have been widely applied to the problem of predicting regulatory elements. Many tools have been proposed. Each has taken a different approach and has been based on different underlying sets of assumptions, frequently similar to those of other tools. To date, the accuracy of each individual tool has been relatively poor. Noting that different tools often report different results, common practice is to analyze a given set of regulatory regions using more than one tool and to manually compare the results. Recently, ensemble approaches have been proposed that automate the execution of a set of tools and aggregate the results. This has been seen to provide some improvement but is still handled in an ad hoc manner since tool outputs are often in dissimilar formats. Another approach to improve accuracy has been to investigate the objective functions currently in use and identify additional informational statistics to incorporate into them. As a result of this investigation, one statistical measure of positional specificity has been demonstrated to be informative.

In this context, this thesis explores the application of three simple models for the positional distribution of transcription factor binding sites (TFBS) to the problem of TFBS discovery. As alternate measures of positional specificity, log-likelihood ratios for the three models are calculated and treated as features to classify TFBSs as biologically relevant or irrelevant. As a verification step, randomly generated positional distributions are analyzed to demonstrate the robustness and accuracy of the log-likelihood ratios at classifying data from known distributions using a simple classifier. To improve classification accuracy, a support vector machine (SVM) approach is used. Subsequently, randomly generated sequences seeded with TFBSs at positions chosen to conform to one of the three models are analyzed as an additional verification step. Finally, two types of sets of real regulatory region sequences are analyzed. First, results consistent with the literature are obtained in three cases for genes experimentally determined to be co-expressed during mouse thymocyte maturation, and a novel role is predicted for three families of TFBSs in single positive (SP) T-cells. Second, the mouse and human "real" sets from Tompa *et al*'s "Assessment of Computational Motif Discovery Tools" are analyzed, and the results are reported.

Table of Contents

Acknowledgements iv

Abstract v

List of Tables vii

List of Figures viii

Introduction 1

Background 5

Methods and Materials 18

- Assumption of the Uniform Model 18
- Assumption of the Normal Model 18
- Assumption of Model 3 19
- Assumed Prior 19
- Maximum Likelihood Parameter Estimation 19
- Confirmation of Likelihood Maximization 22
  - For the Uniform Model 22
  - For the Normal Model 23
  - For Model 3 24
- Classifier Construction and Verification 29
  - Distinguishing Data Sets from the Uniform Model, Normal Model, and Model 3 .. 29
  - Distinguishing Seeds Motifs from Background in Seeded Sequences 35
  - Discovering Motifs in Promoter Regions of Co-Expressed Genes 36

Results 40

- Results For Lymphocyte Data 40
- Results For Seeded Sequences 43
- Results For Benchmark Sequences 49

Conclusions 53

Discussion and Recommendations 55

Works Cited 61

Curriculum Vitae 65

List of Tables

Table #1: Common TFBS Prediction Tools
Table #2: Common Models for Representing Motifs
Table #3: More General Models for Representing Motifs
Table #4: Algorithm Details for Common Combinatorial Tools
Table #5: Algorithm Details for Common Greedy Tools
Table #6: Algorithm Details for Common Probabilistic Tools
Table #7: Model Parameter Bounds for Model and Algorithm Verification
Table #8: Verification of θ^*
Table #9: Puthier *et al* Clusters and Gene Counts
Table #10: "Interesting" Motif Counts and JASPAR Motif Counts and Names by Cluster
Table #11: "Interesting" Motifs Highly Similar to Known Motifs
Table #12: Normalization Constants Used for Input to SVM #2
Table #13: SVM #2 Information Retrieval Metrics For 7-mer Seeds
Table #14: Additional Statistics
Table #15: Best Results from the Tompa Benchmark
Table #16: Information Retrieval Metrics for Best Results from the Tompa Benchmark
Table #17: Predicted Motifs and Known TFBSs for Best Results from the Tompa Benchmark
Table #18: Number of Predicted Motifs and Number of Sequences Per Set

List of Figures

Figure #1: Uniform Model Log-Likelihood Near the Predicted θ^*

Figure #2: Normal Model Log-Likelihood Near the Predicted θ^*

Figure #3: Model 3 Log-Likelihood vs u

Figure #4: Model 3 Log-Likelihood vs μ

Figure #5: Model 3 Log-Likelihood vs σ

Figure #6: Model 3 Log-Likelihood Over the Full Parameter Space

Figure #7: Model 3 Log-Likelihood Nearer the Predicted θ^*

Figure #8: Model 3 Maximum Log-Likelihood Found at the Predicted θ^*

Figure #9: Model 3/Normal LLR vs Model 3/Uniform LLR

Figure #10: "A Simple Classifier"

Figure #11: Classification of Verification Data Sets Using “A Simple Classifier”

Figure #12: ROC Curve for LLR-Based Classification of Normal and Uniform Sets.

Figure #13: Venn Diagram of JASPAR Motifs for “Interesting” Motifs

Figure #14: Optimization of Kernel Parameters g and c for SVM #2 using “Seeded” Random DNA Sequences

Figure #15: SVM #2 Classification Accuracy for Candidate Motifs (7-mer Seeds)

Figure #16: SVM #2 Classification of Candidate Motifs from “Seeded” Random DNA Sequences

Figure #17: Visualization of Predicted and Known TFBSs for hm25r

Figure #18: Visualization of Predicted and Known TFBSs for mus08r

Introduction

The biological sciences, like all branches of science, seek to observe, understand, predict, and, ultimately, manipulate events in the physical world. In the past 400 years, significant progress has been made in advancing these goals. Through the visual observation of microscopic “cells” by Robert Hooke in the mid-1600s and the development of the “cell theory” of Matthias Schleiden, Theodore Schwann, and Rudolf Virchow in the mid-1800s, the cell is accepted as the fundamental structural and functional unit of life. Thanks to the timely rediscovery of Gregor Johann Mendel’s pea plant experiments and the efforts of a host of scientists in the 1900s, the gene has been identified as the fundamental unit of heredity through which traits are passed to offspring, and Watson, Crick, *et al*’s famous DNA double-helix has been identified as the biochemical embodiment of genes. From the “Central Dogma of Molecular Biology” first enunciated by Crick and also an ever-growing body of experimental evidence, the flow of information between DNA, RNA, and proteins has been elucidated, thus providing an understanding of how DNA is transcribed to RNA which is in turn translated into proteins and providing an explanation for the phenotype of any given organism. This understanding has laid the foundation for predicting the phenotypes of organisms, including a variety of prenatal and post-natal genetic screens for humans, and, in the form of genetically modified plants and modern small molecule and protein pharmaceuticals, is beginning to allow human manipulation, beyond selective breeding, to achieve desired outcomes such as increased yield and disease resistance for plants and improved health for humans.

More recently, numerous efforts have been undertaken to sequence all of the DNA, the complete genomes, for a variety of organisms including mice, rats, fruit flies, yeast, chimpanzees, and humans. These massive research efforts have provided vast amounts of information necessary for further progress, driven the development of new technologies to determine genetic sequences, and pushed computer technology and computational methods vital to storage and analysis of this data. In so doing, these efforts have also produced results that have surprised researchers. For example, initial estimates of the number of genes in the human genome were repeatedly revised downward from 100,000 or more to the current estimate of 20,000 to 25,000. With less than twice as many genes as many far simpler organisms such as fruit flies, roughly the same number of genes as mice and rats, nearly exactly as many as chimpanzees, and roughly 95% of the human and chimpanzee genomes having been determined to be identical, biological complexity is clearly not directly related to the number of genes in an organism's genome. To better understand this apparent discrepancy, a great deal of current research involves understanding the significance of those relatively small differences, DNA regions of unknown function, the RNA splice variations that allow the modest human genome to encode a much larger proteome, and the complex regulatory networks that modulate the expression of genes.

As with any complex system, this system can operate in a number of modes, not all of them associated with good health. Many diseases and conditions, including heart disease, cancer, stroke, diabetes, and Alzheimer's disease, have a known genetic component. Viewed through this lens, the nearly 60% of deaths each year in the US attributed to these diseases and conditions are problems of misregulation, or the levels of elements of this regulatory network being outside of "healthy" ranges. The cost of heart disease and stroke alone, including health care expenditures and lost productivity from deaths and disability, is projected to be more than $475 billion in 2009. As the U.S. population ages, the economic impact of cardiovascular diseases on our nation's health care system will become even greater. The cost of cancer care, which excludes the huge economic impacts of deaths and lost productivity, was estimated to have been $72.1 billion in 2004 alone. Estimates for 2009 are currently being calculated by the National Cancer Institute and seem likely to be significantly higher. The other diseases and conditions listed have huge economic and emotional impacts also. Whether measured monetarily, in terms of lost lives, or in terms of decreased quality of life, there are huge potential benefits to be reaped from a better understanding of the basic science behind gene regulation and the application of that knowledge to create interventions that alter gene expression patterns to push the complex regulatory network to a more healthy state.

The sets of genes expressed and their levels of expression typically vary from cell type to cell type, tissue to tissue, and with the age of the organism, from fetus to adult. Common across these cell lineage, spatial, and temporal distances, the expression of genes is largely controlled at the level of transcription, though the set of genes being transcribed and the levels of transcription may be radically different. Due to its key role, this is both a biological process that needs to be well understood and a potential point of intervention to treat or prevent disease development. Challengingly, transcription is a complex process dependent on a myriad of factors. At a gross level, how tightly DNA is coiled around the histone and non-histone proteins that, in addition to DNA, compose chromatin dictates how available the DNA will be to the cellular machinery responsible for transcription. Within regions of DNA that are less compacted, transcription factors (TFs) are known to drive complex patterns of gene expression through their influence on the recruitment of the basal transcription complex to and its activation at particular sites within the genome. TFs are small regulatory proteins that typically have a modest affinity for DNA in general and a high affinity for particular short sequences or families of sequences of DNA. Such a short sequence is often called a DNA motif and has been seen to vary from as few as three or four to more than a dozen nucleotides in length.

Many TFs have been experimentally identified, and, increasingly, computational tools have been used to predict additional transcription factor binding sites (TFBSs), to identify other genes likely to be under the regulatory control of the same known TFs, and, based on patterns of gene co-expression, to identify genes that are likely to have common, yet-to-be-identified TFs. Despite the efforts to date, a recent survey paper estimated that, for more than half of the TFs in the human genome, neither the binding partners nor the binding sites are known. Though steady progress is being made, this is a daunting "needle in a haystack" problem for bench biologists. Similarly, the task of developing effective computational methods, particularly for the identification of yet-to-be-identified TFs, is extremely difficult due to the large size of the human genome, the small size of the DNA "alphabet", the frequent degeneracy of motifs for individual TFBSs, and the complexity inherent in the environment in which transcription is occurring. Though many computational tools have been created in attempts to address this problem and a significant contribution has been made, a 2005 study reported that current performance was poor with only roughly 35% of known binding sites being correctly predicted by even the best publicly available tools. As a result, research has been ongoing to improve this set of tools.

The ideal tool would include, at its heart, a complete mathematical model accurately and completely capturing the biological complexity of transcriptional regulation. Such a model would necessarily include pattern specificity, positional preference, interaction among TFs and the basal transcriptional apparatus, and a variety of other dimensions. When faced with a complex system, projections and first-order approximations are often meaningfully applied to obtain information or make predictions about the system. In this vein and because a complete model of transcriptional regulation is well beyond the scope of this thesis, simple models of position specificity will be considered. By applying model selection via machine learning techniques, these simple models will inform the winnowing of candidate motifs to putative TFBS motifs.

Background

Demonstrating the challenges and perceived potential benefits of a having a good solution, there are more than 120 TFBS prediction tools in the literature according to one recent count. Dozens of tools for *de novo* prediction of TFBS motifs have been proposed and are currently publically available and in wide use. Each tool approaches the problem of TFBS discovery slightly differently in terms of the types of input data required, how binding sites are internally represented, the data structures and algorithms that are used, and how putative TFBSs are scored and ranked. Because of differences in how the problem has been conceptualized, each tool is predisposed to score different candidate sites differently than the other tools and to potentially yield different predictions. Despite their differences, the current tools can be grouped according to a relatively small set of dimensions into a short list of families. Table #1 contains a list of common tools and some relevant details. Cells in Table #1 are blank if the reference did not clearly provide the information. Also, since promoter regions are required by all tools, this is implied.

Table #1: Common TFBS Prediction Tools

	Type Of					
Tool Name	**Algorithm**	**Motif Model**	**Match Model**	**Required Information**	**Objective Function**	**Reference**
A-GLAM	Probabilistic (Gibbs)	string		Positional anchors	position and sequence specificity (e-value)	(Kim, Tharakaraman and Mariño-Ramírez)
AlignACE	Probabilistic (Gibbs)	matrix	PWM	Full genome	motif over-representation (MAP score)	(Hughes, Estep and Tavazoie)
ANN-Spec	Probabilistic (Gibbs)	matrix	PWM	Positive training data Background	sequence specificity (Information content, or IC)	(Workman and Stormo)
BioProspector	Probabilistic (Gibbs)	matrix, dyad	PWM	Background	motif over-representation (z-score)	(Liu, Brutlag and Liu)
Consensus	Greedy (Tree building)	matrix	PWM		sequence specificity (IC)	(Hertz)
cWINNOWER	Combinatorial (Graph-based)	matrix	PWM		pattern specificity	(Liang)

Table #1: Common TFBS Prediction Tools (continued)

Tool Name	Type Of Algorithm	Motif Model	Match Model	Required Information*	Objective Function	Reference
EMD	Ensemble	Multiple	Multiple	Determined by the ensemble of tools used	Multiple	(Hu, Yang and Kihara)
EMnEm	Probabilistic (EM)			Phylogenetic		(Moses)
FMGA	Probabilistic	matrix	PWM			(Liu, Tsai and Chen)
Gibbs Sampler	Probabilistic (EM)	matrix	PWM			(Newberg, Thompson and Conlan)
GibbsST	Probabilistic (Gibbs)					(Shida)
GLAM	Probabilistic (Gibbs)	string				(Frith)
Improbizer	Probabilistic (EM)		PWM	Background (Positional anchors)		(Ao, Gaudet and Kent)
MDScan	Greedy	string	PWM			(Liu, Brutlag and Liu)
MEME	Probabilistic (EM)	matrix	PWM		p-value (Log likelihood ratios)	(Bailey and Elkan)
Mitra	Combinatorial	string, dyad	mismatch			(Eskin)
MotifSampler	Probabilistic (Gibbs)		PWM	Background		(Thijs)
NestedMICA	Probabilistic (Gibbs)		PWM			(Down and Hubbard)
Oligo/Dyad-Analysis	Combinatorial	string, dyad	oligos	Background		(van Helden, Andre and and Collado-Vides)
OrthoMEME	Probabilistic (EM)		PWM	Phylogenetic	p-value (Log likelihood ratios)	(Prakash)
PhyloCon	Greedy (Tree building)		PWM	Phylogenetic		(Wang)
PhyloGibbs	Probabilistic (Gibbs)	matrix	PWM	Phylogenetic		(Siddharthan, Siggia and van Nimwegen)
PhyloScan				Phylogenetic		(Carmack, McCue and Newberg)

Table #1: Common TFBS Prediction Tools (continued)

Tool Name	Type Of Algorithm	Motif Model	Match Model	Required Information*	Objective Function	Reference
PhyME	Probabilistic (EM)			Phylogenetic	sequence specificity (Information content)	(Sinha, Blanchette and Tompa, PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences)
QuickScore		string	IUPAC	Background	motif over-representation (z-score/p-value)	(Regnier and Denise)
SeSiMCMC	Probabalistic (Gibbs)		PWM		sequence specificity (Information content)	(Favorov)
SP-STAR	Combinatorial (Graph-based)	string	mismatch		pattern specificity (sum of pair-wise dissimilarity scores)	(Pevzner and Sze)
Weeder	Combinatorial	string	mismatch		pattern specificity (similarity score and information content)	(Pavesi, Mauri and Pesole)
WINNOWER	Combinatorial (Graph-based)	string	mismatch		pattern specificity	(Pevzner and Sze)
YMF	Combinatorial	string	reg.exp		motif over-representation (z-score)	(Sinha and Tompa, YMF: A program for discovery of novel transcritpion factor binding sites by statistical overrepresentation)

Though genome-wide searches have been conducted in an attempt to identify species-level regulatory networks and biological pathways, TFBS discovery tools are more often applied to sets of co-expressed genes. Commonly, sequence information from the upstream promoter/enhancer regions for genes experimentally determined to be co-expressed, sequence information from upstream promoter/enhancer regions for homologous genes indentified, phylogenic information, or some combination of the aforementioned is used by *de novo* TFBS discovery algorithms. For protozoa, sequence information from the promoter/enhancer regions for genes experimentally determined to be co-expressed has been used by several algorithms that have been observed to perform well in tests to identify TBFSs for genes regulated by known TFs with known TFBSs. For the more complex regulatory networks of metazoans, performance has been more modest. This has been partially attributed to the interplay of a variety of different TFs influencing the expression of genes. Also, binding sites for regulatory proteins have been identified in introns and downstream regions of metazoan genes expanding the size and nature of the regions that must be considered. Phylogenetic footprinting and approaches that search for binding sites for sets of TFs at a time have both been investigated with some success as means to improve performance.

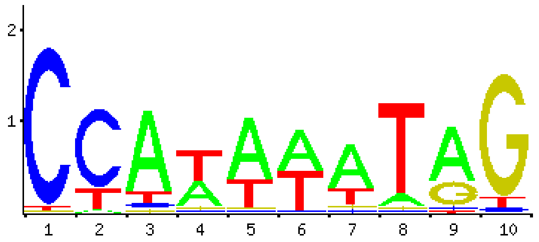
Several models for capturing motifs have been proposed and are described below in Table #2 and Table #3. They are listed in increasing order of generality. Contrary to intuition, successful TFBS predictions have been reported using even the simplest models, primarily in yeast. Position-weight matrices are currently the most common internal representation for motifs.

Table #2: Common Models for Representing Motifs

Motif Representation		
Name	**Description**	**Example***
String	Represents each motif as a consensus sequence of nucleotides	`CCATAAATAG`
IUPAC String	Allows specification of a list of preferred nucleotides for each position within the motif using the IUPAC nucleotide naming convention, but does not allow weighting of the individual nucleotides at each position.	`CYWWWWWWRG`
Regular Expression (RegEx)	In practice, typically used as an alternate representation equivalent to an IUPAC string. RegEx can additionally represent variable-length regions and a wide range of more complex motifs but does not capture positional weighting information.	`C(C|T)(A|T){6}(A|G)G`

*The example motif depicted is the Arabidopsis thaliana motif AGL3 from the Jaspar CORE database.

Table #2: Common Models for Representing Motifs (continued)

Motif Representation		
Name	**Description**	**Example***
Position Frequency Matrix (PFM)	A matrix of the frequency of each nucleotide at each position within a set of occurrences of a given motif. This representation is often used for experimentally determined or verified motifs.	A [0 3 79 40 66 48 65 11 65 0] C [94 75 4 3 1 2 5 2 3 3] G [1 0 3 4 1 0 5 3 28 88] T [2 19 11 50 29 47 22 81 1 6]
Position Weight Matrix (PWM) Position Specific Weight Matrix (PSWM) Position Specific Scoring Matrix (PSSM)	Similar to PFM but no experimental verification is implied. Represents the relative frequency or probability of each nucleotide at each position. This is equivalent to a zero-order Markov Model.	A [.00 .03 .81 .41 .68 .49 .67 .11 .67 .00] C [.97 .77 .04 .03 .01 .02 .05 .02 .03 .03] G [.01 .00 .03 .04 .01 .00 .05 .03 .29 .91] T [.02 .20 .11 .52 .30 .48 .23 .84 .01 .06]
LOGO	A common graphical representation of the information content at each position within a motif. LOGOs are typically generated using PWM and the GC content of the genome of interest. The GC content of the genome of interest is important because, for example, the motif ATAT would have significantly higher information content in a genome with high GC content than the motif GCGC.	

*The example motif depicted is the Arabidopsis thaliana motif AGL3 from the Jaspar CORE database.

The representations in Table #3 are less common and are more general models than those listed in Table #2. The representations in Table #3, due to their complexity and the general support for the assumption of the independence among the sites in a given motif, are not in general use. See Ben-Gal *et al* (Ben-Gal) for examples of these representations.

Table #3: More General Models for Representing Motifs

Motif Representation	
Name	**Description**
Variable-Order Markov Model (VOM)	Though not in common use, a VOM allows the probability or relative frequency of each nucleotide at each position within a motif to be dependent on the nucleotide(s) zero or more immediately preceding positions. Positional independence within motifs is assumed by simpler representations. This assumption is generally supported by current experimental evidence but has been challenged recently. VOMs would allow more accurate representation of motifs that contained adjacent nucleotides that do not co-occur independently.
Bayesian Network (BN)	Extends VOM to allow each position within a motif to be dependent on a fixed number of adjacent and/or non-adjacent positions within the motif.
Variable-Order Bayesian Network (VOBN)	Extends BN to allow each position to be dependent on a variable number of other positions.

TFBS discovery algorithms are exhaustive, heuristic, or probabilistic. Exhaustive solutions typically utilize combinatorial approaches and produce globally optimal solutions for a given algorithm's objective function. Exhaustive algorithms are generally "word-based" and treat the problem of motif identification as a search problem over the alphabet of the algorithm's internal motif representation for over-represented matches within the set of sequences being analyzed. It is worth noting that if any algorithm's objective function correctly scores candidate TFBSs, then this corresponds to finding precisely the true TFBSs. A perfect objective function would be able to perfectly separate in vivo TFBSs from biologically irrelevant sites. Based on the performance of the current set of tools, the common objective functions are far from perfect, but suggestions have been made to improve them. Regardless, many approaches have been reported to demonstrate at least modest success. Table #4 provides some additional details about several common tools that take a combinatorial approach.

Table #4: Algorithm Details for Common Combinatorial Tools

Tool Name	Algorithm Details
cWINNOWER	Improves on WINNOWER and SP-STAR by adding a stronger constraint function.
Mitra	Assumes a hypergeometric distribution of TFBSs and scores results for putative motifs relative to background sequences using a suffix tree based approach.
Oligo/Dyad-Analysis	Relies on a user-selected model of expected frequency of motifs and predicts TFBSs by comparing actual frequencies against expected results. Typically, the expected model is a hidden Markov model of order 0-3. The algorithm was extended to additionally detect dyads spaced by 0-16bp. Only effective at finding short motifs with well-conserved cores because variations within the oligonucleotide is not allowed.
QuickScore	Calculates z-scores and p-values for rare or infrequent patterns relative to the background which is modelled as a Markov model of order up to 3.
SP-STAR	Utilizes a local sum of pairwise score improvement algorithm
Weeder	Makes use of suffix trees to perform an efficient, nearly exhaustive search for candidate motifs that the creators describe as "'almost' exact". Matches are ranked using a "significance" metric and a measure of relative entropy. High-scoring matches are typically clustered to improve results. Requires a trade-off be made between execution time and probability of missing a significant motif. Has scored highly in benchmarks.
WINNOWER	Uses a word-based graph-theoretic method to find motifs by representing candidate binding sites as vertices and pruning edges from the graph to retain a minimal spanning set
YMF	Assumes a binomial distribution of k-mers with a fixed small number of allowed substitutions to estimate the probability of the random occurrence of the detected number of matches to each candidate motif of length k. The probability of each motif can be estimated either assuming independent nucleotides (matching a given or measured GC content) or based on a Markov chain. Shown to perform well on sets of promoter regions for co-regulated yeast genes. (Outperformed MEME and AlignACE for such data sets.) Performance against eukaryotic and mammalian data sets has been weaker.

Greedy algorithms are generally considered a type of heuristic algorithms and have been used by several TFBS prediction tools. Like other greedy algorithms, the idea is to pursue one or more “best” partial solutions to one or more “best” full solution. Though global optimality of the solution is not guaranteed, the resources required, either in storage or computation time, are typically greatly reduced. This reduction in the required resources is often considered acceptable because of the wider range of data sets that can be analyzed. Table #5 provides some additional details about several common tools based on greedy algorithms.

Table #5: Algorithm Details for Common Greedy Tools

Tool Name	Algorithm Details
Consensus	Based on a greedy strategy that progressively extends a bounded number of partial alignments.
MDScan	Uses a word-enumeration strategy to find abundant k-mers using an approximate maximum a *posteriori* scoring function. The enumerated words are used as "seeds" against which to score similarity to other sites, build a list of similar sites, and generate a consensus motif. Performance reported to be comparable to BioProspector but significantly faster
PhyloCon	Groups sequences based on orthology, aligns the sequences, and extracts motifs from the groups of aligned sequences.
PhyloScan	Allows the combination of alignable and non-alignable sequence data from multiple intergenic regions to be used without training data to predict significant motifs.

Like greedy algorithms, probabilistic algorithms make a trade-off between the guaranteed optimality of solutions and the execution time and storage resources required for the algorithm to be applied to data sets. The common feature of all probabilistic algorithms is that some part of the algorithm depends on chance, typically in the form of a (pseudo-)random number generator. They tend to have objective functions that are simpler to describe than combinatorial approaches and simpler data structures since data structures generally do not need to be optimized for efficiency. Gibbs sampling (Gibbs) and Expectation Maximization (EM) form the basis of most probabilistic algorithms for TFBS discovery. Table #6 provides additional details about several common probabilistic tools for TFBS discovery.

Table #6: Algorithm Details for Common Probabilistic Tools

Tool Name	Algorithm Details
A-GLAM	Scores candidate motifs using a Bayesian model based on sequence specificity and occurrence location relative to TSS, or some other positional landmark, for genes of interest. Requires a consistent anchor, such as the correct TSS, be known for the genes of interest. Prefers longer sequences for best performance.

Table #6: Algorithm Details for Common Probabilistic Tools (continued)

Tool Name	Algorithm Details
AlignACE	Gauges over-representation using Gibbs sampling to obtain a maximum *a priori* log-likelihood (MAP) score. A cutoff based on how frequently a motif occurs in the full genome or the specificity of the motif for the genes of interest can be specified to filter the results obtained. The latter is considered more useful than the former since some motifs are observed to occur frequently in some genomes, such as yeast. Clustering of the obtained results, using CompareACE, is typically performed to improve predictions. Outperformed by several tools
ANN-Spec	Uses an artificial neural network and Gibbs sampling to find parameters of a weight martix representing DNA-binding specificity to indirectly learn an ungapped local multiple sequence alignment. Requires positive training data and background data for optimal performance. (Outperformed Gibbs sampler, Consensus, and MEME in terms of specificity and Consensus and MEME overall.)
BioProspector	Uses a variant on Gibbs sampling, a 0- to 3-order Markov model generated from a provided sequence file (typically the intergenic or promoter regions of the full genome of the species of interest), and an algorithm accepting 15 user-provided parameters to find significant motifs. Among the options, gapped or dyad motifs can be detected. Performance reported to be comparable to MDScan but significantly slower
EMnEm	Considers special motifs that are generated from ancestral sequences represented as a two-component mixture of background and motifs.
FMGA	Relies on a genetic algorithm with a rearrangement method to avoid extremely stable local minima. Outperformed MEME and Gibbs Sampler
Gibbs Sampler	Is a stochastic variant of the EM method that uses sampling weighted based on site scores from previous iterations.
GibbsST	Combines Gibbs sampling and simulated annealling in an effort to avoid the tendancy of the former to converge to local maxima.
GLAM	Uses Gibbs sampling and seeks to optimize the alignment and alignment width of candidate sites. Fragments long sequences into shorter ones to find more than one binding site per sequence.
Improbizer	Uses EM to find motifs that occur improbably often relative to an order <=2 Markov model of the background. A Gaussian model of positions of sites can optionally be constructed to add a test of positional specificity.
MEME	Is an expectation maximization (EM) method that uses a product of p-values associated with the information content of the positions within candidate motifs. This statistic implies the assumption that the positions within the motif are independent. Outperformed by several tools
MotifSampler	Extends Gibbs sampling using a higher order Markov model for the background and incorporating a Bayesian mechanism to estimate the number of motifs occurring in each sequence.

Table #6: Algorithm Details for Common Probabilistic Tools (continued)

Tool Name	Algorithm Details
NestedMICA	Treats the problem of motif finding as an independent component analyis, similar to principle component analysis, problem within a Bayesian probabilistic framework. Short motifs are modelled as position weight matrix "voices" in a the sea of "noise" of the remainder of the promoters being analyzed. Nested sampling, a Monte Carlo method more orderly than Metropolis-Hastings and Gibbs Sampling, is used to drive more efficiently convergence.
OrthoMEME	Generalizes MEME's framework and algorithm to allow regulatory regions from two species can be included in analysis for significant motifs.
PhyloGibbs	Combines phylogenetic footprinting and a search for overrepresented sequence motifs in an integrated framework and performs an anneal-and-track strategy to make estimates of the reliability of its predictions.
PhyME	Integrates two different axes of information content, one scoring intra-species motif frequency and one scoring inter-species conservation for the candidate motifs.
SeSiMCMC	Alternates between two-stages. One stage optimizes a candidate motif; optionally assuming the symmetry of a palindrome, a direct repeat, and a spacer; using likelihood relative to a Bernoulli background. Candidate motifs are organized via a Gibbs-like Markov chain. The other stage uses information content of matches and the position of occurrences to find the best matches.

Many objective functions rely on an e-value, p-value, and/or z-score representing the over- or under-representation of a particular motif relative to other motifs or some other background. To obtain the statistic(s), frequencies of occurrence of the candidate motifs are often assumed to be accurately modeled by binomial, hypergeometric, or negative binomial distributions to allow the estimates to be calculated. Most tools additionally consider a measure of sequence specificity such as information content when generating predictions. The choice of the objective function for a tool, since it is the scoring function by which candidate motifs and sites will be judged when making predictions, is obviously critically important. Recent analysis of computational approaches for motif discovery has demonstrated that statistics representing over- and under-representation are not sufficiently informative to allow accurate separation of known TFBS motifs from background. Significant improvements in prediction accuracy were demonstrated using an objective function incorporating information about the position of the binding sites. By the extension of the Bayesian model central to GLAM to include positional specificity, the tool A-GLAM provides at least one example of a tool that has been extended to incorporate this additional type of information.

To demonstrate the potential benefit of a new tool, favorable pair-wise comparisons of tools against generated and/or real data sets have frequently been reported. Though anecdotally supportive, pair-wise comparisons fail to provide a "big picture" of relative performance and were not typically conducted on common data sets for any large set of tools. More recently, comparison of tool performance relative to a set of common benchmark data sets has been reported for a variety of motif finding tools to remedy this deficiency (Tompa, Assessing computational tools for the discovery of transcription factor binding sites). Such tool benchmarking is now relatively common despite the acknowledged challenges in creating benchmarks that accurately gauge tool performance. The central challenge is that the underlying biology of regulatory networks is not well understood. The diversity of algorithms and approaches employed by the different tools is also a challenge because it makes it difficult to know whether a tool's performance is attributable to the relative optimality of the algorithm and internal model or to how a given test set was generated.

While the diversity of approaches presents a challenge, it also provides an opportunity. To exploit the strengths of a variety of different tools, ensemble approaches have been a recent area of focus to improve prediction performance. The use of complementary tools and ensemble approaches potentially provides significantly improved performance over any one tool alone. The EMD algorithm represents one ensemble approach (Hu, Yang and Kihara). In a 2006 test, a set of five TFBS discovery tools showed only 25-35% accuracy at the binding site level for sequences 400bp long and 15-25% accuracy at the nucleotide level for any individual tool, results comparable to those reported in 2005 benchmarks. Significantly, at least one tool in the set was capable of predicting the correct binding site 90% of the time. Overall, the best reported EMD algorithm performed 22.4% better than the best single component tool in terms of nucleotide-level accuracy. The reported ensemble objective function was a weighted aggregate of the scores from the component tools. The authors noted that the performance of the ensemble algorithm might be improved by optimizing the parameters of the component tools, by optimizing the weighting of the predictions from the component tools in the ensemble score, and/or by implementing a position-based voting scheme for candidate binding sites. Further exploring the ensemble approach, Wijaya *et* *al* reported the development of a tool that was able, using a voting scheme, to locate more than 95% of the binding sites found by its component tools and showed significant improvements in sensitivity and specificity.

Though significant steps forward, the refinement of ensemble methods does little to address the underlying problem, namely how to tell which tool (if any) has predicted one or more real TFBS motifs and which of the predicted TFBS motif(s) correspond to biologically relevant TFBSs. With a more complete understanding of how the component algorithms, their internal motif models, and their predicted binding sites are different, the performance of an ensemble approach might be significantly improved. Since each tool implicitly scores putative TFBS motifs against a different mathematical model, the problem of picking which predictions to believe from a set of predictions from different tools can be viewed as a model selection problem. Though the model for each tool might be complex and difficult to represent in a closed, non-algorithmic form, the underlying objective functions are likely simpler and easier to represent. If so, this implies that a common representation could be used to obtain a common set of metrics for a set of tools and their predictions, potentially allowing direct comparison of predictions in a common framework. If not, the outputs from each component tool could serve as features for a pattern classification algorithm to characterize the set of resulting predictions and when each tool's predictions are more likely to be correct.

It is common to use toy problems as both learning aids and to demonstrate the applicability of new approaches. In the latter vein, let us consider a relatively simple set of possible mathematical models, based on the positional occurrence of a perfectly conserved seven base-pair motifs, for TFBSs. It is possible that such simple models might correctly predict some known TFBS, but it is likely that more complex models, such as are implicitly embedded in the best TFBS prediction tools, are required for good results. The use of more complex models will be left to future work.

This thesis seeks to explore the following question:

Using simple mathematical models incorporating both frequency and positional specificity for DNA motifs, can the problem of *de novo* TFBS discovery for co-regulated genes be meaningfully treated as a model selection problem?

Methods and Materials

Three simple models will be considered. For simplicity, each model will use strings to represent candidate motifs and will represent a different positional distribution of occurrences of each candidate motif, relative to the nearest TSS, within the set of DNA promoter regions being analyzed. Though an exact match with the candidate motif will be required, reverse, complement, and reverse-complement patterns are treated as equivalent to the candidate motif to allow a small amount of biologically relevant flexibility. The three models will be referred to as “the uniform model”, “the normal model”, and “model 3”. The models will be described in subsequent sections. Since each model represents a hypothesis about the positional distribution of candidate TFBSs, the models will represented as h_1, h_2, and h_3 respectively. The position of occurrences of each candidate motif within any given promoter region will be represented as a position x relative to TSS. Positions upstream of the TSS for the gene of interest are defined to have negative values of x; positive values of x are downstream of TSS.

Assumption of the Uniform Model

The uniform model is the simplest of the three models and assumes that positions of occurrences of matches to the candidate motif are uniformly distributed. Represented mathematically, the underlying assumption of model h_1 is a position probability mass function

$$P_{h_1}(x) = 1/(l - k + 1)$$

where l is the common length of all promoter regions, or strands, being analyzed and k is the length of the motif. Because this model assumes that the probability of occurrence of the motif is independent of position, this model corresponds to the case of a uniform background “noise” of occurrences of the motif in question.

Assumption of the Normal Model

The normal model represents the hypothesis that occurrences of matches to the motif are normally distributed with an unknown mean μ and standard deviation σ, or

$$P_{h_2}(x) = \int_{y=x-\frac{1}{2}}^{x+\frac{1}{2}} \frac{e^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} dy$$

A motif that exhibits strong positional preference would have a set of occurrences, or a “signal”, clustered about the position $x = \mu$. In principle, μ is unbounded, but, given the nature of the problem, reasonable bounds can be imposed at the ends of the strands. Similarly, a reasonable upper bound can be imposed on σ, namely $\sigma \leq l$.

Assumption of Model 3

Model 3 is the most complex of the models being considered. This model corresponds to a normal signal with some amount of uniform noise and literally combines the uniform and normal models by assuming that the occurrences of matches to the given motif are the result of some linear combination of the uniform and normal models, or

$$P_{h_3}(x) = u\, P_{h_1}(x) + (1-u)P_{h_2}(x)$$

where $u \in \mathbb{R} \mid 0 \leq u \leq 1$ and is a measure of how much uniform character is present. This model explicitly adds a third unknown real-valued parameter to the set of unknown real-valued parameters implicitly inherited from the normal model.

Assumed Prior

To compare the quality of fit of data sets to the models and thereby enable us to choose the model that best explains the data, we will find the maximum-likelihood parameters for each model and calculate the corresponding model likelihoods. Recall that the posterior probability of any given model h_i is given by Bayes rule

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{P(D)}$$

where D is the data, or evidence; $P(D|h_i)$ represents the likelihood of h_i given D, the strength of the evidence supporting model h_i; $P(h_i)$ is our subjective prior; and $P(D)$ is effectively a normalizing constant. Notice that under the assumption that each model is equally probable (i.e., $P(h_i) = P(h_j)\ \forall i,j$), Bayes rule reduces to

$$P(h_i|D) = c\, P(D|h_i)$$

where c is a constant independent of the model but dependent on the data D.

Maximum Likelihood Parameter Estimation

Let Θ represent a parameter vector for any one of the three models and Θ^* represent a maximum-likelihood parameter vector for a given model. The maximum likelihood for model h_i is then directly proportional to $P(D|h_i, \Theta^*)$. Once Θ^* is determined for the given D, we can perform a maximum-likelihood model selection by directly comparing values of $P(D|h_i, \Theta^*)$ for each model h_i. Other measures, such as the Bayesian information criterion (BIC) or Akaike's information criterion (AIC) or related measures, could be used instead but are not since the maximum-likelihood model parameters and maximum likelihood are to be part of a feature set for a non-linear classifier.

Based on the previously described position probability mass functions for the three models, we can represent $P(D|h_i, \Theta^*)$ in closed form for each model. Letting n_x represent the total number of occurrences of a given motif at position x in the set of strands corresponding to D and $n = \sum_x n_x$, we obtain

$$P(D|h_i, \Theta^*) = \prod_x \binom{n}{n_x} \left(P_{h_i}(x)\right)^{n_x} \left(1 - P_{h_i}(x)\right)^{n-n_x}$$

For computational convenience, we use the standard technique of working with log-likelihoods. For the uniform model, the parameter vector has dimension zero and, thus, the determination of Θ^* for this model is trivial. For the normal model, the maximum-likelihood parameters are simply

$$\mu = \frac{\sum_x x\, n_x}{n}$$

and

$$\sigma = \sqrt{\frac{\sum_x (x\, n_x - \mu)^2}{n}}$$

For model 3, the determination of Θ^* is significantly more difficult and requires exploration of the three-dimensional parameter space to maximize the likelihood function. A modified Bees Algorithm, using the log-likelihood function as a fitness function, is seeded with a reasonable estimate of the parameters and is used to explore the parameter space until it converges to a local maximum. Though not proven, evidence will be presented supporting the assumption that the maximum found is the global maximum.

Since model 3 reduces to the normal model when $u = 0$, estimates for h_3's maximum-likelihood parameters can be obtained by assuming $u = 0$ and calculating the mean and standard deviation for D as in the normal model, but these estimates might vary significantly from the true Θ^*, particularly if the maximum-likelihood value of u isn't close to 0. To obtain a better estimate, consider the effect of increasing u from zero toward one. As u is increased, the contribution of the uniform model is increased thereby increasing the expected number of occurrences of the given motif at every position. The expected portion of the motif occurrences that are then attributed to the normal model at each position x must necessarily decrease by a constant amount for a given choice of u. If all of the motif occurrences attributable to the uniform model could be correctly subtracted, then the maximum-likelihood choices of μ and σ could be readily calculated as for the normal model. Obviously, this is not possible in general, but we can apply the idea to obtain an initial estimate of Θ^* by arbitrarily assuming that the median n_x is the level of the uniform background. Call this uniform background level n_b. This then implies that

$$u \approx \frac{n - \sum_x \max(0, n_x - n_b)}{n}$$

and estimates of μ and σ can be calculated as just described to obtain an estimate of Θ^*. Since the parameter space for model 3 is three dimensional, regions within the parameter space defined by bounds on each parameter independently correspond to right rectangular prisms. Though not technically correct, we will refer to such regions as "cubes" for convenience.

In the vicinity of the best estimate, a swarm of 125 points is generated. To avoid local maxima, 125 additional points are selected at equal intervals to span the cube of interest, initially the full parameter space. The corresponding log-likelihood for each point in the swarm is then calculated, and the points in the swarm are rank-ordered according to fitness. The two points in the swarm that are most fit for each parameter are then used to define bounds for the next cube of interest and the process is repeated until the swarm converges to a solution. Since convergence is never slower than penta-section of the parameter space, the number of repetitions is bounded above by $\log_5 l/\varepsilon$, where ε is a constant representing the maximum numeric precision required. For $\varepsilon = 1.0e-14$ and $l = 3300$, $\log_5 l/\varepsilon \approx 25$, or approximately 25 repetitions.

The models, a library of the log-based functions used to calculate the model likelihoods, and the algorithms to optimize the model parameters were implemented in Java. Perl was initially considered because of the ease with which DNA sequence data can be manipulated as strings, but this language was rejected for this use due to the challenges that Perl's "loose typing" model presented in maintaining numeric precision and avoiding computational errors. C/C++ was also briefly considered, but Java's far superior platform-independence and high-quality standard math libraries made Java the natural choice. The math libraries, particularly a good method for generating samples from a normal distribution, were critical to constructing classes to generate random sets of positional tallies from known distributions to verify the proper operation of the classes encapsulating the models, especially the normal model and model 3, and their likelihood calculations and parameter optimization algorithms. The class for each model includes a test method to provide verification that θ^*, or, in the case of model 3, at least a locally optimum θ, is being correctly determined.

Confirmation of Likelihood Maximization

The following figures show examples of how log-likelihood for each model varies over the parameter space given typical randomly generated data sets. Markers are colored by log-likelihood from blue to red, from least to greatest. The maximum likelihood encountered in the parameter space is shown in green to highlight its location.

For the Uniform Model

For the uniform model, the log-likelihood function is dependent only on $P_{h_1}(x)$ and n and n_x determined from D. Unsurprisingly, varying $P_{h_1}(x)$ such that the expected n differs from the actual n has a significant effect on the log-likelihood. Figure #1 shows results for a typical randomly-generated uniformly distributed data set. On the x-axis, "p-adj" is an additive term, and translational adjustment, of $P_{h_1}(x)$ from its optimal value.

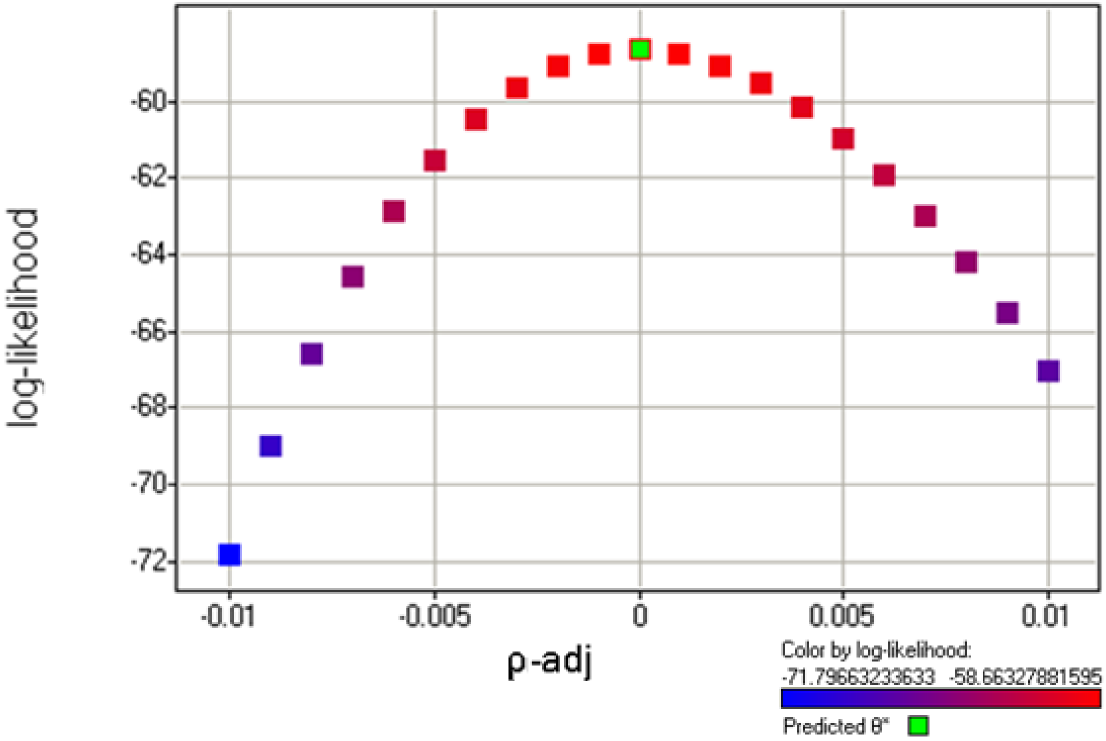


Figure #1: Uniform Model Log-Likelihood Near the Predicted θ^*

For the Normal Model

For the normal model, μ and σ are verified to be the maximum-likelihood parameters. Given the exponential nature of the probability mass function's dependency on μ and σ, the smoothness and concavity of the log-likelihood function are unsurprising. Figure #2 shows results for a typical randomly-generated normally distributed data set. Similar to the axes in the previous figure, "mean-adj" and "stddev-adj" are additive terms used to translate on μ and σ. This choice of axes highlights that the choice of parameters that maximizes the log-likelihood does not differ from the predicted θ^*.

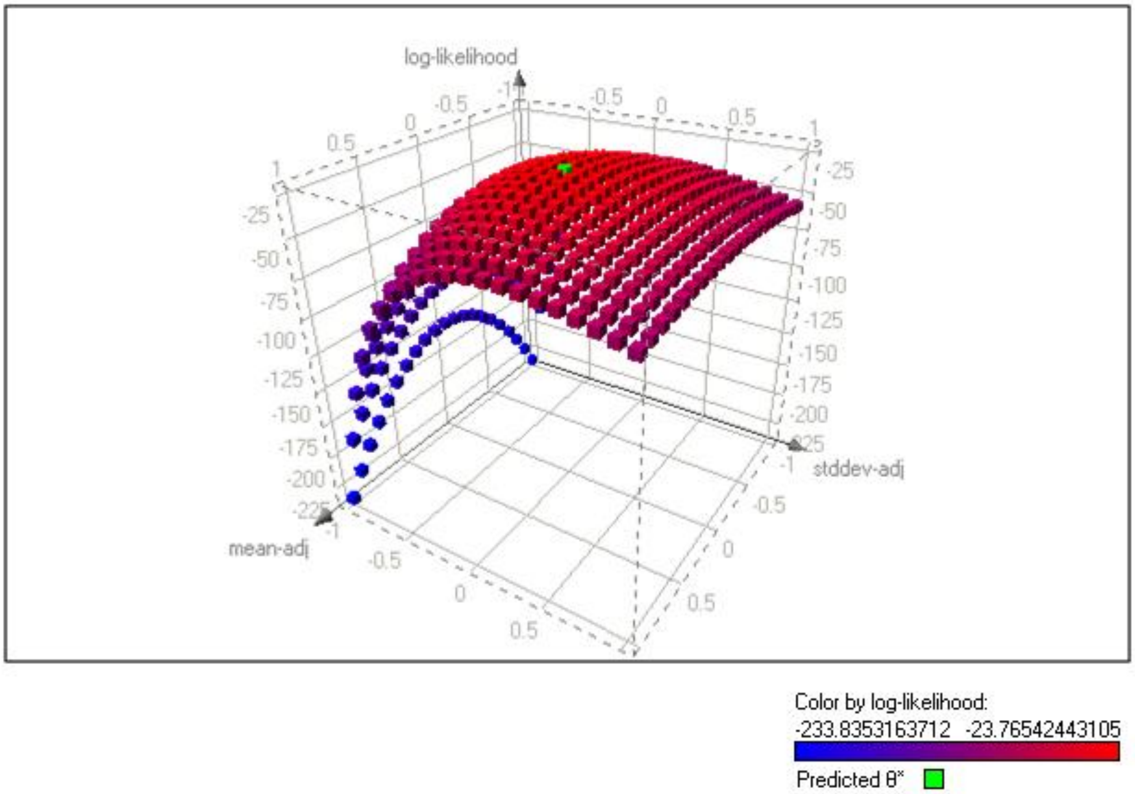


Figure #2: Normal Model Log-Likelihood Near the Predicted θ^*

For Model 3

To visualize the model 3's four-dimensional log-likelihood function, we'll look at each of the two-dimensional projections. Figure #3, Figure #4, and Figure #5 show different projections of the same results for a typical randomly-generated data set generated with equal contributions of normal and uniform character. Similar to the axes in the previous figures, "u-adj", "mean-adj", and "stddev-adj" are additive terms used to translate u, μ, and σ and highlight that the predicted θ^*produce at least a locally maximized value of the model's log-likelihood.

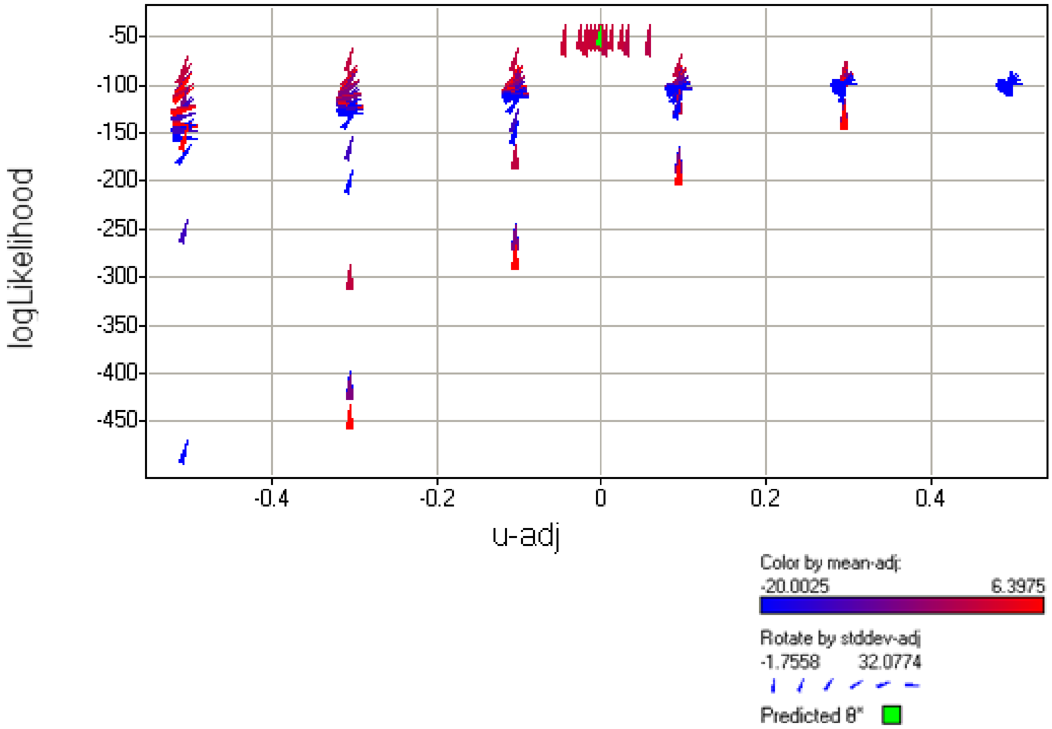


Figure #3: Model 3 Log-Likelihood vs u

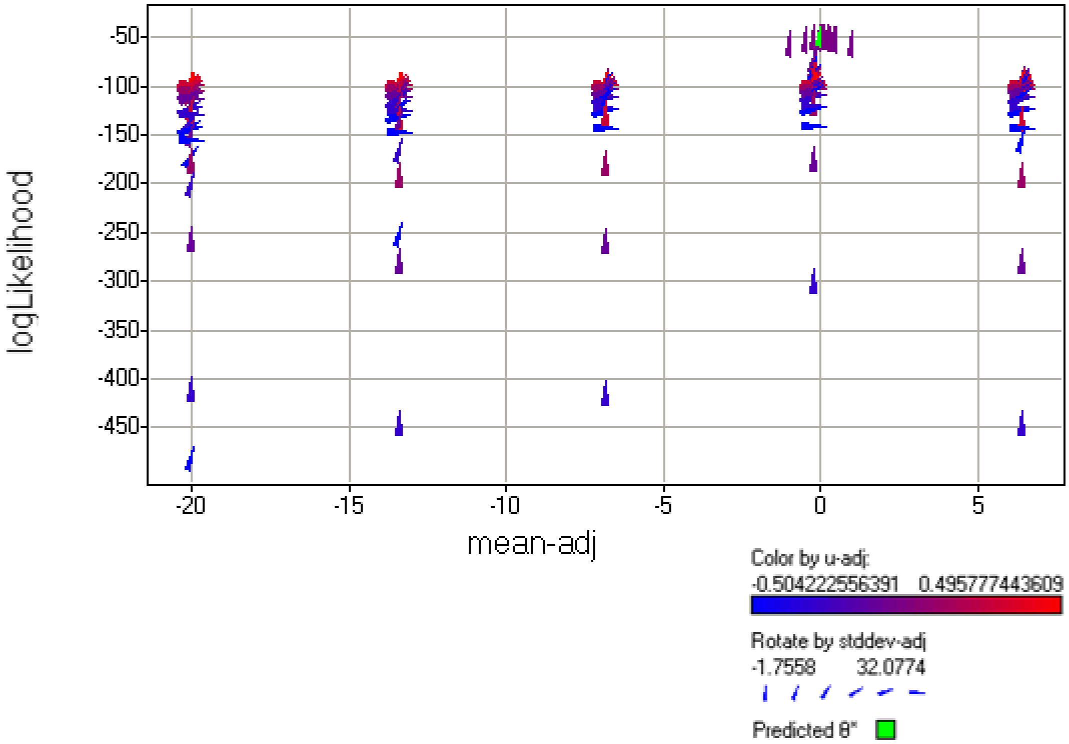


Figure #4: Model 3 Log-Likelihood vs μ

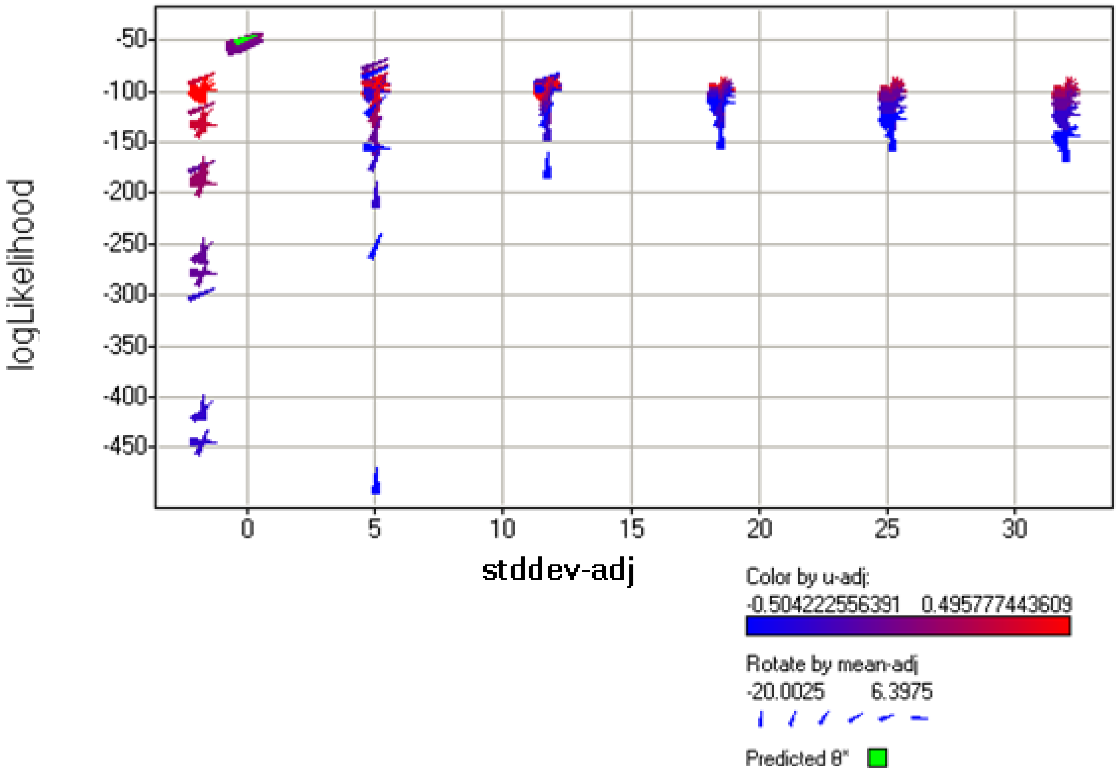


Figure #5: Model 3 Log-Likelihood vs σ

Observe that the log-likelihood for model 3 seems particularly sensitive to the choice of u and σ but is less sensitive to changes in μ. Note also that the log-likelihood trends upward as u and σ increase. The data set fit here was generated from a known distribution with $u = 0.5$, a significant amount of uniform character. The resulting positional scattering of occurrences, though best fit by only one choice of θ^*, can be well fit by any set of parameters that does not include too small a choice of σ or too small a choice of u.

The next figure, Figure #6, shows the set of points in the parameter space that have been tested during the parameter optimization process for a particular data set. As in the previous figures, the parameters have been translated such that the predicted θ^* is located at (0,0,0). The points sampled can be seen to span the parameter space with a cluster near (0,0,0). Note also that the points sampled are not exclusively on a three-dimensional grid. Typical of a “Bees” approach, the algorithm favors points near the best estimate discovered thus far by taking additional samples in its vicinity. In the middle of each iteration, a set of points, each of which is a small distance in a pseudo-random direction from the best estimate, is tested. These additional samples are the points that are not aligned to the grid on which the other points lie.

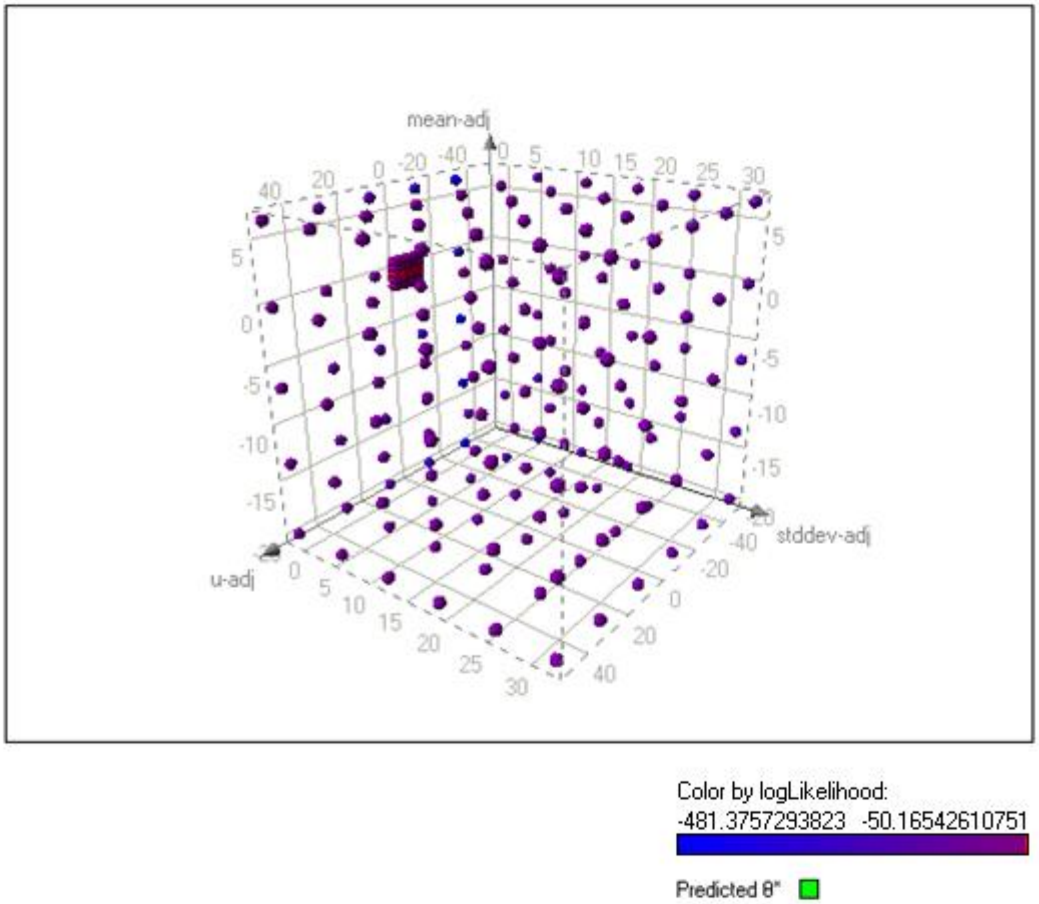


Figure #6: Model 3 Log-Likelihood Over the Full Parameter Space

Zooming into the apparently solid cube near $(0, 0, 0)$, the iterative nature of the optimization algorithm is visible. Again, observe that the samples are not exclusively drawn from the parameter space along the lines of penta-section. This is due to the additional sampling conducted in the vicinity of the current best estimate during the subsequent iteration.

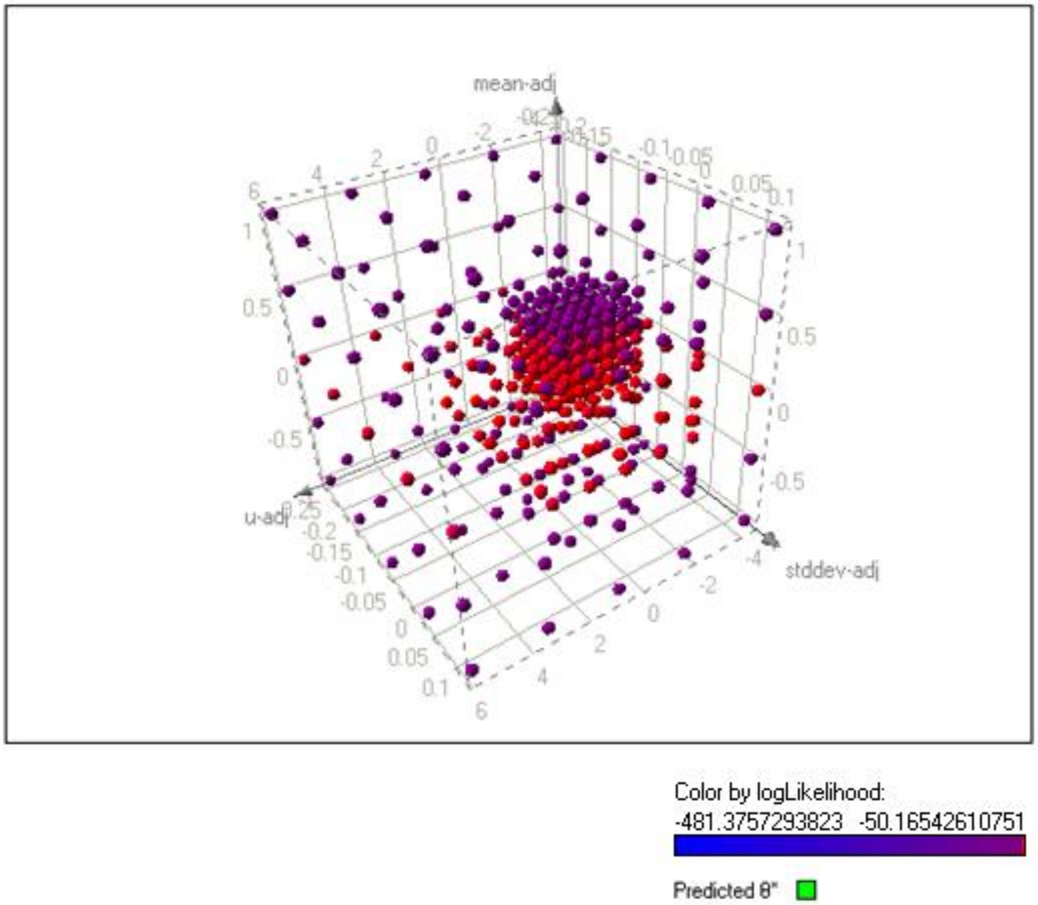


Figure #7: Model 3 Log-Likelihood Nearer the Predicted θ^*

In subsequent iterations, the estimate of the maximum-likelihood parameters is refined. The points sampled in the subsequent iterations are located within the region of dense sampling in Figure #7. Looking at (0,0,0) in Figure #8, we find the maximum likelihood parameters, as expected.

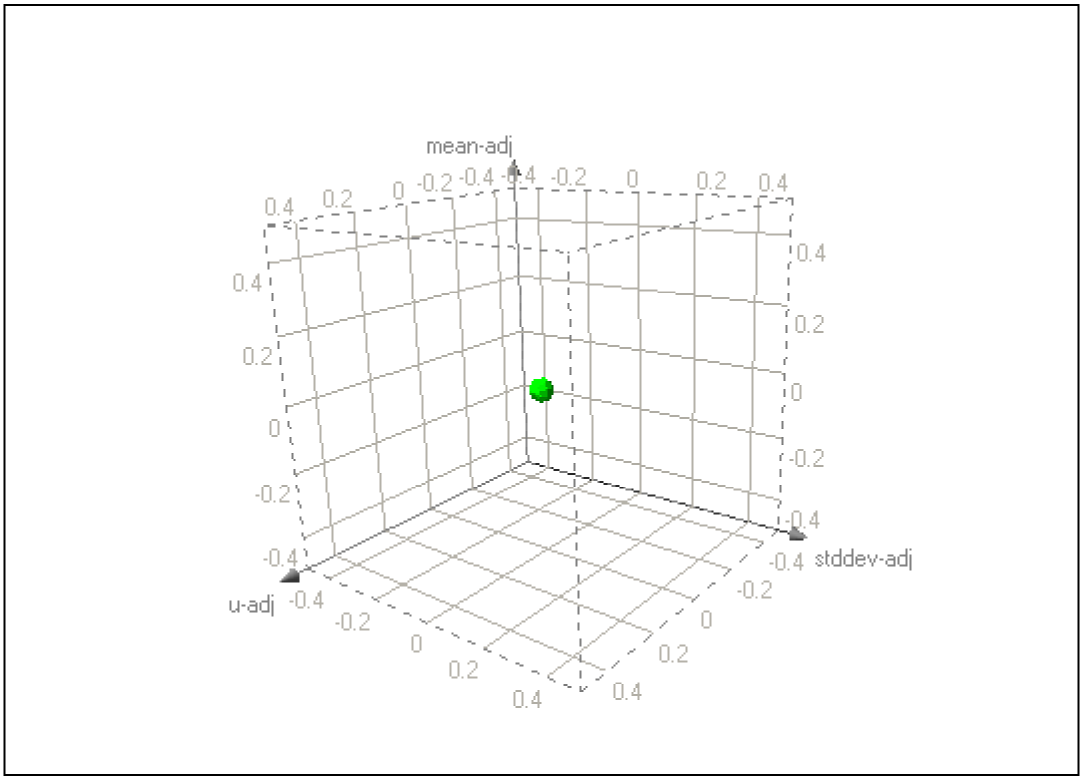


Figure #8: Model 3 Maximum Log-Likelihood Found at the Predicted θ^*

Classifier Construction and Verification

Distinguishing Data Sets from the Uniform Model, Normal Model, and Model 3

Once the maximum-likelihood parameters and maximum log-likelihood for each model have been determined for the occurrences of each candidate motif, selecting the model with the maximum log-likelihood yields the maximum-likelihood model selection. The log-likelihood ratio provides an additional measure of information by quantifying how much more likely one model is than another. Particularly if a candidate motif shows strong positional preference, this information alone might be sufficient to identify interesting candidate motifs.

To provide additional verification that the models and algorithms are behaving as expected, a set of 112136 sets of randomly generated positional tally data was generated and analyzed. Each set was generated using one of the three models and a known set of parameters. The parameters were selected from pseudo-uniform distributions over the ranges in Table #7.

Table #7: Model Parameter Bounds for Model and Algorithm Verification

Parameter	Minimum	Maximum
n	1	1000
μ	-3000	300
σ	0.0	600.0
u	0.0	1.0

The three models were fit to each of the sets, and the known "actual" model and parameters and the value of θ^* obtained for each model for each data set were recorded. It is worth noting that the modified Bees algorithm to find θ^* for model 3 typically converged in six to twelve iterations. Though a systematic analysis of error was not performed, simple linear regression comparing maximum-likelihood and known parameters was performed.

Table #8: Verification of $\boldsymbol{\theta}^*$

Model	Parameter	R value
Normal	μ	0.996
	σ	0.951
Model 3	μ	0.896
	σ	0.717
	u	0.943

The relatively low R values for known versus maximum-likelihood values of μ and σ were noticeably dependent on n. As tally sets were excluded from the calculation by progressively more stringent minimum values of n, model 3's predicted parameter values tended to better correlate with the actual values. Excluding tally sets with progressively smaller actual values of u produced a similar result.

Satisfied that the maximum-likelihood parameters are being correctly calculated, we consider the type of classifier that we will use to identify interesting candidate motifs. If a candidate motif shows strong positional preference, simply knowing which model has the greatest log-likelihood for that motif might be useful. Typically, to gain a relative measure of the likelihood of one model relative to another, the log-likelihood ratio (LLR) is calculated. Figure #7 shows a scatter plot of model 3's likelihood relative to the other two models as a means to visually separate the verification sets.

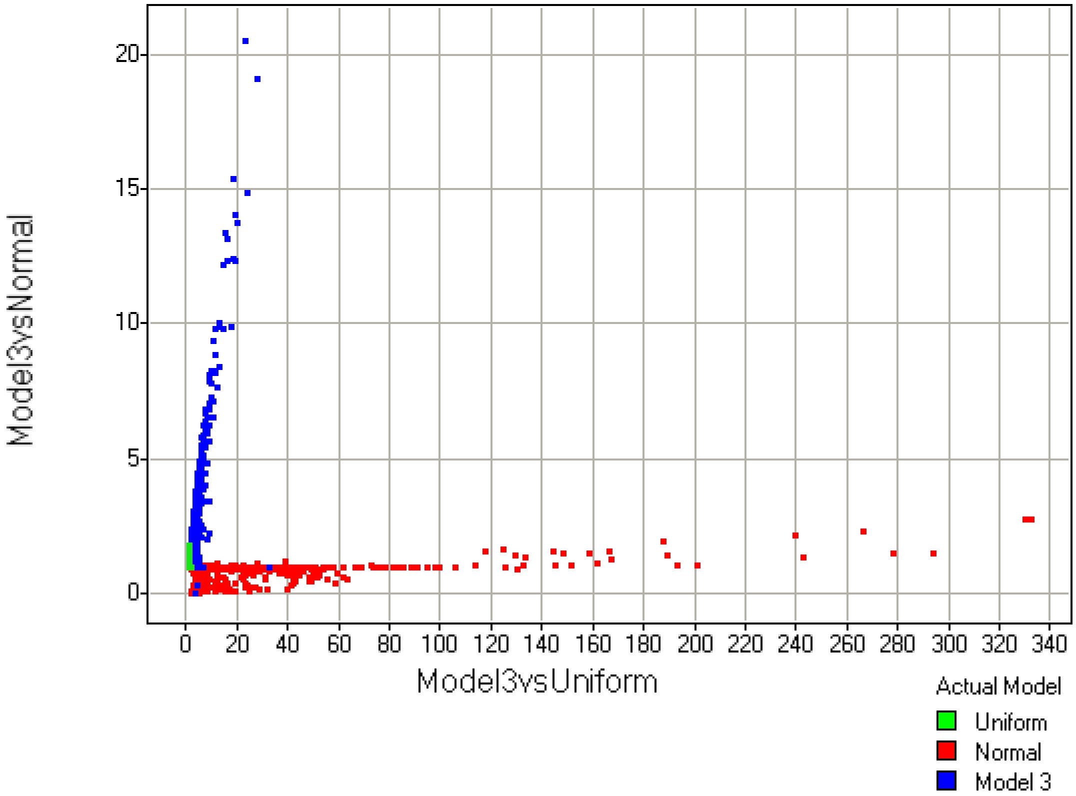


Figure #9: Model 3/Normal LLR vs Model 3/Uniform LLR

In the above figure, each marker represents a data set; with green, red, and blue representing sets from the uniform model, the normal model, and model 3, respectively. At a glance, it seems surprisingly promising that a simple linear classifier based only on these two features might reliably identify the actual model of occurrences for any given candidate motif. On closer inspection, this continues to appear likely, but the classification is not perfect. Consideration of additional features might provide superior classification.

Presuming that over-training can be avoided, the performance of a classifier based on just these dimensions should provide a reasonable lower bound for the expected performance of a classifier based on a larger feature set. Rather than make any attempt to find an optimal linear classifier in these two dimensions, we can just freehand two lines to partition the space into four regions as shown below to obtain an estimate.

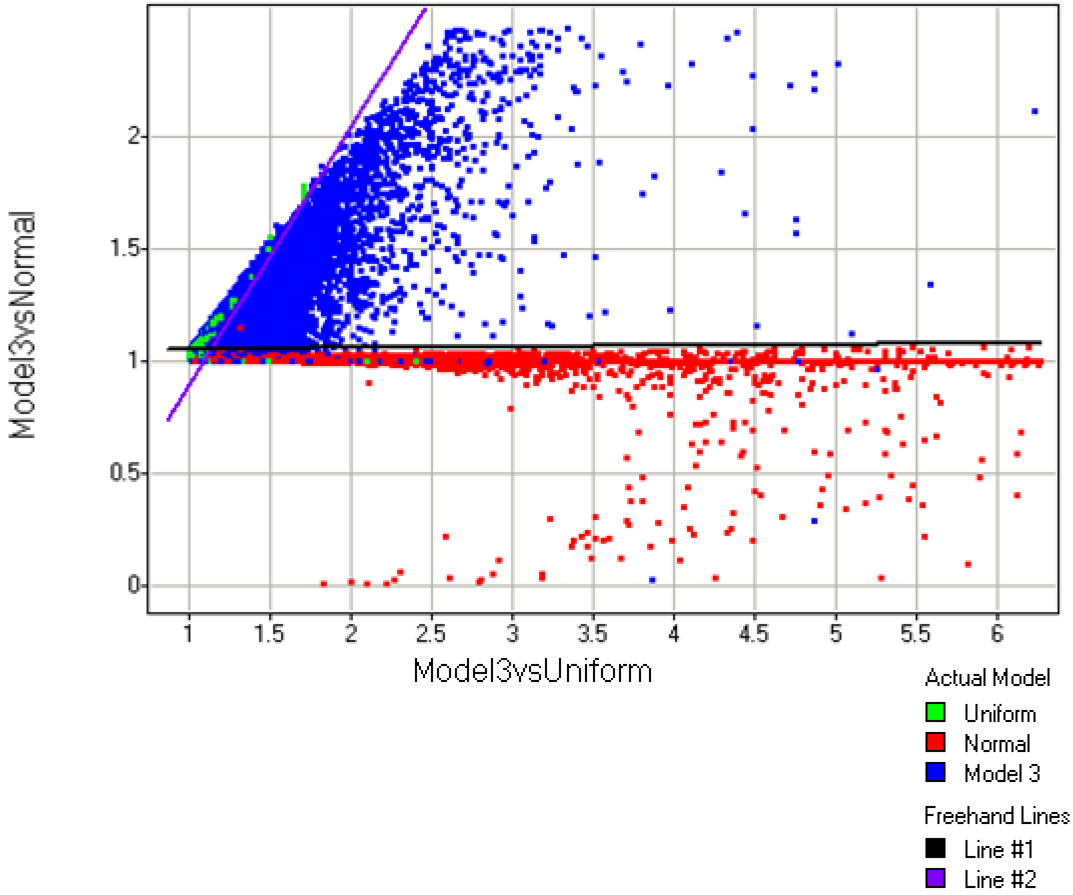


Figure #10: "A Simple Classifier"

The classifier in Figure #8 classifies all points to the left of the purple line as "uniform", all points to the right of the purple line and above the black line as "model 3", and all points to the right of the purple line and below the black line as "normal". Despite its simplicity, this classifier is reasonably accurate in classifying the sets of test data, as seen below, and achieves better than 73% accuracy.

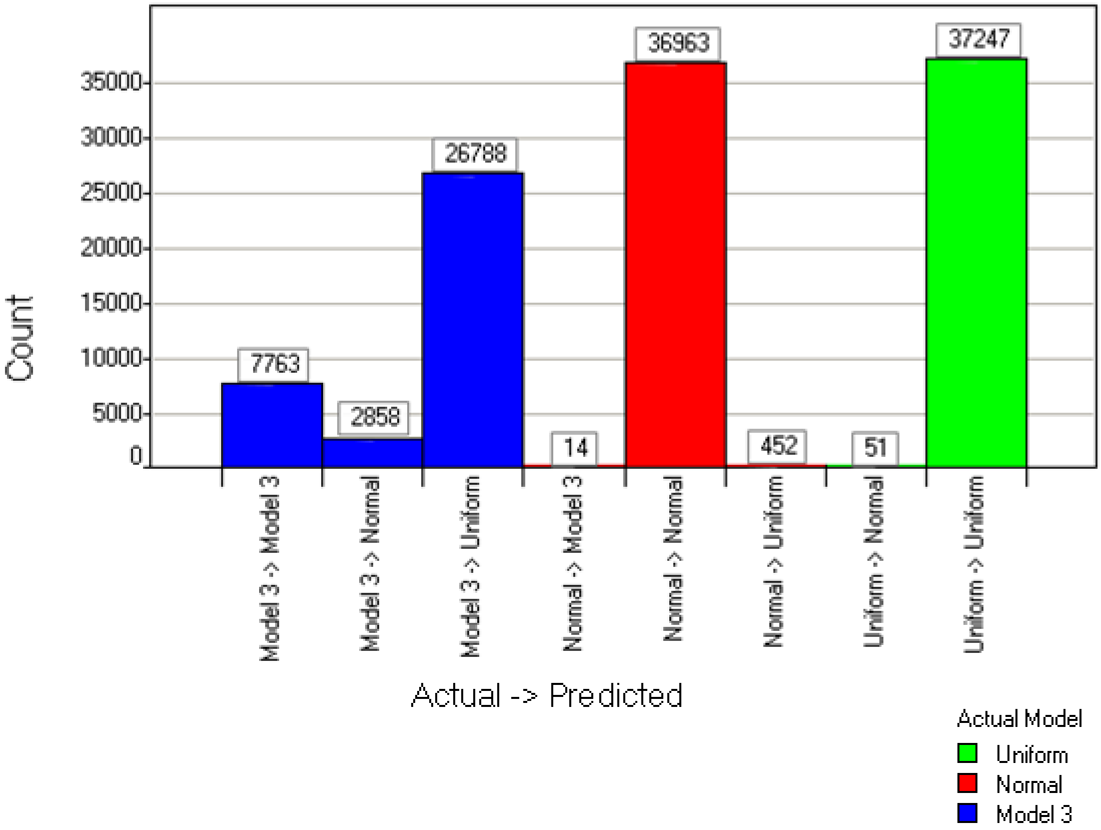


Figure #11: Classification of Verification Data Sets Using “A Simple Classifier”

Considering only test sets from the normal and uniform models, the classification is extremely good, better than 99.3% accurate. Consider the subset of the test sets that excludes test sets for which the actual model was model 3. Figure #10 shows a receiver operating characteristic (ROC) curve for a classifier operating on this subset of data. The ROC curve in Figure #10 was generated using only the normal-vs-uniform log-likelihood ratio as the predictor, we can see that the level of performance demonstrated by the simple ad hoc classifier should not be unexpected; test sets from these two models are generally easily distinguished from one another.

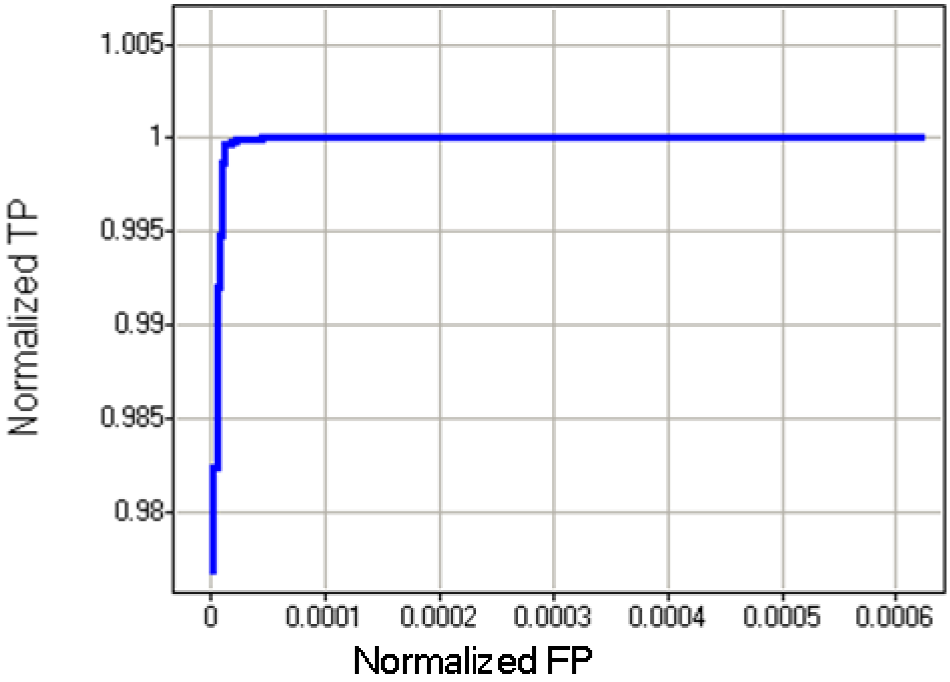


Figure #12: ROC Curve for LLR-Based Classification of Normal and Uniform Sets.

To provide more accurate classification, we choose to use a support vector machine (SVM). Rather than create our own SVM implementation, LibSVM 2.86 (Chang) is used. To gauge the potential improvement in classification accuracy, a radial-basis function (RBF) kernel with default parameters is used initially. The feature set used is shown below.

SVM #1 Features (13 total features):

- Actual Model (required for training but not a feature)
- n
- Uniform Model output
  - Log-Likelihood
- Normal Model output
  - Log-Likelihood
  - Mean
  - Standard deviation
- "Model 3" output
  - Log-Likelihood
  - % Uniform
  - Mean
  - Standard deviation
  - "n of signal" (defined to be $n * (1 - u)$)
- Log-Likelihood Ratios
  - "Model 3" vs Normal
  - "Model 3" vs Uniform
  - Normal vs Uniform

For this default SVM, the accuracy was significantly better than the simple classifier, slightly better than 91%, when trained on the full set of test sets. For subsequent data sets, we follow the following steps to prepare our SVM, per the recommendation of the authors:

1. Transform our data to the format required by LibSVM
2. Conduct simple scaling on the data
3. Use the RBF kernel to provide the flexibility of a non-linear classifier
4. Use cross-validation to find kernel parameters to maximize accuracy while avoiding over-fitting
5. Use the best kernel parameters and train the SVM on the whole training set
6. Classify

Following the recommended procedure above yields a classifier with accuracy from five-fold cross-validation greater than 90% and accuracy of 98.5% when trained on the full set of randomly generated data sets, reliably identifying the model from which a set was generated.

Distinguishing Seeds Motifs from Background in Seeded Sequences

Though important as a validation test set, the set of randomly generated test sets has neither biological context nor relevance. To address this deficiency, sets of strands of

DNA are generated and the strands in each set are seeded with occurrences of a particular motif. As in the case of the set of sets of tally data, the set of positions for seed occurrences in each set of strands are generated using the three models. Once the seed motifs of length six or seven are generated and planted, the resulting sets of strands are analyzed. For each set, the set of unique 7-mer candidate motifs is identified, the positional occurrences of each candidate motif are tallied, and the model fitting is performed to obtain the maximum likelihood and θ^* for each model. This is analogous to running multiple TFBS discovery tools on promoter regions for sets of co-expressed genes and obtaining metrics for sets of putative motifs. The next step is to prepare an SVM to separate the seed motifs from the irrelevant candidate motifs. The features used for the second SVM are shown below.

SVM #2 Features (14 total features):

- 0 – not a seed; 1 – a seed (equivalent to “Actual Model” from SVM #1)
- Number of strands in the set
- n
- Uniform Model output
  - Log-Likelihood
- Normal Model output
  - Log-Likelihood
  - Mean
  - Standard deviation
- “Model 3” output
  - Log-Likelihood
  - % Uniform
  - Mean
  - Standard deviation
  - “n of signal” (defined to be $n * (1 - u)$)
- Log-Likelihood Ratios
  - “Model 3” vs Normal
  - “Model 3” vs Uniform
  - Normal vs Uniform

Discovering Motifs in Promoter Regions of Co-Expressed Genes

After acceptable discriminating power has been demonstrated in separating seed motifs from the irrelevant candidate motifs, promoter regions for two kinds of sets of genes determined experimentally to be co-expressed are analyzed. The first type of data set is a set of five sets, obtained from the literature, composed of promoter regions for sets of genes determined experimentally to be co-expressed in mouse thymocyte cells developmentally blockaded at different stages of development (Puthier). The second kind of set is composed of the 12 mouse and 26 human sets of type “real” from the Tompa benchmark (Tompa, Assessing computational tools for the discovery of transcription factor binding sites).

In Puthier *et al*’s “A General Survey of Thymocyte Differentiation by Transcriptional Analysis of Knockout Mouse Models”, thymocyte-laden thymus samples were collected from mice representing six mouse lines, one wild-type and five knockout models that impose different developmental blockades at various stages of thymocyte development. The stages of thymocyte maturation toward mature T-cells in the thymus, in chronological order, and the mouse models used are listed here for reference.

Lymphocyte maturation stages
- CD44$^{\text{high}}$CD25^{-} (DN1)
- CD44$^{\text{high}}$CD25^{+} (DN2)
- CD44$^{\text{low}}$CD25^{+} (DN3)
- CD44$^{\text{low}}$CD25^{-} (DN4)
- CD4^{+}CD8^{+} (DP)
- CD4^{+}CD8^{-} / CD4^{-}CD8^{+} (SP)

Mouse lines
- Wild type
  - C57BL/6
- DN enrichment
  - RAG1°
  - LAT°
  - CD3-$\varepsilon^{\Delta5}$ (CD3ε°)
- Lack of medullary dendritic cells
  - TCRα°
  - RelB°

Puthier *et al* used fluorescence-assisted cell sorting (FACS) to isolate cells in the various stages of development. RNA was then extracted from the purified samples and tested using a microarray prepared from publically available mouse cDNA libraries to obtain gene expression data. The resulting data was hierarchically clustered to obtain six clusters of preferentially expressed genes. The authors identified the clusters as being specific to thymocyte proliferation, DN T-cells, TCR rearrangement, DP T-cells, SP T-cells, and cells of the stromal compartment. Names for the genes in each of the latter five clusters were obtained from the supplemental materials.

To obtain the promoter regions for each of the genes, MGI (Bult CJ and Group) is searched for each of the gene names and, if the gene name is recognized as a valid gene name or synonym, the location of the gene in the mouse genome is retrieved. Using the location information, the promoter sequences, arbitrarily defined to be $x \in [-3000,300]$, are then retrieved from MGI for analysis. Since the true regulatory motifs, if any, are not known, we use the first of the two previously trained SVM to identify the model which best explains the occurrence of each motif in each of these sets of co-expressed genes. Motifs that demonstrate positional specificity are expected to be best explained by either the normal model or model 3 and, based on this classification, are tentatively deemed interesting. Interesting motifs that are present in one cluster but not any of the others are most likely to represent TFBSs for TFs specific to the stage of thymocyte maturation associated with the cluster. The interesting motifs are then scored for similarity with known mouse TFBS motifs from the JASPAR Core database to look for possible matches with known motifs.

To obtain the second type of co-expressed gene sets, the mouse and human "real" sets are retrieved from the Tompa benchmark website (Tompa, Assessment of Computational Motif Discovery Tools). To verify that the sequences represent the same positions relative to the nearest TSS, the locations of the retrieved sequences are determined by performing a genome-wide Blast search and matching these to the locations of the nearest known genes.

The mouse and human genomes, respectively, were downloaded from
ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/ (dated Jul 05, 2007) and
ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ (dated Mar 24, 2008)

Blast was performed locally using Blast 2.2.19 downloaded from
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/

The locations of the best hits were checked against the locations of known genes in
ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz (dated Mar 4, 2009)

Promoter regions from the Tompa sets that are offset more than 16bp from the expected TSS are discarded. The retained DNA sequences are then analyzed. The classifier trained on the seeded test sets is used to perform the classification to produce a small set of putative TFBS motifs. The best putative motif, if any, for each of the sets is then submitted to the benchmark for scoring.

Results

Results For Lymphocyte Data

To obtain maximum likelihood values and corresponding sets of maximum-likelihood parameters, the three models were fit to each candidate motif identified in the promoter regions, defined as -3000 to 300, for each set of co-expressed genes identified by Puthier *et al*. Puthier *et al* did not provide a list of the genes identified in the cluster associated with thymocyte proliferation, so this cluster was not considered. For the remaining clusters, not all of the identifiers provided in the supplemental materials were recognized by MGI or could be uniquely resolved. Gene identifiers that could not be uniquely resolved were discarded. The "catch all" cluster associated with generic "cells of the stromal compartment" was also discarded.

Table #9: Puthier *et al* Clusters and Gene Counts

Cluster Description	Cluster Identifier	Number of Genes in Cluster	Number of Genes Considered
DN T-cells	S1	39	39
TCR Rearrangement	S2	50	47
DP T-cells	S3	56	54
SP T-cells	S4	69	67

Exploring the hypothesis that the simple models themselves might be able to identify TFBS motifs, the normalized feature sets for all of the candidate motifs in each cluster were generated and classified using SVM #1. The classified results were then filtered to obtain lists of potentially interesting 7-mers, where "interesting" was defined as being equivalent to satisfying the following set of filters.

Filters to Identify "Interesting" Motifs:

- Classified as either "normal model" or "model 3" by SVM #1
- $u \leq 0.50$
- At least one occurrence in at least half of the strands of the set
- Motif not considered "interesting" in any of the other clusters

As seen in Table #10, significant numbers of candidate motifs remain in each cluster after these filters are applied. At a minimum, the full set of features used by SVM #1 might meaningfully be used to identify putative motifs among the candidate motifs. Given the large number of variables by which the candidate motifs might meaningfully be filtered and the intuitive but relatively arbitrary means by which the filters were chosen, the set of putative motifs could readily be reduced further. It seems difficult to know if any biologically relevant motifs have been retained. Though still an impractically large set for one-off experimental confirmation, comparison against known motifs provides supporting evidence that the relatively small fraction of motifs retained from the full set of candidate motifs in each cluster contains at least some biologically relevant TFBS motifs. Toward this end, the JASPAR Core database was searched for mouse motifs highly similar to the interesting putative motifs remaining in each cluster and the sets of highly similar motifs were generated.

Table #10: "Interesting" Motif Counts and JASPAR Motif Counts and Names by Cluster

Cluster Identifier	Number of 7mers, Post-Filter	7mers Matching Jaspar Motifs at >= 0.9	Jaspar Motifs Matched at >= 0.9
S1	140	5	Prrx2 Gata1
S2	87	7	Gata1 Fos
S3	110	4	Gata1
S4	98	12	Gata1 Bapx1 Hand1-Tcfe2a Sox17

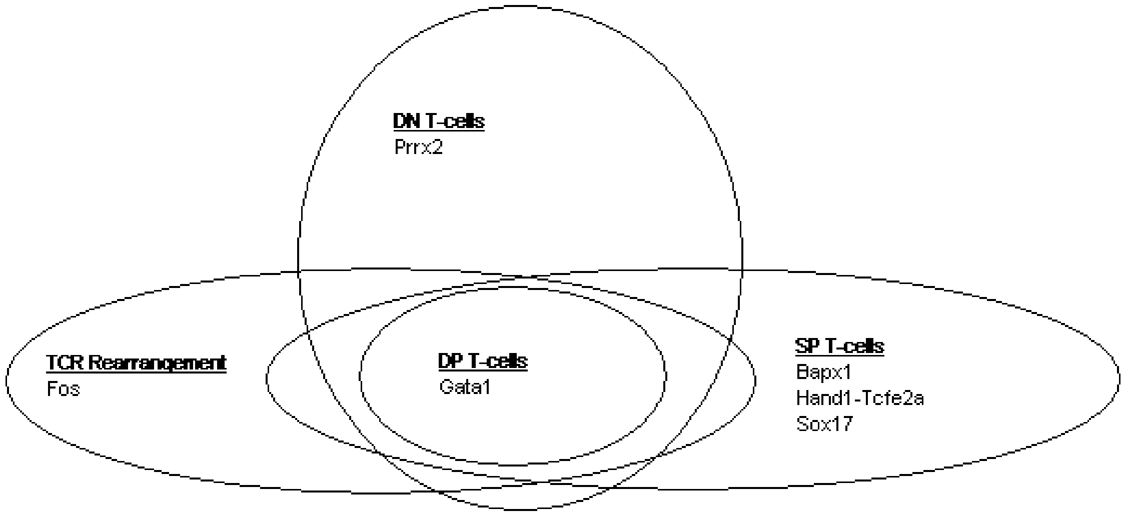


Figure #13: Venn Diagram of JASPAR Motifs for “Interesting” Motifs

Anecdotal support for the associations between the putative TFBS motifs and the sets of known TFBS motifs in JASPAR Core was obtained by electronically searching the literature. In support of the DP and DN regions of the Venn diagram, the literature suggests that T-cells “exhibit a distinct molecular expression pattern” and links this expression pattern to both the Gata1 and Prrx2 motifs. Also, based on experimental evidence, c-Fos has been associated with TCR rearrangement. Neither supporting nor contradictory evidence was obtained for the known motifs exclusive to SP T-cells in the Venn diagram. Though this is likely simply due to a deficiency in the search methodology employed, it is possible that one of more of the identified motifs play a role in SP T-cell gene expression that has not yet been experimentally identified. Given the extreme simplicity of the models being considered, this support is both surprising and encouraging.

Table #11 shows the motifs in each cluster that were identified as “Interesting” and provides the details behind Table #10 and Figure #13. When looking at the “Number of Promoters in Set With At Least One Occurrence” column, it may be beneficial to compare the number of promoter regions in which each motif was found to the total number of sequences in each set shown in Table #9.

Table #11: "Interesting" Motifs Highly Similar to Known Motifs

				Model 3			JASPAR			
Cluster	pattern	n	Number of Promoters in Set With At Least One Occurrence	u	x	s	Model ID	Model Name	Rela-tive Score	Site Sequence
S1	taattta	40	20	0.0	-1737.3	908.5	MA0075	Prrx2	1.0	AATTA
	ggatgct	23	18	0.5	-387.7	339.0	MA0035	Gata1	1.0	GGATGC
	catctat	23	18	0.2	-1345.8	911.2	MA0035	Gata1	0.9	AGATGN
	aaacatc	30	20	0.2	-1218.6	851.5	MA0035	Gata1	0.9	NGATGT
	gattgtg	26	18	0.0	-1774.5	822.6	MA0035	Gata1	0.9	NGATTG
S2	gatggga	39	25	0.3	-1350.0	939.0	MA0035	Gata1	1.0	NGATGG
	tggatgc	32	24	0.2	-843.8	992.4	MA0035	Gata1	1.0	GGATGC
	accaatc	33	23	0.2	-1899.0	955.2	MA0035	Gata1	0.9	NGATTG
	tgagtaa	39	23	0.1	-1759.3	920.5	MA0099	Fos	0.9	NTGAGTAA
	tatcctg	36	25	0.4	-2198.5	778.0	MA0035	Gata1	0.9	GGATAN
	gaaggat	37	25	0.2	-1685.7	1002.0	MA0035	Gata1	0.9	GGATNN
	aagggat	36	25	0.1	-1612.0	901.9	MA0035	Gata1	0.9	GGATNN
S3	ccatcac	38	28	0.2	-1397.8	927.5	MA0035	Gata1	1.0	TGATGG
	gatgatg	35	29	0.2	-1636.5	926.9	MA0035	Gata1	0.9	TGATGN
	aagatac	44	27	0.3	-1698.5	863.7	MA0035	Gata1	0.9	AGATAC
	agatgac	55	36	0.1	-1759.2	925.1	MA0035	Gata1	0.9	AGATGA
S4	cagcatc	46	34	0.4	-1076.0	832.3	MA0035	Gata1	1.0	NGATGC
	tgagatg	59	42	0.3	-1270.2	935.3	MA0035	Gata1	0.9	AGATGN
	aacatcc	46	33	0.4	-2150.2	810.5	MA0035	Gata1	0.9	GGATGT
	cacttag	57	36	0.3	-1339.6	918.9	MA0122	Bapx1	0.9	CTAAGTGNN
	gtcatcc	53	38	0.4	-1852.2	897.8	MA0035	Gata1	0.9	GGATGA
	tggatga	56	38	0.4	-1802.9	921.7	MA0035	Gata1	0.9	GGATGA
	ttgagtg	101	51	0.3	-1823.9	887.0	MA0122	Bapx1	0.9	TTGAGTGNN
	gaagtgg	69	50	0.4	-979.4	971.9	MA0122	Bapx1	0.9	NGAAGTGGN
	tgccaga	53	39	0.4	-1370.0	954.3	MA0092	Hand1-Tcfe2a	0.9	NNTCTGGCAN
	attgtgt	69	39	0.4	-1786.7	991.7	MA0078	Sox17	0.9	NNNATTGTG
	caaggat	51	34	0.5	-1160.7	643.8	MA0035	Gata1	0.9	GGATNN
	atcctgt	60	40	0.4	-1254.2	830.2	MA0035	Gata1	0.9	GGATNN

Results For Seeded Sequences

For use in training SVM #2, 256 sets of random DNA sequences were generated and seeded with known 7-mer motifs to simulate promoter regions for sets of co-expressed genes. Similar to the Tompa benchmark sets, sequences of length 2007 were used to allow putative motifs to be identified in the region [-2000, 0] relative to TSS.

Sets of between one and thirty-five sequences were considered. This matches the range of numbers of genes per set in the Tompa benchmark and was thus taken to be a biologically and experimentally relevant range of set sizes to be considered.

A GC content of 50% was assumed both for the sequences and the 7-mer seed motifs. One seed motif was planted per set. The frequency and locations of the seed motif for each set were dictated by random samples from a distribution corresponding to one of the three models and with the necessary model parameters, if any, selected uniformly from the ranges in Table #7. Once the 7-mer seeded sets of sequences were generated, all candidate motifs were identified for each set, and the occurrences of each candidate motif were fit to the three models to obtain the maximum likelihood value and corresponding sets of maximum-likelihood parameters for the motif in the set. To the obtained data set, an attribute was added to indicate whether or not the given motif matched the seed motif for that set. In total, this constituted a training set of 938408 points, 938152 negatives and 256 positives, for training SVM #2.

Per the procedure suggested by LibSVM's authors, the training set was converted to a normalized feature set and formatted to match LibSVM's required input format. Ensuring that the SVM is no more complex than required is important to reduce the likelihood of over-fitting. For the RBF kernel, two parameters, c and g, may be specified to adjust the complexity of the projection of the data to a higher dimension space and the cost of adding additional support vectors to the classifier. Though not stated explicitly, the authors imply that reasonable ranges are $c \in [2^{-5},\ 2^{15}]$ and $g \in [2^{-15}, 2^{3}]$. To obtain a good estimate of the best SVM parameters, three-fold cross validation was used to identify regions of the RBF parameter space that provided optimal accuracy during the cross validation. The normalization constants shown in Table #12 were used to maintain a consistent scale for the parameters to SVM #2. These values were selected because they were simple fractions that scaled the training set data very near to a range of $[0, 1]$.

Table #12: Normalization Constants Used for Input to SVM #2

Feature Category	Feature	Normalization Constant
Set-specific	Cardinality of the set	1/35
	n for the candidate motif	1/125
Uniform model	Log-likelihood	1/500
Normal model	Log-likelihood	1/500
	x	1/2000
	s	1/1000
Model 3	Log-likelihood	1/500
	x	1/2000
	s	1/1000
	u	1
Derived features	"n of signal"	1/125
	Model3vsNormal	1
	Model3vsUniform	1
	NormalvsUniform	1

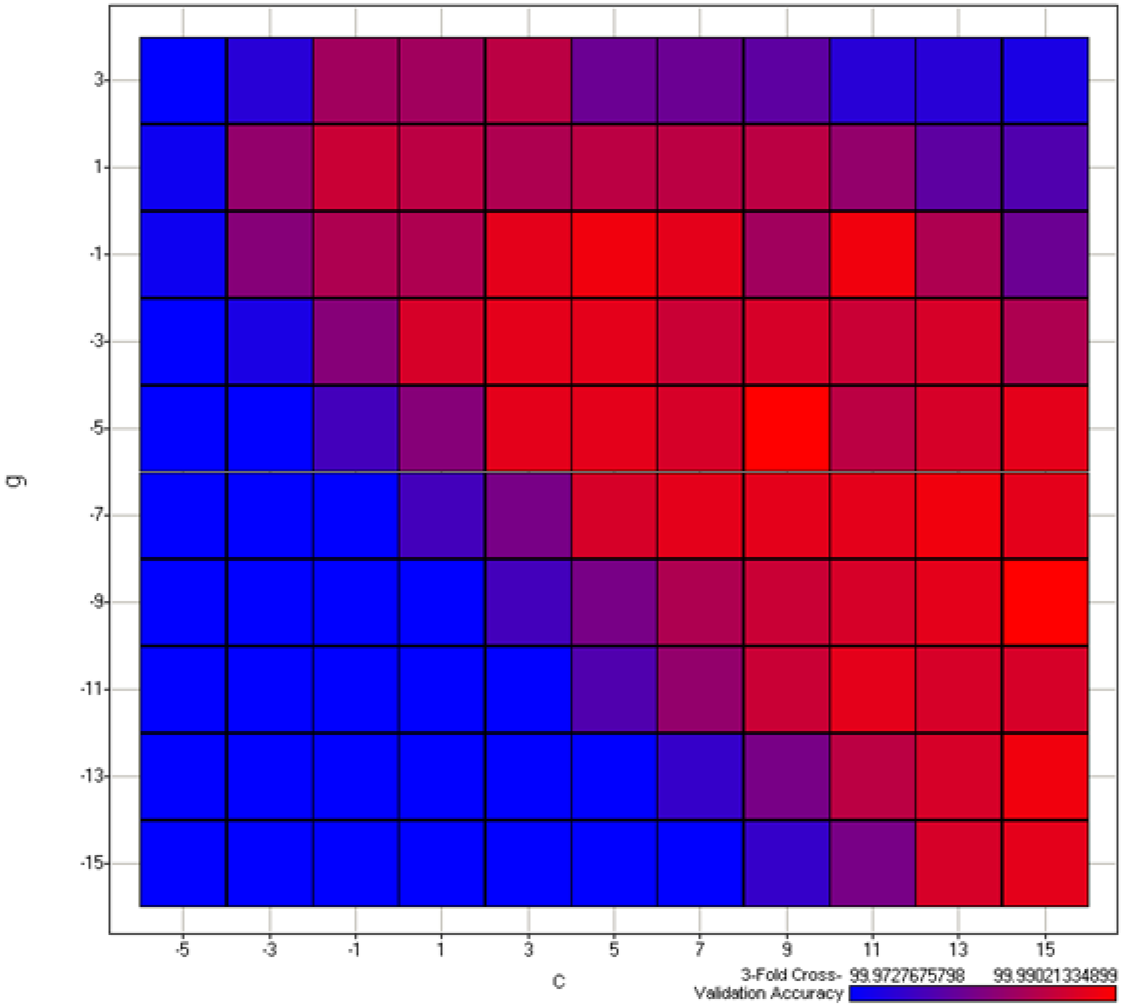


Figure #14: Optimization of Kernel Parameters g and c for SVM #2 using “Seeded” Random DNA Sequences

Notice that the point $(9, -5)$ in Figure #12 is the brightest red. This point showed the highest 3-fold cross-validation accuracy during training of SVM #2. Based on the results of the exploration of the parameter space, the kernel parameters $c = 2^9$ and $g = 2^{-5}$ were selected. SVM #2 was then generated by training on the full set of training data using these kernel parameters. SVM #2 was then used to classify the training set, and the classification results were compared to the known values. To translate into standard information retrieval concepts, we consider the seed motifs to be “positives” and non-seed motifs to be “negatives”. Since SVM #2 seeks to classify motifs as either seeds or non-seeds, correct classifications are “true” while incorrect classifications are “false”.

Following these naming conventions, Figure #13 shows the performance of SVM #2 classifying the seeded sequences of DNA.

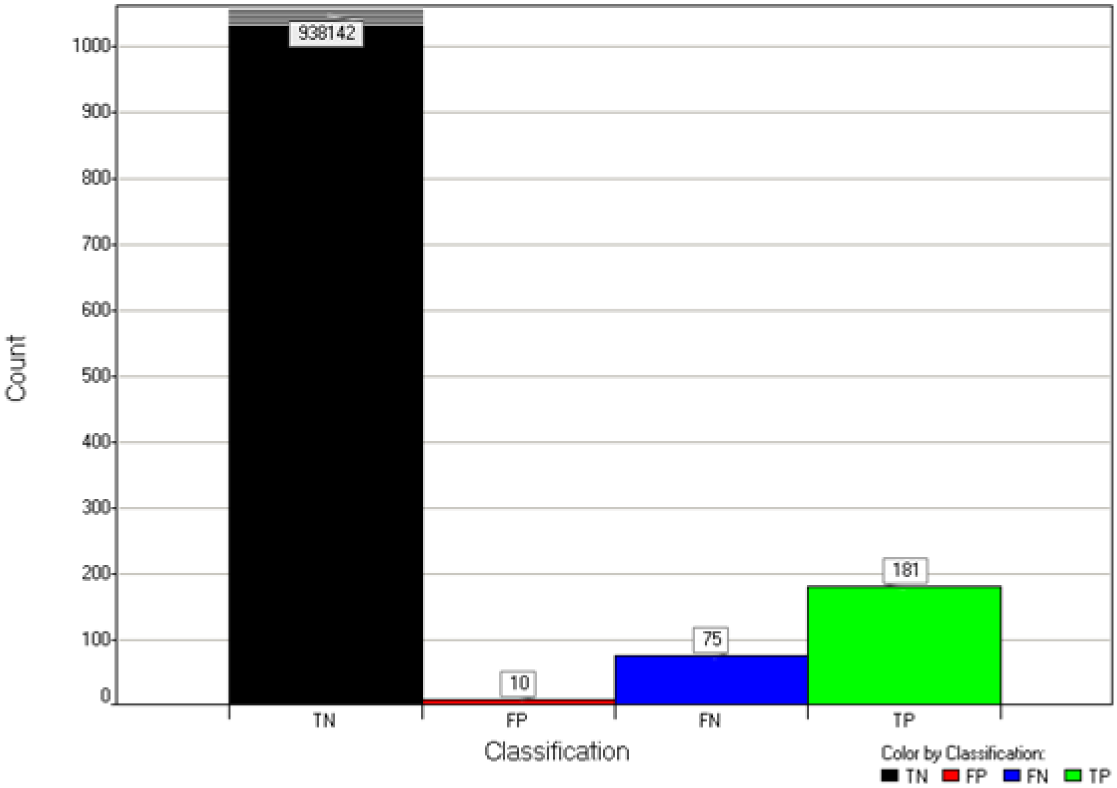


Figure #15: SVM #2 Classification Accuracy for Candidate Motifs (7-mer Seeds)

Using the tallies shown in Figure #13, a common set of information retrieval statistics were calculated to better characterize the performance of the classifier. Table #13 shows the calculated values of these statistics.

Table #13: SVM #2 Information Retrieval Metrics For 7-mer Seeds

Metric	Count
Precision (PPV)	0.94764
Sensitivity (Sn)	0.70703
Specificity (Sp)	0.99999
NPV	0.99992
F-score	0.80984

Unsurprisingly, SVM #2 performs extremely well at classifying the set on which it was trained, as demonstrated by the information retrieval metrics in Table #13. The extremely high values of the specificity and negative predictive value (NPV) metrics are due in part to the small fraction of positives relative to negatives. Though the set could have been constructed such that this fraction would have been closer to 0.5, such a training set would less accurately reflect the sets that would be analyzed in practice. The other metrics, which are less influenced by this feature of the training set, are also demonstrative of the good performance of SVM #2 at classifying this set. Viewed on the same axes that were used when viewing the results for model identification using SVM #1, Figure #14 seems to be a favored region for positives, but, as expected, positives and negatives seem to be intermixed when viewed according to these dimensions.

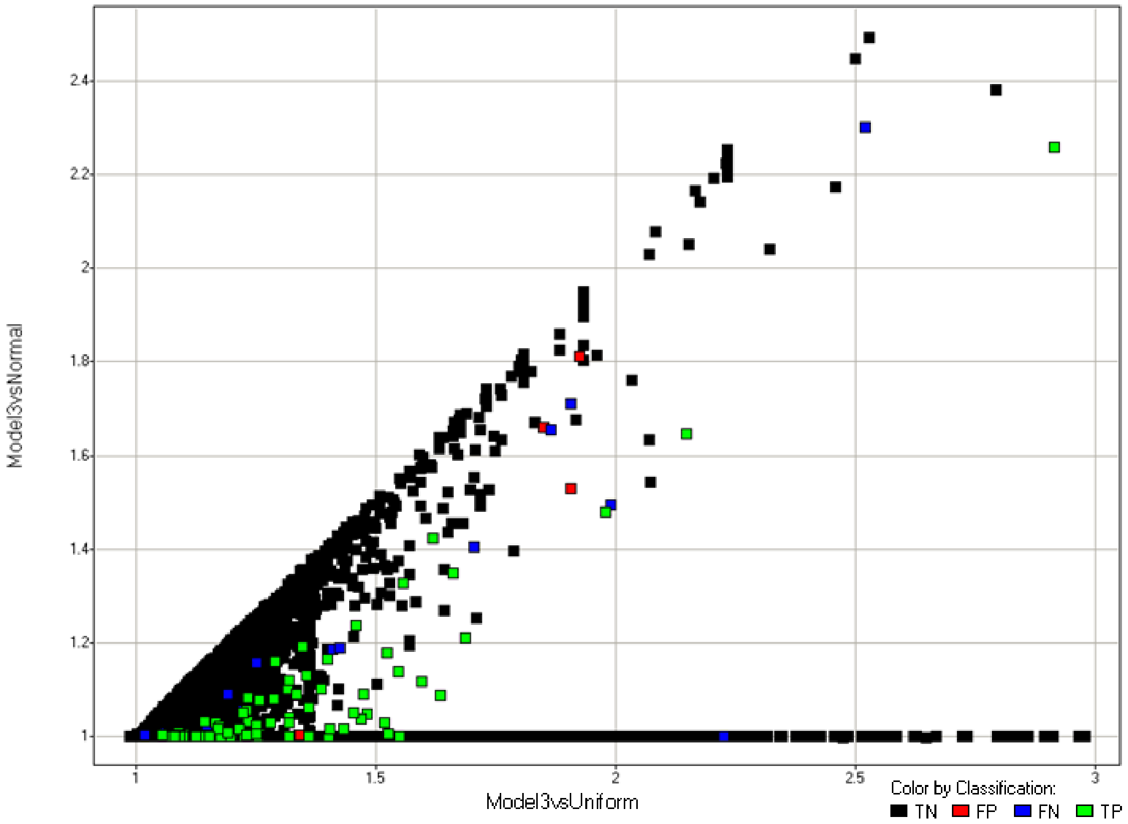


Figure #16: SVM #2 Classification of Candidate Motifs from "Seeded" Random DNA Sequences

For this data set, the three simple models and SVM #2 have correctly identified nearly 71% of the seed motifs with very few false positives, performance far beyond what would have been expected from tools in the current tool set and improbably good given the simplicity of the models.

Results For Benchmark Sequences

To provide a test against sets of known data, the mouse and human "real" sets of sets of promoter regions for co-expressed genes were downloaded from the Tompa benchmark. Given the relatively high degeneracy of the known TFBSs in these sets, the variability in binding site length, and the relatively poor performance of the current set of tools against these sets, performance was expected to be extremely poor. Motifs were predicted for only 13 of the 38 sets. The best two, the only two sets for which any true positives were predicted, were human set #25 and mouse set #8. The scoring results for these two sets are shown in Table #15, Table #16, and Table #17. Tompa *et al* define several additional statistics, as seen in Table #14.

Table #14: Additional Statistics

Metric Name	Definition
Nucleotide-level Performance Coefficient	$\mathrm{nPC} = \frac{\mathrm{nTP}}{\mathrm{nTP} + \mathrm{nFN} + \mathrm{nFP}}$
Nucleotide-level Correlation Coefficient	$\mathrm{nCC} = \frac{\mathrm{nTP}\,\mathrm{nTN} - \mathrm{nFN}\,\mathrm{nFP}}{\sqrt{(\mathrm{nTP}+\mathrm{nFN})(\mathrm{nTN}+\mathrm{nFP})(\mathrm{nTP}+\mathrm{nFP})(\mathrm{nTN}+\mathrm{nFN})}}$
Site-level Average Site Performance	$sASP = \frac{sSn + sPPV}{2}$

Table #15: Best Results from the Tompa Benchmark

Data set	nTP	nFP	nFN	nTN	sTP	sFP	sFN
hm25r	2	12	68	918	0	2	5
mus08r	2	128	39	4331	0	93	3

Table #16: Information Retrieval Metrics for Best Results from the Tompa Benchmark

Data set	nSn	nPPV	nSp	nPC	nCC	sSn	sPPV	sASP
hm25r	0.0286	0.1429	0.9871	0.0244	0.0340	0	0	0
mus08r	0.0488	0.0154	0.9713	0.0118	0.0114	0	0	0

As expected, performance against the benchmark sets was extremely poor. Comparing the sets of predicted motifs against the known TFBSs, there appears to be very little similarity. Figure #17 and Figure #18 show the locations of the predicted motifs and the known sites for the "hm25r" set and the "mus08r" set respectively.

Table #17: Predicted Motifs and Known TFBSs for Best Results from the Tompa Benchmark

Data set	Predicted Motif	Known TFBSs
hm25r	ACTGCTG	ATTACACCAAGTACC GGAATTTCCTGTTGATCC ACCTAAGCTG CTAAAGGACGTCACATTGC ATATAGGA
mus08r	AAGGAAG AGAAGAG CACCACT CTCTCTC GATTAGG TCTCTCT	AGGGGGATTTTCCCT CTGGGGACTCTCCCT GGGGGCTTTCC

Observe in Table #17 that more motifs were predicted for the mouse data set than for the human data set. As Table #18 indicates, this was generally the case regardless of the number of sequences in the set being considered. This may be due to inherent differences in the gene sets considered or between mice and humans or may simply be a result of the choice of training data. Though the reason is not clear, the difference was determined to be significant at a confidence level of p=0.018 using a one-tailed t test.

Table #18: Number of Predicted Motifs and Number of Sequences Per Set

Set Name	Sequence Count	Number of Predictions
hm01r	18	5
hm02r	9	0
hm03r	10	0
hm04r	13	0
hm05r	15	0
hm06r	9	0
hm07r	5	0
hm08r	15	0
hm09r	10	5
hm10r	6	0
hm11r	8	0
hm12r	2	0
hm13r	6	0
hm14r	2	0
hm15r	4	0
hm16r	7	5
hm17r	11	4
hm18r	5	0
hm19r	5	0
hm20r	35	0
hm21r	5	0
hm22r	6	0
hm23r	4	0
hm24r	8	0
hm25r	2	6
hm26r	9	0

Set Name	Sequence Count	Number of Predictions
mus01r	3	0
mus02r	9	0
mus03r	5	0
mus04r	7	1
mus05r	4	1
mus06r	3	0
mus07r	4	8
mus08r	3	5
mus09r	2	9
mus10r	13	4
mus11r	12	6
mus12r	3	14

Looking at the locations of the predicted TFBSs, in green, relative to the known TFBSs, in blue, there is no clear relationship between the positional occurrences of the predicted motifs and locations of the known TFBSs.

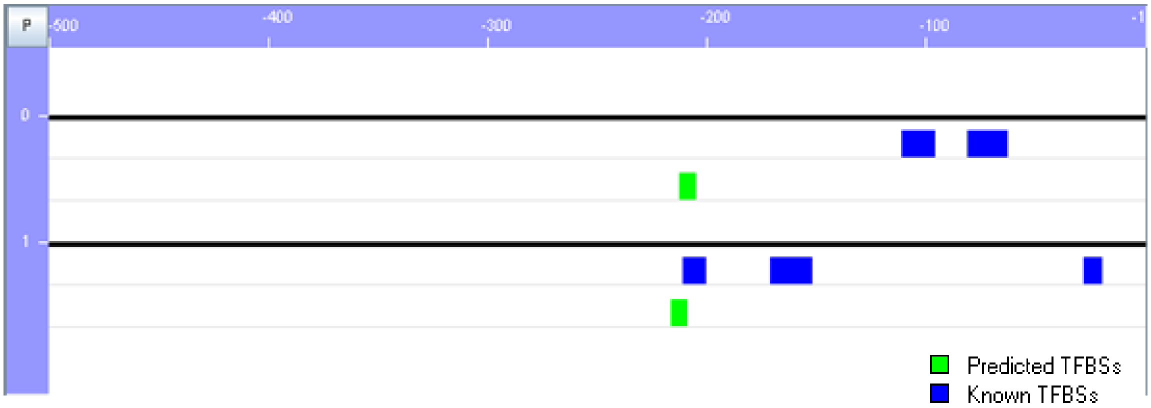


Figure #17: Visualization of Predicted and Known TFBSs for hm25r

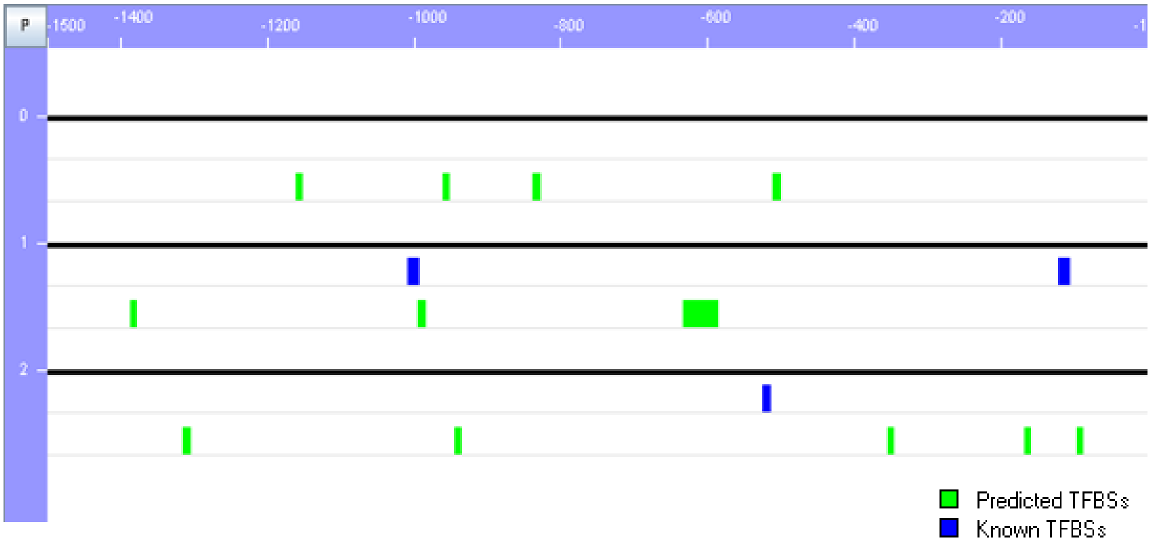


Figure #18: Visualization of Predicted and Known TFBSs for mus08r

Conclusions

Three extremely simple mathematical models incorporating frequency and positional and sequence specificity were applied to the problem of *de novo* TFBS discovery for simulated and real promoter regions of co-expressed genes. For simulated sets of co-expressed genes, the resulting analyses showed promise by performing extremely well on a training set and, for a confirmation set, correctly identifying several seed motifs without producing any false positives. For sets of genes determined experimentally to be preferentially co-expressed in mouse lymphocytes at different stages of thymocyte maturation, simple model selection and subsequent filtering resulted in a set of predicted TFBS motifs for each of the stages considered. The motifs predicted by this approach included motifs highly similar to a small number of known TFBS motifs. For three of the four stages of maturation investigated, the predicted motifs correctly matched with the TFBS motifs known to be associated with the stage of development; for the fourth stage, neither confirmation nor refutation was uncovered in a search of the literature. Based on these results, we conclude that even such simple models are informative and provide information that can be applied to improve the predictions of current *de novo* TFBS discovery tools.

The performance of the models was extremely poor when used to predict known TFBS motifs in sets from the Tompa benchmark, specifically sets of promoter regions for mouse and human genes determined experimentally to be co-expressed and regulated via TF interactions at occurrences of exactly one known TFBS motif per set of genes. Comparing the known binding sites to the predicted motifs, the poor performance of the TFBS motif discovery approach employed is seen to largely be due to the simplicity of the models investigated and, more specifically and most significantly, the inability of the chosen motif representation to capture the variable-length and degenerate nature of the motifs for the known TFBSs. Accurate prediction is further complicated by the possibility that there are unknown but biologically relevant TFBS motifs in the benchmark sets analyzed. If these unknown motifs are discovered and unknown sites are reported by a tool being scored, the tool would be incorrectly penalized for reporting "false positives" since the results would not match the known motifs.

These results imply that machine learning approaches such as supervised learning and model induction may yet yield better objective functions for TFBS discovery than are currently in common use. Though models as simple as those investigated here can be informative, far superior results could be expected for more complex models that are capable of representing more biologically relevant information, such as the models underlying the best of the TFBS discovery tools currently in general use. Significant improvements in accuracy are to be expected if machine learning methods are applied to statistics generated by the best of the currently available tools or models mimetic of their objective functions.

Discussion and Recommendations

In many areas of research, a problem will often have several common approaches if the problem is sufficiently trivial to allow many paths to a solution or, alternately, if the problem is difficult and no clear "best approach" has been identified. *De novo* TFBS discovery currently represents an example of the latter. From 1995 to present, there have been several waves of new TFBS prediction tools resulting in the solution space being quite saturated. Despite the many efforts, the overall performance of the tools in this space remains poor. This implies that the correct information is not available or is simply not being incorporated in the prediction process. Recently, positional specificity, in the form of the Kolmogorov-Smirnov (K-S) statistic, has been shown to provide a valuable additional dimension for use in separating TFBSs from background (Kim, Tharakaraman and Mariño-Ramírez). In the work presented here, the likelihoods of the various models serve a very similar function. Though not demonstrated here, there would be expected to be a strong correlation between the K-S statistic, calculated with the uniform distribution as the null model, and the log-likelihood score obtained for the uniform model. This was likely a contributing factor in the positive results obtained for the mouse thymocycte data. Future methods that incorporate other types of information might demonstrate further improvements in performance. For example, considering multiple TSSs per gene and distinguishing among upstream, downstream, intronic, and extronic regions might be beneficial. Semi-empirical methods that incorporate high-throughput amino acid/DNA affinity data may provide another boost to prediction accuracy and enable prediction of the structure of novel DNA binding domains to predict unknown TFs for putative TFBS motifs.

In addition to needing to assess how informative additional dimensions may be, the possibility must be considered that one or more of the assumptions currently being made by most tools is not valid. The "hard" assumption of independence among positions within a motif has received some scrutiny and is one assumption that merits further study. Even considering only the possibility that coupling between adjacent positions in a motif might be biologically relevant, a more general representation for motifs than the ubiquitous PWM would be necessitated. Informed by experimental evidence, options to search for direct and mirrored repeats have been included in several tools, and work on tools to detect cis-regulatory modules is on-going. Both of these efforts would be well served by a comprehensive approach to detecting more complex motifs.

Though the different approaches to TFBS discovery have each reported successful prediction of known and novel TFBS motifs, the optimality of predictions is guaranteed only for combinatorial approaches. With quantum computing on the horizon and as the Moore's Law trends in computational power and storage capacity continue, the resource constraints that often necessitate greedy and probabilistic approaches can be expected to relax and allow combinatorial approaches to be applied more widely. Though these trends are a boon to computational biology, we obviously cannot trust to faster hardware and more storage to eliminate the challenge of unraveling transcriptional regulation. Even with no performance or resource constraints, scoring candidate motifs using an objective function that does not include necessary and sufficient information to distinguish biologically relevant motifs from background will produce erroneous predictions unacceptably often. Extending proven algorithms to include additional information, such as considering positional specificity in a MDScan-like algorithm to grow motifs from "seeds", exact n-mers found in the set of regulatory sequences, can be expected to provide some improvement and seems a promising direction for future investigation.

Ultimately, the synergy between computational investigation and experimental investigation of transcriptional regulation must be exploited to allow better characterization of an ever-larger set of known TFBSs so that either the fully biological mechanism or an optimal set of features for prediction can be identified. Only through the repetitive cycle of hypothesis generation and confirmation using both computational and experimental approaches will we gain a more complete understanding of the underlying mechanics of transcriptional regulation. Though largely neglected here, the experimental front has not been silent while efforts on the computational front have continued. Experimental approaches, such as genome-wide screening for binding sites for known transcription factors using ChIP-on-Chip or ChIP-Seq, are powerful tools that have become widely used. Though each method, computational or experimental, has limitations, each method has a useful niche to fill. ChIP-on-Chip and ChIP-Seq, for example, are invaluable tools when a transcription factor is known. Motif finders, such as discussed in this thesis, are more relevant to cases in which a common transcription factor for a set of co-expressed genes is not known or to cases in which regulatory networks are being predicted for entire genomes.

Three areas for extension of this work are immediately obvious:

- Improving the training data used with the current models and approach
- Comparison of performance on training data with a small set of the best tools
- Improving the set of models

The randomly generated sequences seeded with relatively plentiful exact occurrences of a perfectly conserved motif of a fixed width should have provided a very easy test case, perhaps too easy. For such a simple and non-biological test set, a classifier trained on the test set might classify exclusively on a dimension such as the number of exact matches detected and ignore features that would be relevant for real data. Obviously, the training set must be representative of the data that will actually be encountered in normal use of the tool if the classifier is to learn the distinctions between TFBSs and background. By this argument, the Tompa benchmark would be an excellent training set. By performing automated cycles of training reserving one set from the benchmark as a test set, the optimal SVM parameters could be determined, and a good estimate of the performance of this approach using the current models and features could be obtained.

It is possible that the construction of the randomly generated test sets systematically incorporated unexpected characteristics that would make it difficult to predict the seeded TFBSs. If these sets are to be used for training, it would be beneficial to use at least one of the common tools to predict TFBSs in them as a verification step. A comparison of the performance of this approach to the performance of tools that have been demonstrated to perform well on real data sets should provide a meaningful measure of how difficult the TFBSs are to predict in the test sets, regardless of the type of test set being analyzed.

Improving the set of models is the most urgent direction for future work. One common feature of the three models currently being considered, the reliance on a string representation of the 7-mer candidate motifs, is extremely inflexible and unrealistic. At a minimum, the models need to be extended to represent motifs as PWM. This will require that the models incorporate probability-weighted candidate TFBSs in the model and in the likelihood calculations. Adding additional models, models mimetic of the objective functions of the best current tools, should also be explored.

Based on recently published results, better than a 20% improvement is possible using simple voting schemes based on the binding site predictions of current tools. For the same set of tools, at least one tool in a set of five predicted the correct motif 90% of the time for the data sets analyzed. Based on these results, it seems reasonable that a 20% improvement, and possibly as high as a 300% improvement, in accuracy may be possible by applying machine learning methods to the tool outputs or, equivalently, to statistics generated by models mimetic of the objective functions of the best current tools.

If models mimetic of the objective functions of a few of the most accurate current tools are created, it would be desirable to work with all of the models in a single statistical framework. There is a challenge to be overcome if the current approach were to be carried forward. The current approach is based on a point estimate of the maximum likelihood parameters. As a result, the current approach suffers from a problem similar to most EM algorithms, namely difficulty obtaining z-scores, e-values, p-values, or some other standard statistic and the resulting need for one or more custom statistics to fill this void. To avoid this difficulty, a purely Bayesian approach to obtaining model likelihoods would be preferable and should be pursued.

In addition to the results from fitting the current models to the data, the fit results for new models or the outputs from a diverse set of tools should be included as features for classification. The Tompa benchmark and ensemble approaches such as EMD provide side-by-side performance comparisons for tools that utilize diverse approaches to predict TFBSs and a common set of benchmark data. Though not novel, at least a modest improvement over previous performance could be expected. After bloating the feature set and demonstrating improved performance, the aim would be to trim the feature set, to induce a simple predictive model that demonstrates comparable performance with a reduced feature set.

Particularly given the initial promise shown in using simple maximum-likelihood model selection to obtain meaningful TFBS motif predictions, the performance of the approach against the Tompa benchmark was, if not unexpected, all the more disappointing. Though the poor performance was almost exclusively due to the inflexibility of the models used, improving the models cannot be expected to fully ameliorate the problem of poor performance; Tompa *et al* discuss the failings of the benchmark and how it should be changed to be more fair and reflective of the relative performance of the tools being assessed. Even with the authors' suggestions for future benchmarks, each of the subcategories of benchmark sets will present a different set of difficulties. Tompa *et al* acknowledge that the "real" data sets may contain unknown TFBSs and that the other sets in the benchmark are potentially flawed in different ways. The "Markov" sets, for example, were generated using a Markov chain of order three, and, as a result, it may be very easy for some tools to find such motifs because the tools may be based on an approach that, intentionally or unintentionally, exploits this fact. The order zero Markov chain used to generate the training and confirmation sequences in this thesis similarly produce output that likely lacks key characteristics of real data and embeds non-biological characteristics that tools might exploit. It is an unfortunate Catch-22 that good test sets cannot be created without an accurate model of the biology of transcriptional regulation but the goal of the tools is to answer questions raised by the lack of just such a model.

A perfect generative model for sets of regulatory regions for co-regulated genes would necessarily be a complex model that includes both variables that are currently unknown and those that are known to be informative in motif discovery or predictive modeling. Since the true model for TFBS generation *in vivo* is not completely understood, good experimentally generated test sets and test sets derived from generative models of varying complexity will be the training and validation test sets. Whatever approach is taken to predict motifs, it must be robust enough to accommodate the known data and some amount of noise.

It is worth noting that the Tompa benchmark sets presume that 2000bp upstream of TSS contains all relevant regulatory regions. In contradiction to this assumption, regulatory regions have been identified in exonic and intronic regions as well as downstream of the gene of interest. The relative dearth of sets of fully characterized co-regulated genes for which transcriptional regulation is well understood poses a significant but not insurmountable challenge to progress.

Works Cited

Ao, Wanyuan, *et al*. "Environmentally Induced Foregut Remodeling by PHA-4/FoxA and DAF-12/NHR." Science (2004): 1743-1746.

Bailey, Timothy L and Charles Elkan. "The value of prior knowledge in discovering motifs with MEME." ISMB-95 Proceedings (1995): 21-29.

Ben-Gal, I *et al*. "Identification of transcription factor binding sites with variable-order Bayesian networks." Bioinformatics (2005): Vol 21, No 11, 2657-2666.

Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA and and the members of the Mouse Genome Database Group. "The Mouse Genome Database (MGD): mouse biology and model systems." Nucleic Acids Research (2007 35 (Database issue)): D724-8.

Carmack, C Steven, *et al*. "PhyloScan: identification of transcription factor binding sites using cross-species evidence." Algorithms for Molecular Biology (2007): 1-17.

Chang, Chih-Chung and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. 21 May 2009. 29 September 2008 <http://140.112.30.28/~cjlin/papers/libsvm.pdf>.

Down, Thomas A and Tim JP. Hubbard. "NestedMICA: sensetive inference of over-represented motifs in nucleic acid sequence." Nucleic Acids Research (2005): 1445-1453.

Eskin, Eleazar and Pavel A Pevzner. "Finding composite regulatory patterns in DNA sequences." Bioinformatics (2002): 354-363.

Favorov, AV *et al*. "A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length." Bioinformatics (2005): Vol 21, 2240-2245.

Frith, M.C. *et al*. "Finding functional sequence elements by multiple local alignment." Nucleic Acids Research (2004): 189-200.

Hertz, G. Z. and Stormo, G. D.Stormo, Gary D. "Identifying DNA and protein patterns with statistically significant alignment of multiple sequences." Bioinformatics (1999): 563-577.

Hu, Jianjun, Yifeng D Yang and Daisuke. Kihara. "EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences." BMC Bioinformatics (2006): 7:342.

Hughes, Jason D, *et al*. "Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." Journal of Molecular Biology (2000): 1205-1214.

Kim, Nak-Kyeong, *et al*. "Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites." BMC Bioinformatics (2008, 9): 262.

Liang, S. *et al*. "cWINNOWER Algorithm for Finding Fuzzy DNA Motifs." Journal of Bioinformatics and Computational Biology (2004): 47-60.

Liu, F.F.M., *et al*. "FMGA: finding motifs by genetic algorithm." Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on (2004): 459-466.

Liu, X Shirley, Douglas L Brutlag and Jun S. Liu. "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nature Biotechnology (2002): Vol 20, 835-839.

Liu, X, DL Brutlag and JS Liu. "Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput (2001): 127-138.

Moses, A. M. *et al*. "Phylogenetic motif detection by expectation-maximization on evolutionary mixtures." Pacific Symposium on Biocomputing (2004): 324-335.

Newberg, Lee A., *et al*. "A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction." BIOINFORMATICS (2007): 1718-1727.

Pavesi, Giulio, Giancarlo Mauri and Graziano. Pesole. "An algorithm for finding signals of unknown length in DNA sequences." Bioinformatics (2001): Vol 17 Suppl 1, S207-S214.

Pevzner, Pavel A and Sing-Hoi Sze. "Combinatorial Approaches to Finding Subtle Signals in DNA Sequences." Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (2000): 269-278.

Prakash, A *et al*. "Motif Discovery in Heterogeneous Sequence Data." Pacific Symposium on Biocomputing (2004): 348-359.

Puthier, Denis *et al*. "A General Survey of Thymocyte Differentiation by Transcriptional Analysis of Knockout Mouse Models." The Journal of Immunology (2004): 6109-6118.

Regnier, Mireille and Alain Denise. "RareEvents and Conditional Events on Random Strings." Discrete Mathematics and Theoretical Computer Science (2004): 191-214.

Shida, Kazuhito. "GibbsST: a Gibbs sampling method for motif discovery with enhanced resistance to local optima." BMC Bioinformatics (2006): 7:486.

Siddharthan, Rahul, Eric D Siggia and Erik van Nimwegen. "PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny." PLOS Computational Biology (2005): 0534-0556.

Sinha, Saurabh and Matrin Tompa. "YMF: A program for discovery of novel transcritpion factor binding sites by statistical overrepresentation." Nucleic Acids Research (2003): 3586-3588.

Sinha, Saurabh, Mathieu Blanchette and Martin Tompa. "PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences." BMC Bioinformatics (2004): 186/1471-2105-5-170.

Thijs, Gert *et al*. "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling." Bioinformatics (2001): 1113-1122.

Tompa, Martin *et al*. "Assessing computational tools for the discovery of transcription factor binding sites." Nature Biotechnology (2005): 137-144.

—. Assessment of Computational Motif Discovery Tools. 26 August 2004. 19 February 2009 <http://bio.cs.washington.edu/assessment/>.

van Helden, J, B Andre and J. and Collado-Vides. "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." Journal of Molecular Biology (1998): 827-842.

van Helden, Jacques, Alma F Rios and Julio. and Collado-Vides. "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads." Nucleic Acids Research (2000): 1808-1818.

Wang, Ting and Stormo, Gary D. "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics (2003): 2369-2380.

Wijaya, E *et al.* "MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders." Bioinformatics (2008): 2288-2295.

Workman, CT and GD. Stormo. "ANN-Spec: A method for discovering transcription factor binding sites with improved specificity." Pacific Symposium on Biocomputing (2000): 5:464-475.

Curriculum Vitae

James E Scherschel
jeschers@iupui.edu

Education

Master of Science in Bioinformatics, Expected May 2009
School of Informatics, Indiana University Purdue University at Indianapolis (IUPUI)
Thesis: De Novo Transcription Factor Binding Site Discovery: A Machine Learning and Model Selection Approach
Advisor: Dr Narayanan B Perumal

Bachelor of Science in Mathematics and Computer Science, May 2000.
Ball State University; Muncie, Indiana USA
Thesis: Modeling heat flow in a thermos
Advisor: Dr Michael A Karls

Bachelor of Science in Chemistry, May 1999.
Ball State University; Muncie, Indiana USA

Experiences

- **Software Engineer**, Eli Lilly and Company, Indianapolis, IN
- **Team member** of Ball State University's high-energy physics research group; presented "Polarization of Neutrons" to the Ohio section of the American Assn of Physics Teachers
- Developed an interactive artificial neural network application which illustrates computer "learning" through playing tic-tac-toe; presented "Playing Tic-Tac-Toe with a Neural Network" at Argonne National Laboratory