IDENTIFYING DNA SEQUENCE MOTIFS OF PDX-1 AND NEUROD1 TRANSCRIPTION FACTORS

_____

A Project

Presented

to the Faculty of

California State University, Chico

_____

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Computer Science

_____

by

Hassan Aldarwish

Fall 2014

IDENTIFYING DNA SEQUENCE MOTIFS OF PDX-1 AND NEUROD1 TRANSCRIPTION FACTORS

A Project

by

Hassan N. Aldarwish

Fall 2014

APPROVED BY THE DEAN OF GRADUATE STUDIES
AND VICE PROVOST FOR RESEARCH:

_____
Eun K. Park, Ph.D.

APPROVED BY THE GRADUATE ADVISORY COMMITTEE:

_____
Melody Stapleton, Ph.D.
Graduate Coordinator

_____
Elena Harris, Ph.D., Chair

_____
Melody Stapleton, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A - Adenine

C - Cytosine

CHIP-SEQ - Chromatin Immunoprecipitation Sequencing

DNA - Deoxyribonucleic acid

EM - Expectation Maximization

FDR - False Discovery Rate

G - Guanine

IC - Information Content

NEUROD- Neuronal differentiation

PWM - Position Weight Matrix

PDX - Pancreatic and duodenal homeobox

T - Thymine

UCSC - University of California, Santa Cruz

K-MER - A sequence of characters of size k

STD - standard deviation

TSS - Transcription Starting Site

ABSTRACT

IDENTIFYING DNA SEQUENCE MOTIFS OF PDX-1 AND

NEUROD1 TRANSCRIPTION FACTORS

by

Hassan N. Aldarwish 2014

Master of Science in Computer Science

California State University, Chico

Fall 2014

Transcription is a biochemical process in which genes are copied to produce proteins. Transcription is initiated by the binding of special proteins called transcription factors, at specific sites in the promoters of genes. Transcription factors binding sites are short patterns of consecutive characters, hereafter called motifs. This paper discusses the design, implementation, and application of a word-based approach to detect motif pairs in a set of co-regulated promoters of a reference genome. More specifically, the approach forms motif pairs using distinct strings of lengths ranging from 6 to 8 characters from the co-regulated prompters. Then it uses the hypergeometric probabilistic model to measure the p-value of each motif pair, which indicates the motif pair's statistical significance. Furthermore, the method clusters the statistically significant motif pairs using the Tanimoto distance to eliminate possible duplicates. Moreover, it uses a phylogenetic conservation analysis, which examines the statistical significance of the motif pairs in several different genomes. Lastly, it uses randomized analysis to control the false

discovery rate (or to limit the number of motif pairs that are found to be significant at random). To demonstrate the biological relevance of the results, the approach measures the information content, which shows the conservation level of the nucleic-characters at each position in a motif pair. Finally, the method investigates the positional bias of the resulting motifs relative to the transcription start sites.

We have applied the approach discussed in this paper to detect motif pairs of pdx-1 and Nureod1 transcription factors, which regulate the production of insulin. We have evaluated a total of 4465 motif pairs, which were formed using distinct strings obtained from the set of co-regulated promoters from the mouse gnome. As a result, we have detected 178 motif pairs that are statistically significant and conserved in the rat or human genomes.

CHAPTER I

INTRODUCTION

Background

Diabetes mellitus (or diabetes) is a disease reported to be the 8th leading cause of death across the world. Nearly 38 million people worldwide have Type I diabetes caused by a dysfunction of beta cells that impairs insulin production. A better understanding of mechanisms related to gene expression in beta cells might help in the development of novel strategies for the effective treatment of diabetes. Gene expression is the process of copying (transcribing) genes from DNA and then using these copies to build (translate) the corresponding proteins that are necessary for normal cell functioning. In order for gene expression to take place, special proteins called transcription factors bind to designated transcription factor binding sites that are short motifs, 6-15 characters in length; this binding preludes the process of gene expression. Two known transcription factors, Pdx-1 and NeuroD1, are shown to regulate gene expression in beta cells. Only recently have gene targets that are regulated by both Pdx-1 and NeuroD1 been identified experimentally [1]. However, the motifs (6-8 characters substrings, to which these factors bind within DNA) for this set of genes have not been found. Here we undertake the task of finding overrepresented motifs for Pdx-1 and NeuroD1 given a set of their gene targets (genes whose gene expression are regulated by Pdx-1 and NeuroD1). The challenge of this project is to identify statistically significant pairs of motifs: one motif of each pair is

for Pdx-1 and the other for NeuroD1. Commonly known motif-finding methods are usually restricted to finding a set of potential candidates, each of which is a single motif.

Statement of the Problem

Given a set of sequences of 2000 characters in length using the alphabet (A, C, G, T), identify statistically-significant pairs of motifs overrepresented in this set compared to a background set of sequences of the same length, where a motif is a sequence of consecutive characters (base pairs or nucleotides) of length 6-8. Each sequence in the given set of sequences includes binding sites of both Pdx-1 and NeuroD1 (as was determined experimentally). More specifically, we intend to use a hypergeometric probabilistic model to identify statistically overrepresented pairs of motifs in the given set of sequences. After pairs of motifs have been found, we will analyze their biological significance using phylogenetic conservation analysis (measuring the statistical significance of overrepresentation of the motifs in other related organisms).

Purpose of the Study

The most important contribution of this project is to expand biological knowledge about gene expression mechanisms in beta cells, the failure of which causes diabetes. Secondly, we will research and review the methods for finding pairs of motifs in a set of co-regulated genes. This will contribute to the organization of knowledge and will make further research on motif-finding more efficient. Finally, we will design a tool

for finding overrepresented pairs of motifs given a set of sequences and background sequences. This tool might be useful for other biologists to facilitate their research.

CHAPTER II

LITERATURE REVIEW

DNA Sequence Motifs

DNA sequence motifs are used to profile genomic sites in DNA. They are characterized as short nucleotide patterns, 6 to 8 characters long, conserved in biologically related genomes. The structure of these patterns ranges from simple consecutive subsequences to more complex such as palindromic. In essence, motifs are considered as a means of deciphering DNA; they unveil regulatory networks and provide insights into the relations between transcription factors and their binding sites [2].

Motif Models

There are several models used to represent DNA motifs, including consensus sequences and position weight matrices, PWM. Consensus sequences depict the type of nucleotides of motifs using symbolic codes. These codes, which are standardized by IUPAC [3], reflect the certainty of the nucleotides occurring at a particular position in a motif. For example, the code (A) refers to Adenine, whereas (Y) stands for Cytosine or Thymine. Consensus sequences are compact and suit enumerative based analysis, where a binary decision is sufficient (either a match or a mismatch). However, in some cases it is desirable to measure how well a genomic site matches a motif; it indicates the activity level of promoters as well as binding affinity of transcription factors [2]. To this end,

PWMs encompass probabilistic measures of nucleotides occurrences in motifs. This model is a matrix consisting of nucleotide types (rows) and their indices of occurrences (columns). In practice, these models are intensively modified and extended so as to enhance their robustness and their ability to express complex motifs. For example, some nucleotides' weights are biased in PWM to compensate their "noisy" abundance in a genome [4].

Motif Analysis

Motif analysis is a complex process that consists of several stages. Various algorithms and methodologies have been developed for motif discovery and analysis. The following describes three stages of motif analysis in DNA.

Stage 1: Preprocessing

Co-regulated genes (genes that perform the same biological function) share motifs in their promoters [5], where a promoter is usually defined as a sequence of 2000 base pairs upstream of the start of a gene. Thus, in order to find motifs, the first step is to identify a set (or a cluster) of co-regulated genes. Here, we mention a couple of widely used methodologies for discovering co-regulated genes. One well-known technique employs ChIP-seq (Chromatin immunoprecipitation sequencing) [6]; biologists identify and select the regions to which specific transcription factors bind. These regions are then used to identify co-regulated target genes: given a region to which a transcription factor binds, determine whether there is a gene within 2000-5000 base pairs from this region. Another method uses dynamic chromatin structure, i.e. the changes of nucleosome landscape over time (locations where DNA is wrapped tightly and where it is loose –

gene expression usually is associated with loose DNA). A recent study proposed a relation between binding sites in P. falciparum genome and the evident variance of chromatin availability in malaria-infected erythrocytes [7]. In this study, the regions with the highest variance of loosening of DNA (nucleosome landscape) over time were used to identify clusters of genes with similar behavior, and this was further used to identify motifs. It is worth noting that optimization techniques (such as phylogenetic footprinting) are often used to improve and accelerate the process of identifying co-regulated genes, especially in complex genomes (e.g. human genome) [8]. Once a set of co-regulated genes has been identified, their promoter sequences are extracted from the genome reference for further motif analysis.

Stage 2: Motif Detection

After obtaining a set of sequences that might share the same motifs (from Stage 1), the next step is to identify common motifs overrepresented in the given set of sequences compared to a background. A set of sequences refers to promoter sequences of co-regulated genes, and a background to promoter sequences of all genes in an organism of interest. This is a challenging task that depends on various factors such as the quality of the obtained sequences or quality of conservation of a motif, for example. Moreover, this task has some unknown parameters such as the size or orientation of the motifs. Following is a description of two widely adopted strategies to address these challenges.

The word-based analysis approach searches for overrepresented patterns by exhaustively inspecting all possibilities. Accordingly, globally optimal motifs are likely to be detected. Thorough enumeration, however, demands high computational resources, thus leading to spurious motifs that might be overrepresented by chance [8].

Another approach uses stochastic analysis. In this approach sample motifs are randomly selected and evaluated based on probabilistic assumptions. Then, probabilistic parameters are updated and the process is repeated. It is expected that the random sampling becomes more accurate after each iteration and ultimately converges to the motifs. Statistical analysis is known for its ability to detect lightly conserved motifs. However, it is very sensitive towards local maxima signals. Examples of sampling procedures include EM (Expectation-maximization) and Gibbs sampling [8].

Stage 3: Postprocessing

The performance of motif detection algorithms is affected by various factors. Therefore, it is important to evaluate motifs detected by the algorithms. The evaluation of detected patterns, typically, involves clustering and scoring. Clustering improves the significance of similar patterns and filters out spurious motifs [5]. To check that motifs are indeed special recurring patterns in a genome, their statistical distribution can be compared with the distribution of the genomic background. There have been several statistical distributions used as a genomic background, including random distribution and hypergeometric distribution [9]. The statistical significance of motifs is measured by the p-value [9].

CHAPTER III

METHODOLOGY

Methodology Outline

To detect motif pairs, we enumerate all distinct 6-, 7-, 8-mers in the given set of co-regulated promoters in the mouse genome. The k-mers are, then, filtered according to a scoring criterion, which is described in section 3.2. Following, we evaluate every possible pair combination of the remaining k-mers, again, in the mouse genome. The evaluation includes identifying (section 3.5) and scoring each of the motif pairs. Since our approach is enumerative, it is very likely to find duplicate motif pairs. Therefore, we cluster highly similar motif pairs according to the Tanimoto distance (section 3.6). To further assess the significance of motif pairs, we analyze their conservation in two different genomes, namely rat and human. In particular, we score the motif pairs and keep those that pass a specified threshold in either of the genomes. Lastly, to demonstrate the quality of the final results, we measure the information content of the motif pairs and analyze their positional biases in the co-regulated promoters of the mouse genome. The flow of the analysis is shown in Figure1.

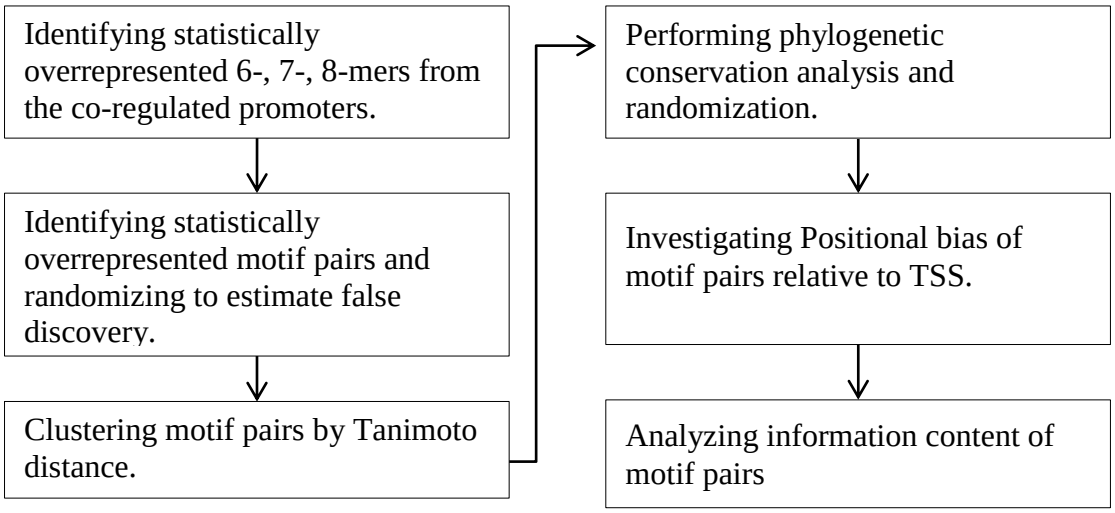


Figure 1. The flow of the various analyses to detect motif pairs.

Motif Pairs Scoring

The primary goal of this analysis is to measure the statistical significance of occurrences of motif pairs in the promoters of the co-regulated set of genes as compared to the background set of promoters of all genes. We assumed that motif pairs' occurrences in promoters follow the hypergeometric probabilistic distribution. In particular, we calculated the probability of a motif pair being observed in the set of the promoters of the co-regulated genes if the set of co-regulated genes had been selected randomly. Specifically, let y be the number of promoters of the set of co-regulated genes G in which a motif pair m occurs and n be the size of G. Further, let r be the number of promoters in the background set S where m occurs and N be the size of S. Then the p-value of m is given by the following formula:

$$P(N,r,n,y) = \sum_{i=y}^{min(n,r)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}$$

The over-representation of a motif pair in the co-regulated promoters is indicated by its p-value, the lower the p-value the higher is the significance of over-representation.

Due to huge search space, we restricted our search to the motif pairs that were formed from single statistically significant motifs. First, we identified single statistically significant motifs using the p-value calculated similarly to the formula above (**y** is the number of the co-regulated promoters, where a motif occurs; and r is the number of promoters of the background set, where the motif occurs) and using the threshold of 0.01. Then we formed all possible motif pairs from the set of statistically significant single motifs, and calculated the p-value for each motif pair. A motif pair was considered statistically significant if its p-value was less than or equal to a threshold that was found using a randomized analysis and that was corresponding to a 5% of false discovery rate, 5% FDR threshold. We considered different combination of lengths for each motif in a pair: (6-mer, 6-mer), (6-mer, 7-mer), (6-mer, 8-mer), (7-mer, 7-mer), (7-mer, 8-mer) and (8-mer, 8mer). For each combination of lengths, we calculated the corresponding 5% FDR threshold for p-value.

False Discoveries Estimation

Here we discuss our randomized analysis to identify 5% FDR threshold. We addressed multiple testing problem by adjusting the p-value threshold according to the

distribution of the p-values calculated for motif pairs occurring at a random set of promoters (corresponding to a random set of co-regulated genes). To estimate the null distribution of such p-values, we computed p-values of the significant motif pairs by repetitively sampling a random set of promoters that is of the same size as the set of co-regulated promoters. Depending on the precision of the simulation, randomized analysis is proven to be efficient in controlling the false discovery rate [10]; thus, we can limit the number of falsely detected motif pairs according to the adjusted threshold. For each combination of lengths of two motifs in a pair, we repeated our analysis 100 times by choosing randomly a set of co-regulated genes of the size of the real set of co-regulated genes used in this research. For each randomly selected set, first, we identified statistically significant overrepresented single motifs (using the p-value threshold of 0.01), and then formed all possible motif pairs formed from these single motifs (a valid motif pair consists of two non-overlapping single motifs that occur in a single promoter), and finally calculated p-value for these motif pairs. We chose the 5th percentile of the p-values of motif pairs from the random sets to be the threshold, which means that no more than 5% of the significant motif pairs found by our method are spurious.

Phylogenetic foot-printing

Here we discuss orthologous analysis of the selected statistically significant motif pairs. Orthologous analysis is a technique that examines conservation of motifs in the promoters of orthologous species. It is assumed that binding preferences of transcription factors are conserved across orthologous species. The rationale behind this is that cis-regularity elements (e.g. promoters) diverge very slowly as they evolve [11].

So, to examine biological significance of the identified statistically significant motif pairs, we calculated their p-value using orthologous promoters of rat and human. We required that the detected motif pairs were statistically significant in at least one orthologous specie with the threshold on p-value corresponding to 5% FDR threshold. The threshold on the p-value was chosen according to a randomized analysis similar to one described to identify 5% FDR threshold to distinguish statistically significant motif pairs.

Motif Pairs Identification

A motif pair consists of two k-mers. An instance of a motif pair is any non-overlapping occurrence of its k-mers in a promoter. To be able to detect conserved motif pairs, we consider non-exact matches for each k-mer of a motif pair. Specifically, we define a mutant M (i, j) of a motif pair MP as any other motif pair, in which the first k-mer has at most i mismatches with the first k-mer of MP (mismatches are identified according to the Hamming distance) and, similarly, the second k-mer has at most j mismatches with the second k-mer of MP. In our analysis, we set both i and j to 1 (one-one-mismatch neighborhood).

The set of mutants that maximizes the enrichment score (or equivalently that minimizes p-value) of a motif pair MP is determined as following. First, we enumerate all possible mutants of MP and calculate their p-values in promoters of mouse. Then, the ten most significant mutants, having p-value less than 0.01, are processed by a heuristic from [7] to choose mutants that maximize the enrichment score of MP. Briefly, the heuristic

uses the dynamic programming to compute p-value of combinations of the 10 lowest p-value mutants and choses the combination that has the lowest P-value.

Motif Pairs Clustering

In In order to weed out highly similar motif pairs, we clustered them. In general, the first requirement to cluster data is the metric that describes the distance between two objects. For this purpose, we used the Tanimoto distance [7], which indicates how well two motif pairs align. For two aligned motif pairs, M1 (a, b) and M2 (a, b), the Tanimoto distance (T) is the ratio of the number of overlapping characters to the size of the alignment. We take the complement of the ratio to immediately interpret the result; the smaller the Tanimoto distance, the higher similarity of two motif pairs.

$$T(M1, M2) = 1 - \frac{(M1.a \cap M2.a) + (M1.b \cap M2.b)}{(M1.a \cup M2.a) + (M1.b \cup M2.b)}$$

The following properties intuitively follow from the definition of Tanimoto distance:

- Non negativity: T(M1,M2)∈[0,1]
- Symmetry: T(M1,M2)=T(M2,M1)
- Identity: T(M1,M1)=0

TABLE 1
An illustration of the Tanimoto distance T between motifs M1 and M2. (a) M1 and M2 are the same motif; they perfectly align, resulting in the minimum distance 0. (b) M1 and M2 partially overlap. (c) M1 and M2 cannot be aligned, resulting in the maximum distance.

	(a) Identical						(b) Partially overlapping									(c) Non-overlapping											
M1	a	c	g	t	c	a	a	c	g	t	c	a				a	c	g	t	c	a						
M2	a	c	g	t	c	a				t	c	a	c	c	g							g	g	t	t	g	g
T	1 – (6/6) = 0						1 – (3/9) = 0.66									1 – (0/12) = 1											

The second ingredient of any clustering experiment is the procedure that divides a set of data objects into groups according to some given criteria. The clustering heuristic we have implemented can be classified as a hierarchical complete linkage. Hierarchical implies that each motif pair belongs to its own cluster at the begging of the process. Complete linkage dictates that the maximum pairwise distance of motif pairs in a cluster is less than the given threshold. The implementation of the clustering heuristic is discussed in section 4.3.

Positional Analysis

To study positional preferences of a motif pair within a promoter relative to a transcription start site, TSS, we simulated the positional distribution of each of the k-mers of a motif in a window of a given size. In particular, for each occurrence of each k-mer of a motif pair inside a single promoter such that both k-mers of the motif pair occur in this promoter, we incremented the count at the starting position of the k-mer inside the promoter and the count of w/2 bases around this position, where w is the size of a window in this research. Initially we subdivided 2000bp promoter into 2000 bins, and initialize the count for each bin to zero. The frequency at each position of a k-mer's window w is obtained from the k-mer's occurrences in a set of promoters P as follows: for each promoter in P, find the occurrences O of the k-mer (consider only those promoters where both k-mers from a motif pair are present). Then, for each occurrence in O, increment the count of all bins centered at the starting position of the occurrence and within distance of w/2 from the starting position. This approach spreads the impact of the

signal (or occurrence) and eliminates sharp cutoffs while preserving the trend of the signal. We used the following sizes for window w (5, 10, 25, 50, and 100) in our analysis. Also, we included the mutants of each k-mer of a motif pair that resulted from Motif Pairs Identification in finding the occurrences.

Information Content

In addition, we carried out position conservation analysis of each motif pair using the information content as the measurement of conservation of bases in a motif pair. The information content of a position i in a motif pair is the sum of the relative entropies of each base (A, C, G, T) at that position, as shown below, where p_i(x)is the probability of base x in position i and p_b(x) is its probability in the background (all promoters) [12]:

$$IC_i = \sum_{x \in \{a,c,g,t\}} p_{i(x)} \times log \frac{p_{i(x)}}{p_{b(x)}}$$

We obtained the base probabilities p_i(x) for each motif pair from all non-overlapping and co-existing (occurring in the same promoter) occurrences of its k-mers in the co-regulated promoters. On the other hand, the background probabilities of each base p_b(x) were calculated based on their frequencies in the set of all promoters. IC reflects the certainty of the content of positions of motif pairs. If the IC of a position is found to be zero, then the probabilities of the characters occurring in the position are no different than their probabilities in the background. As the value of IC increases, the number of expected characters decreases, and thus we become more certain about the content of the position. Since motif pairs are made up of four characters, the theoretical maximum IC is

2; it indicates absolute certainty, meaning only a single character is expected to occur in the position.

CHAPTER IV

IMPLEMENTATION

Programming was a major task in conducting this study. We implemented most of the discussed algorithms and calculated the various attributes of motif pairs using our own tools. The programming was done in c++, using gnu's compiler. The data structures we implemented were an extension of the Standard c++ data structures (e.g. vector). In addition, we used openMP, a thread management library, to parallelize the execution of independent sections of code.

Storing Occurrences of Motif Pairs

This module is designed to store occurrences of k-mers in all promoters of a genome in memory. In order to perform the matching operation rapidly, occurrences of each k-mer were stored in a hash table. The hash function encodes characters of a k-mer as follows (A = 00, C = 01, G = 11, T = 10) and produces the address of a table's entry containing the occurrences of the k-mer. Since we only consider k-mers of size at most 8 characters and each character is 2-bits encoded, the table will have a maximum of $2^{2\times 8}$ entries (approximately 65000 entries).

The hash table is populated by processing the promoters' sequences. As can be seen in Figure 1, a fixed size window is slid over the promoters, encoding all k-mers and inserting their indices into the corresponding entry in the hash table. The time consumed by the encoding and updating operations (lines 10 and 11 in Figure 1.) is constant. Therefore, the hash table is built in linear time in the size and number of

sequences (promoters). Once all promoters are processed and the hash table is filled, the occurrences of any k-mer can be directly obtained by querying the table with its hash value. For example, the hash value of k-mer (acgtacgt) is (0001111000011110), which is the hash table entry containing the k-mer's occurrences. The hash table is collision free and its space complexity depends on the k-mer's size and the number of k-mers in the sequences. The following pseudo code demonstrates the processing of a promoter (Seq) and storing occurrences of all motifs of size (k).

```
1. KmerCounter(Seq, k)
2.   SEQ_LENGTH = 2000 // characters
3.   CHAR_LENGTH = 2   // bits
4.   TABLE_SIZE= 2^(k*CHAR_LENGTH)
5.   Hash_Table[TABLE_SIZE]
6.   Kmer = encode(Seq,0,k)
7.   Hash_table[Kmer].insert(0)
8.   for i = 1 to (SEQ_LENGTH - k):
9.        Kmer_8 = (Kmer<<2) + encode(Seq,i,1)
10.       Hash_table[Kmer].insert(i)
11.  return Hash_table
// The table is built.
// To obtain a motif occurrences:-
Motif_Hash = encode(motif)
Motif_Indices = Hash_Table[Motif_Hash]
```

Figure 2. KmerCounter populates a hash table of k-mers.

Identifying and Scoring Motif Pairs

The primary purpose of this module is to form and identify motif pairs. Since a k-mer might belong to different motif pairs, we designed two data structures motif and mpair. The former data structure maintains the occurrences of a k-mer in memory. The latter, mpair, is a motif pair object, which contains two motif objects' references, statistical attributes, as well as a set of mutants. The statistical attributes of a motif pair

are (1) the number of the co-regulated promoters where the motif pair occurs (cluster_count), (2) the number of promoters in the background set where it occurs (promoters_count), and its p-value (P-value).

TABLE 2
Motif (left) and motif pair (right) data structures.

motif:	motif_pair:
string name	motif motif1
short indices[][]	motif motif2
	double P-value
	int cluster_count
	int promoters_count
	string[] mutants_1
	string[] mutants_2

Since we enumerate all pairs of k-mers to identify motif pairs, the number of motif pairs to be evaluated is $\binom{n}{2}$, where n is the number of k-mers. Moreover, the process of evaluating a motif pair is time consuming; it involves computing of p-value of each of its mutants, sorting the mutants, and executing a heuristic to choose the set of mutants that minimizes the overall p-value of the pair.

We addressed the problem of evaluating large number of motif pairs by distributing the motif pairs among a number of threads, which perform the normal procedure to identify the motif pairs. A noteworthy optimization that we applied was using pre-calculated table of p-values. Note that the computation of p-value of a motif pair requires four inputs (number of promoters, number of co-regulated promoters, promoters_count, and cluster_count). For a given genome where co-regulated promoters have been determined, the first two arguments are constant, promoters_count ranges from zero to the number of all promoters in the genome, and cluster_count ranges from zero to

the number of co-regulated promoters. This means there are only two variables in computing P-value of motif pairs in a genome, namely cluster_count and promoters_count. So, instead of computing p-values during the execution of MotifID, we pre-calculated and stored p-values in a two-dimensional table. As a result, we were able to obtain p-values in constant asymptotic time, simply by referencing the table of p-values.

It should be noted that the table of p-values is specific to a genome, since different genomes might have different number of promoters. The size of the table is the number of promoters multiplied by the number of co-regulated promoters. Typical values of the number of promoters and the number of co-regulated promoters are 39000 and 300, respectively, which result in a table of 11,700,000 p-values. Such table consumes approximately 93.6MB of memory, since the data type of a p-value is double (8 bytes).

Clustering Motif Pairs

Pseudo code for the clustering algorithm is shown in Figure 4. Briefly, cluster refers to a group of motif pairs where the maximum pairwise distance is less than the threshold, compare(cluster C1, cluster C2) is a routine that returns the maximum distance between motif pairs in cluster C1 and motif pairs in cluster C2, and join(cluster C1, cluster C2) is a routine that combines motif pairs from C2 and C1. Lines 1 through 8 preprocess the input set of motif pairs (MP), assigning each motif pair to a cluster, and construct the proximity matrix (P), where P[i][j] is the Tanimoto distance, T, between MP[i] and MP[j]. It should be noted that P is symmetric along its diagonal. Next, starting with the cluster at the first index, the heuristic iterates over all clusters, comparing the ith

cluster with all succeeding clusters and joining those that are compatible. Note that at each iteration, a maximum of one motif pair might change its cluster (be joined with another). In addition, once a motif pair had been joined, it is not re-considered again (line11).

```
1. MotifCL(MP[N], Threshold)
2.    P[N][N]
3.    CLUSTER[N][]
4.    for i = 0 to N:
5.        CLUSTER[i].insert(MP[i])
6.    for j = i+1 to N:
7.            P[i][j] = T(MP[i], M[j])
8.    for i = 0 to N:
9.        if(CLUSTER[i] is not empty):
10.       for j = i+1 to N:
11.           if(CLUSTER[j] is not empty):
12.               distance = compare(CLUSTER[i], CLUSTER[j])
13.           if(distance is less than Threshold):
14.               join(MP[i], MP[j])
15.   return CLUSTERS
```

Figure 3. MotifCL clusters motif pairs using the Tanimoto distance according to the given threshold.

CHAPTER V

RESULTS

A recent study has revealed around 300 genes with binding sites of Pdx-1 and NeuroD1 transcription factors in the mouse genome [1]. Accordingly, we applied our strategies to detect motif pairs for this set of the co-regulated promoters. Throughout our study, we referred to UCSC Genome Bioinformatics website to obtain several data sets, including promoters of mouse, promoters of rat, and promoters of human [13].

TABLE 3
Sizes of promoters in the used genomes.

Genome	Total number of promoters	Number of co-regulate promoters
Mouse	21,204	277
Rat	14,205	201
Human	39,275	214

Evaluation of Motif Pairs

We started the analysis by extracting all distinct k-mers of sizes 6, 7, and 8 characters from the co-regulated promoters in the mouse genome. After that, we measured their distribution in all promoters of mouse in order to calculate their hypergeometric probability. The intent was to reduce the size of the input to the next stage by selecting only statistically significant overrepresented k-mers. So, discarding all k-mers of p-value greater than (0.01), we have retained a total of 30 6-mers, 25 7-mers, and 40 8-mers, which were used to identify motif pairs.

Next, we evaluated the distribution of each combination of two significant k-mers in the promoters. The evaluation involved computing the enrichment score and

selecting mutants of motif pairs. The ten mutants with the lowest p-values of each motif pair were processed by the dynamic programming heuristic to choose those that maximize the enrichment score (or minimize p-value). In addition to the standard p-value threshold such as 0.01, we required that a motif pair must occur in at least five promoters, and used 5% FDR threshold. Table 2 depicts the total number of motif pairs (column 2), together with the number of significant motif pairs (column 3).

TABLE 4
The number of identified motif pairs, which are classified based on the size of their k-mers into: 6-6, 6-7, 6-8, 7-7, 7-8, and 8-8. The total number of motif pairs at each stage, which is the summation of each column, is shown in the bottom row.

Pairs Size	Total Motif Pairs	Significant Motif pairs	Clustered	Significant In Human Or Rat
6-6	435	389	179	39
6-7	750	637	217	26
6-8	1200	927	286	30
7-7	300	269	121	26
7-8	1000	872	387	31
8-8	780	645	366	26
Total	4465	3739	1556	178

Carrying out the analysis, we further investigated biological evidence for the significant motif pairs. First, motif pairs of Tanimoto distance less than 0.5 were filtered out. The resulting number of clusters is shown in Table 1 (column 4). After that, we examined the statistical distribution of the most significant motif pair in each cluster in the mouse and rat genomes and kept only those motif pairs that had the p-value passed the 5% FDR threshold obtained in randomized analysis and that had the p-values less than 0.01 in either genome (column 5).

Discussion

To summarize, we have evaluated a total of 4465 motif pairs, which were formed using all significant distinct k-mers in the co-regulated promoters of mouse. As a result, we have detected 178 motif pairs that are statically significant in at least two of three different genomes. Figure 3 shows the maximum and minimum amount of each class of the detected motif pairs in the co-regulated promoters of mouse. Note that the number of detected motifs decreases as the size of the motif increases. The p-values of the detected motif pairs range from 1.89e-15 to 1.78e-4, as observed in the mouse genome. Tables with various characteristics of the detected motif pairs are shown in appendices A through F.

A summary of the null p-values distribution obtained in randomized analysis in mouse, rat and human genomes is provided in Tables 5, 6, and 7. Note that for each class of motif pairs the 5th percentiles are more lenient than the standard p-value (less than 0.01), suggesting that the number of false discoveries doesn't exceed 5%.

TABLE 5
Null p-value distribution of motif pairs in promoters of mouse.

Pair Size	6-6	6-7	6-8	7-7	7-8	8-8
Mean	0.5600	0.5187	0.5639	0.5714	0.6005	0.6091
Std.	0.2912	0.2823	0.2961	0.2878	0.3051	0.3156
10^{th} centile	0.1154	0.1204	0.1284	0.1514	0.1490	0.1500
5^{th} centile	0.0581	0.0637	0.0671	0.0811	0.0819	0.0780

TABLE 6
Null p-value distribution of motif pairs in promoters of rat.

Pairs Size	6-6	6-7	6-8	7-7	7-8	8-8
Mean	0.5140	0.5386	0.5762	0.5529	0.6185	0.7211
Std.	0.2890	0.2902	0.2960	0.2966	0.3125	0.3344

Table 6 (Continued)

	6-6	6-7	6-8	7-7	7-8	8-8
10^{th} centile	0.1062	0.1196	0.1449	0.1292	0.1571	0.1881
5^{th} centile	0.0529	0.0612	0.0772	0.0659	0.0829	0.1039

TABLE 7
Null p-value distribution of motif pairs in promoters of human.

Pairs Size	6-6	6-7	6-8	7-7	7-8	8-8
Mean	0.5171	0.5385	0.5637	0.5321	0.6012	0.7025
Std.	0.2837	0.2917	0.2957	0.2926	0.3082	0.3321
10^{th} centile	0.1211	0.1229	0.1321	0.1211	0.1525	0.1825
5^{th} centile	0.0646	0.0639	0.0705	0.0581	0.0755	0.0967

To measure the conservation level of the detected motifs, we calculated the probability of the genomic bases (A, C, G, and T) based on the motif pairs' instances in the co-regulated promoters of mouse. Further, we calculated the probability of each of the genomic bases in all promoters, and calculated the Information Content for the positions as described in section (3.2). The basic statistics calculated for the Information Content for different size motif pairs are shown in Table 8. The maximum possible values of Information Content for a motif pair consisting of two motifs of size x and y respectively are calculated as $2(x + y)$ and shown in the last row of Table 8.

TABLE 8
Distribution of the information content of the motif pairs in promoters of mouse.

Pairs Size	6-6	6-7	6-8	7-7	7-8	8-8
Mean	20.2759	22.2975	22.4770	24.6801	24.1030	24.3801
Std.	1.2476	1.1231	1.4684	1.5208	1.2863	1.7470
Maximum	22.6133	24.7762	25.9939	27.3523	26.5545	27.7737

Table 8 (Continued)

	6-6	6-7	6-8	7-7	7-8	8-8
Minimum	17.9329	20.4624	19.7696	22.2074	21.2759	21.5338
Maximum Possible	24	26	28	28	30	32

Table 9 illustrates the conservation level of the motif pairs using sequence logos. The first column of the table shows the most conserved motif pairs in each class, whereas the second column shows the least conserved.

TABLE 9
Sequence logos of some of the detected motif pairs. Seq2Logo2.0 [14].

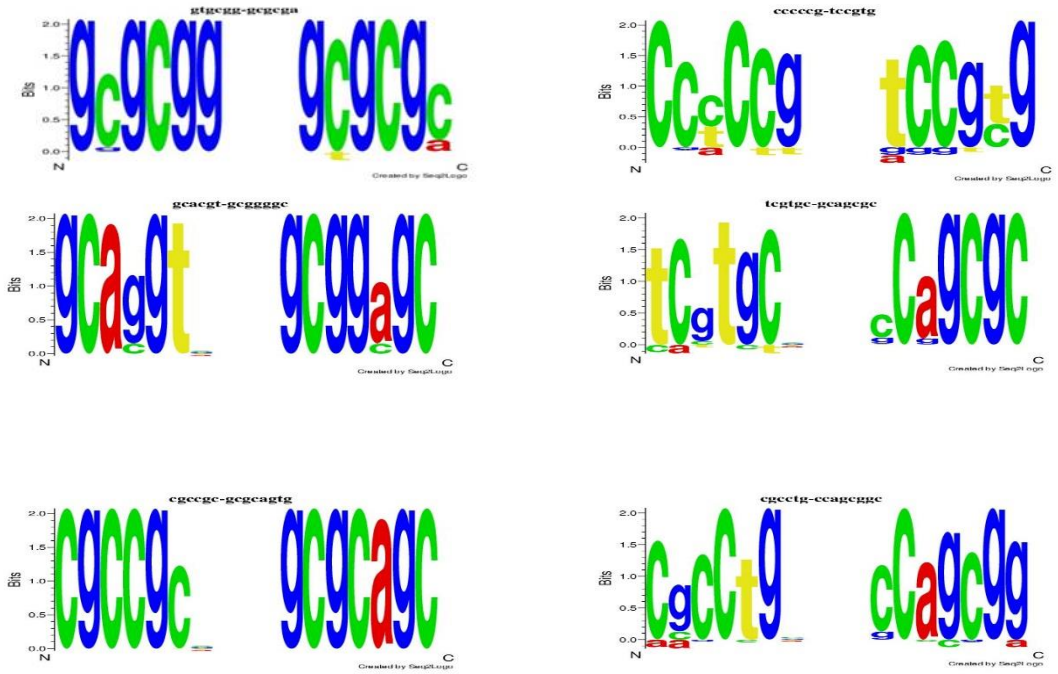
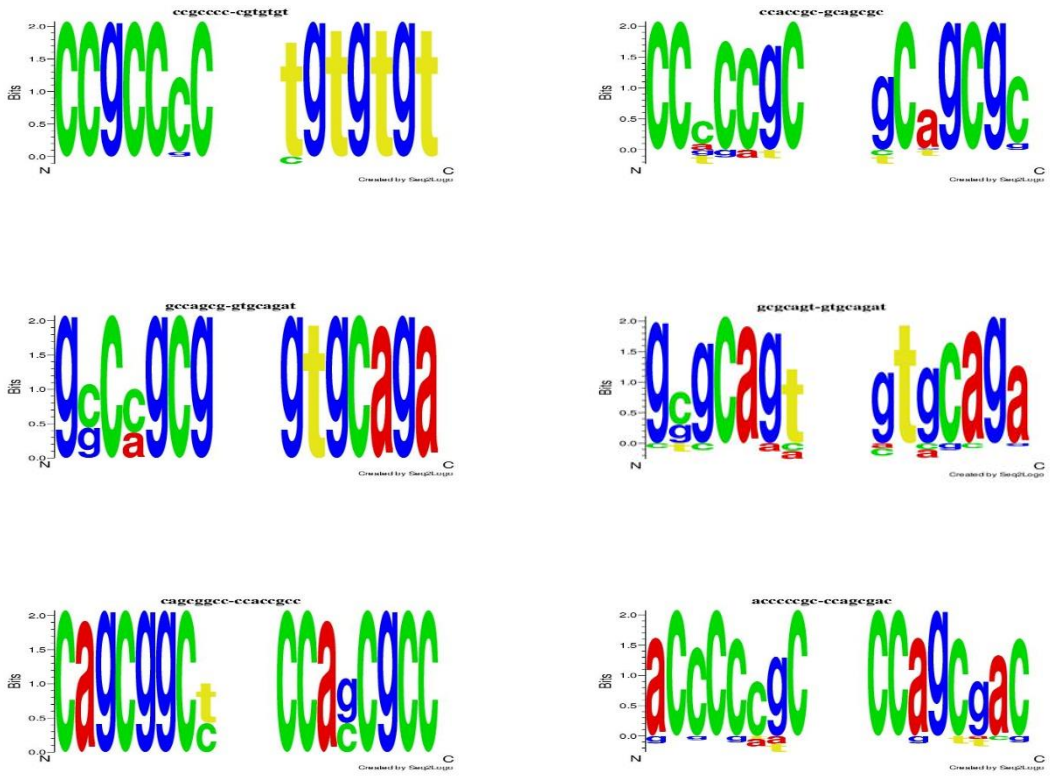


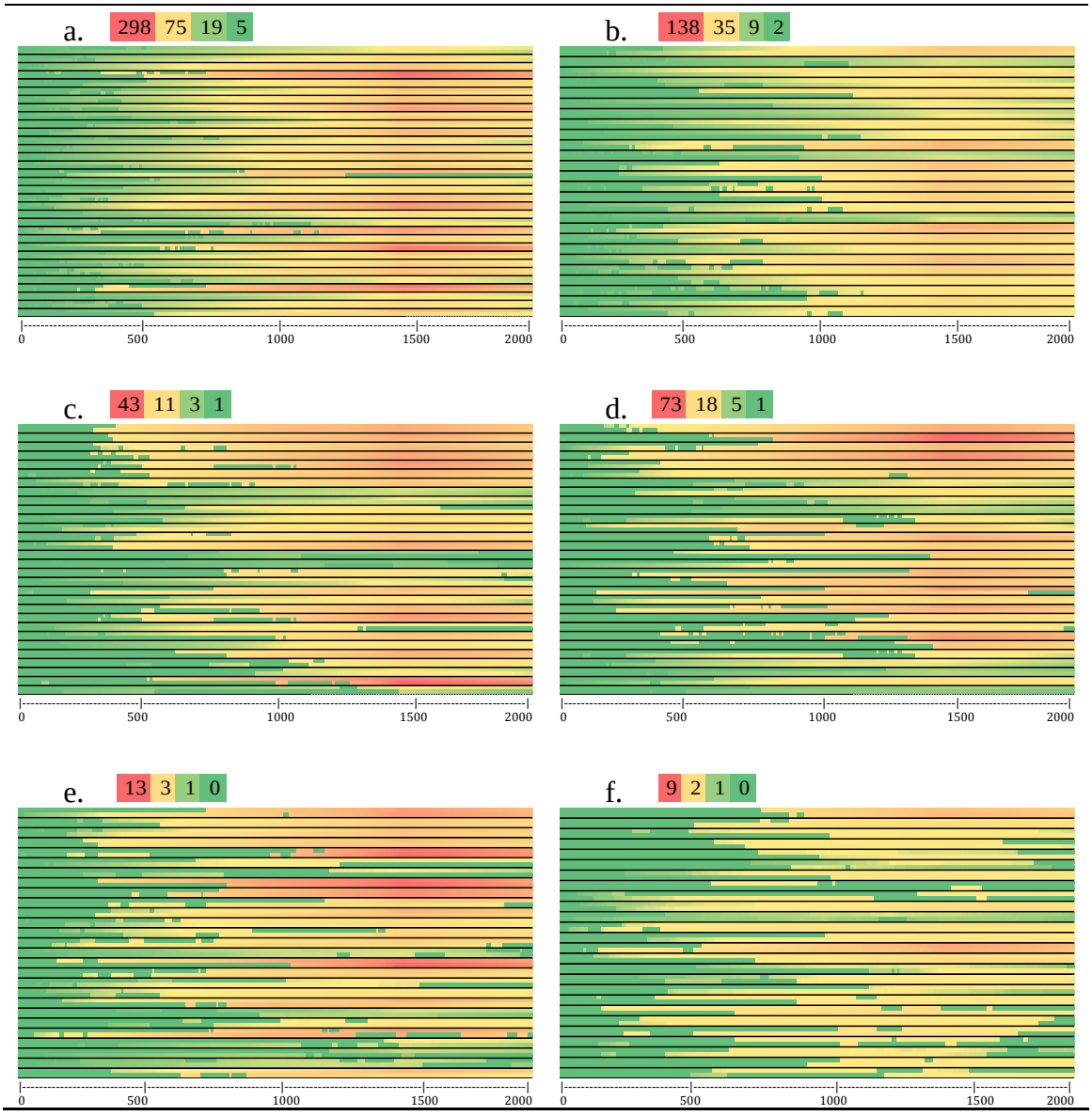
Table 9 (Continued)



The results of the positional analysis in promoters of mouse are shown in Table 10. The analysis includes all identified motif pairs and the window width in the displayed maps was set to 50. As can be seen, all classes of motif pairs tend to occur within the same region, which is the upper half of the promoters, and they peak at the center of the region (approximately at position 1500 corresponding to -500 base pairs upstream of the transcription start site).

TABLE 10

Heat maps of each class of motif pairs: (a) 6-6 (b) 6-7 (c) 6-8 (d) 7-7 (e) 7-8 (f) 8-8. Motif pairs are separated by the horizontal dark lines. The index of promoter is depicted in the scales beneath the maps. Color encoding is shown above each map. Colors correspond to the observed frequency of occurrences, which is also shown inside the color scale.



CHAPTER VI

CONCLUSIONS

Conclusion

Using the results from [1], in particular, using the set of co-regulated genes for Pdx1 and NeuroD1 transcription factors, we identified a total of 178 motif pairs that are statically significant in the mouse genome (P-value ranges from 1.89e-15 to 1.78e-4) and conserved in either of the rat or the human genomes. Though we considered mutated instance (or non-exact matches) of motif pairs in the evaluation, the information content of the detected motif pairs indicate a high level of specificity; most of the positions of motif pairs are well conserved. Also, the detected motif pairs have a strong positional bias inside the co-regulated promoters of mouse; they are likely to be found at around -500 base pairs upstream of TSS in the co-regulated promoters.

Future Research

In this project we have established a word-based approach to detect motif pairs in a set of co-regulated promoters. Despite the demand of computational resources imposed by the problem, we have implemented a fully functional program with no exceptional execution requirements. In addition, the program uses standard input/output formatted files, meaning it can be easily used to detect motif pairs of other transcription factors. This project focused on the computational aspects of the motif detection problem. A logical next step would be to affirm the correctness of the approach, which can be done by validating the detected motif pairs using biological experiments.

REFERENCES

REFERENCES

[1] Keller, D. M., S. Mcweeney, A. Arsenlis, J. Drouin, C. V. E. Wright, H. Wang, C. B. Wollheim, P. White, K. H. Kaestner, and R. H. Goodman. "Characterization of Pancreatic Transcription Factor Pdx-1 Binding Sites Using Promoter Microarray and Serial Analysis of Chromatin Occupancy." Journal of Biological Chemistry 282.44 (2007): 32084-2092. Web.

[2] D'haeseleer, Patrik. "What Are DNA Sequence Motifs?" Nature Biotechnology24.4 (2006): 423-25. Web.

[3] Stothard P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104.

[4] Stormo, G. D. "DNA Binding Sites: Representation and Discovery." Bioinformatics 16.1 (2000): 16-23. Web.

[5] Macisaac, Kenzie D., and Ernest Fraenkel. "Practical Strategies for Discovering Regulatory DNA Sequence Motifs." PLoS Computational Biology 2.4 (2006): E36. Web.

[6] Bailey, Timothy, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. "Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data." Ed. Fran Lewitter. PLoS Computational Biology 9.11 (2013): E1003326. Web.

[7] Harris, Elena Y., Nadia Ponts, Karine G Le Roch, and Stefano Lonardi. "Chromatin driven De Novo Discovery of DNA Binding Motifs in the Human Malaria Parasite." BMC Genomics 12.1 (2011): 601. Web.

[8] Das, Modan K., and Ho-Kwok Dai. "A Survey of DNA Motif Finding Algorithms." BMC Bioinformatics 8.Suppl 7 (2007): S21. Web.

[9] Barash Y, Bejerano G, Friedman N (2001) "A simple hyper-geometric approach for discovering putative transcription factor binding sites". Algorithms in bioinformatics: First International Workshop, WABI 2001, Aarhus, Denmark, August 28–31, 2001. Berlin: Springer. pp. 278–293.

[10] Westfall, Peter H., and S. Stanley Young. Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. New York: Wiley, 1993. Print.

[11] Zhang, Zhaolei, and Mark Gerstein. "Of Mice and Men: Phylogenetic Footprinting Aids the Discovery of Regulatory Elements." Journal of Biology 11th ser. 2.2 (2003): n. pag. Web.

[12] Durbin, Richard. Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK: Cambridge UP, 1998. Print.

[13] UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

[14] Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Martin Christen Frolund Thomsen; Morten Nielsen, Nucleic Acids Research 2012; 40 (W1): W281-W287.

APPENDIX A

List of Tables

TABLE A-1
Attributes of the detected motif pairs of class 6-6 as observed in the mouse genome

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
cgtgca	tccggc	18.9134	1.94e-14	5272	127
cagcgc	gcacgt	19.7123	3.39e-12	4549	110
ccgccc	tccggc	19.0762	5.05e-11	6184	132
gtgcgg	gggcgc	22.4372	4.96e-08	4150	92
aggtac	tcgtgc	20.2541	1.78e-13	3758	100
cccccg	tgagcg	19.7906	3.02e-13	6390	141
ggggcg	tgggcg	19.6199	2.89e-08	9783	173
cccccg	tccgtg	17.9329	1.88e-13	6152	138
gcgcga	tgagcg	20.5267	1.12e-09	4726	106
cgcctg	gggcgc	20.5556	3.62e-09	6493	131
gccgcc	gcgcga	21.6689	2.92e-08	6419	127
aggtac	gcgggc	20.6304	9.12e-12	5061	117
cgcctg	gcgcga	21.0508	4.50e-09	5696	119
ggggcg	tcgtgc	19.0406	1.89e-15	4563	118
gcgggc	gggcgg	19.6695	4.11e-09	6853	136
cccccg	gcacgt	20.1346	1.42e-10	5206	116
cagcgc	tccggc	19.1461	1.01e-14	6480	146
cgcctg	gcgggc	20.1327	8.31e-11	6499	136
cgcctg	tgggcg	20.9334	3.49e-08	7639	144
gggcgg	tgagcg	19.1706	8.58e-11	7412	149
cccccg	gcgggc	21.2858	5.78e-12	5480	124
tgggcg	gggcgc	19.6758	1.10e-09	6590	134
cccccg	gggcgc	21.1308	7.86e-11	7908	156
gccgcc	gcgggc	18.6975	4.79e-12	10396	192
cagcgc	gcgggc	21.9295	4.47e-10	5562	120

Table A-1 (Continued)

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
gcgggc	gcgcga	22.208	5.33e-08	4093	91
gcgggc	cgtgca	18.8854	4.87e-10	4724	107
ggggcg	gcgggc	20.7011	1.06e-08	8686	160
gtgcgg	gcgcga	22.6133	2.57e-08	3661	85
cccgcc	gcgcga	21.6958	9.73e-08	6542	127
gcgcga	tccgtg	19.6936	5.17e-10	5508	119
cagcgc	tgggcg	22.2352	2.49e-09	6875	137
cagcgc	gggcgc	22.1311	1.62e-08	8360	155
ggccct	gcacgt	19.3683	4.63e-10	4912	110
gggcgg	tgggcg	18.0595	5.24e-14	10079	193
ggtgcg	ggggcg	21.24	8.33e-08	5035	105
ggtgcg	gcgggc	20.2236	2.45e-09	7226	142
cgccgc	tgggcg	19.5815	2.59e-10	6876	140
cccccg	tgggcg	19.0092	6.18e-14	8323	170

Table A-2
Attributes of the detected motif pairs of class 6-7 as observed in the mouse genome

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
gcgcga	ccgcccc	22.7634	2.59E-14	3763	102
ggtgcg	gcgggct	21.5743	4.35E-13	1531	57
tccgtg	gcgggggc	20.7786	1.33E-11	4211	103
tcgtgc	gcagcgc	20.4624	2.49E-11	2517	73
cgcctg	gcagcgc	21.0714	2.60E-10	3628	90
gtgcgg	cgcctgc	23.0983	2.61E-10	2267	66
tcgtgc	gcgggggc	21.6051	3.16E-10	1808	57
gtcgct	ccgccgc	21.729	7.86E-10	3705	90
cggctc	gcagcgc	22.1276	1.49E-09	2255	64
ggggcg	ccgcccc	21.4947	1.72E-09	5346	115
gcacgt	ccaccgc	22.2001	2.25E-09	1751	54
ccgccc	ccccgcc	21.8489	7.13E-09	4622	102
ccgccc	cgcctgc	23.9457	1.49E-08	2560	67
gggcgg	ccccgcc	23.7913	1.77E-08	4253	95
cagcgc	ccccgcc	23.462	8.79E-08	3947	88

Table A-2 (Continued)

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
gcgggc	gcgcagt	23.1004	1.68E-07	3261	76
gcacgt	gcgggggc	24.7762	2.66E-07	1640	47
gcgcga	ccccgcc	23.509	4.03E-07	4726	98
ggtgcg	cgcctgc	21.35	2.36E-12	3916	100
tccgtg	cgcctgc	20.9773	2.49E-12	3051	85
tgggcg	cgcctgc	21.7811	9.86E-11	4228	101
ggtgcg	gctcgcc	21.6678	2.28E-10	4165	99
tccgtg	ccccgcc	21.2049	2.35E-10	3681	91
cagcgc	cgcctgc	23.3848	3.63E-08	2448	64
ggtgcg	gctctcc	22.9326	7.15E-08	3497	81
gcgggc	gcgcagt	23.1004	1.68 E-08	3261	76

TABLE A-3
Attributes of the detected motif pairs of class 6-8 as observed in the mouse genome

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
tgagcg	ccccagct	21.9234	3.83e-14	2358	76
tgggcg	ccccagct	21.3443	9.73e-14	2046	69
cgcctg	ccagcggc	19.7696	2.13e-12	1406	53
cagcgc	gcaggggc	22.0212	6.89e-12	1883	62
cgccgc	ccagcggc	21.2666	1.92e-10	1394	49
gcgggc	gcaggggc	22.5219	3.39e-10	1370	48
tgggcg	ccagcggc	21.1979	3.73e-10	1280	46
ggtgcg	aacggttc	22.559	2.29e-09	236	18
gggcgg	aacggttc	21.8229	6.47e-09	252	18
gtcgct	acccccgc	22.9555	1.18e-08	501	25
gtgcgg	gcaggggc	22.5397	1.59e-08	586	27
gcgggc	aacggttc	21.3797	1.61e-08	509	25
cgcctg	ccccagct	23.326	1.02e-07	949	34
gggcgc	ccagcggc	22.3562	2.47e-07	1428	43
gccgcc	tgcgacag	22.4536	5.43e-06	124	10
tccgtg	tgcgacag	22.8663	6.63e-05	77	7

Table A-3 (Continued)

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
tgagcg	ttcagtgt	24.4059	3.96e-04	703	21
attcgt	gcgcagtg	20.0815	3.00e-12	429	28
cagcgc	ccccagct	22.5303	4.24e-12	1619	57
gcgcga	tgcttaga	20.7517	2.08e-11	642	33
gccgcc	ccagcggc	20.6882	1.58e-10	1156	44
cgccgc	ccagcggc	21.2666	1.92e-10	1394	49
gcacgt	tgtgtact	22.9874	4.11e-10	494	27
gtgcgg	cagcggcc	22.5157	1.14e-09	1049	40
cccgcc	cagcggcc	24.8087	1.77e-09	387	23
gcgcga	ggcagcgc	22.9681	1.95e-07	929	33
cgccgc	gcgcagtg	25.9939	2.76e-06	304	16
gtcgct	ggcagcgc	23.8128	7.82e-06	330	16
gggcgc	ggggcggg	23.7781	9.60e-06	2809	63
cccgcc	tggtctgc	25.4197	1.77e-05	240	13

TABLE A-4
Attributes of the detected motif pairs of class 7-7 as observed in the mouse genome

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
ccaccgc	gcagcgc	22.2074	1.49E-12	2426	74
ccaccgc	cccccgc	22.7426	4.79E-12	3373	90
ccccgcc	gcgggct	22.6273	8.22E-10	2439	68
ccgcccc	gctcgcc	23.5178	1.14E-09	3432	85
gcagcgc	gctcgcc	23.7185	1.16E-09	2028	60
cccccgc	gcagcgc	23.4937	3.54E-09	1521	49
ccgccgc	cgactcg	24.8409	2.51E-08	275	18
gcggggc	ctcgggt	22.3637	3.41E-08	1137	39
ccaccgc	cgactcg	26.1932	2.81E-06	235	14
cccccgc	cgactcg	26.7212	9.42E-06	193	12
ccccgcc	tgcgaca	24.9072	1.18E-05	637	23
cgtgtgt	taagtcg	26.7623	2.09E-05	439	18

Table A-4 (Continued)

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
cgcctgc	ggggcgg	24.1088	9.20E-11	2810	77
ccaccgc	gcgcagt	25.4507	1.63E-09	1153	42
ccgccgc	gcagcgc	23.9706	2.25E-09	1500	49
gcagcgc	gcgggct	22.8733	3.15E-09	1087	40
cgcctgc	ggagggg	25.3318	7.58E-09	2974	75
gctcgcc	gctctcc	23.9343	1.38E-08	2555	67
ccgcccc	cgtgtgt	27.3523	1.94E-08	2634	68
ccccgcc	ggccctg	25.1145	3.30E-08	1378	44
ccgcccc	gcgcagt	24.0816	3.10E-07	2195	57
cgcctgc	gcgggct	25.3398	5.96E-07	929	32
gctcgcc	gtagctg	25.6107	1.34E-06	1111	35
ccccgcc	cgcctgc	27.2526	1.60E-06	1802	48
ccgcccc	gcagcgc	26.2603	2.45E-06	1398	40
ccccgcc	tgcgaca	24.9072	1.18E-05	637	23

TABLE A-5
Attributes of the detected motif pairs of class 6-6 as observed in the mouse genome

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
gcgcagt	gtgcagat	21.2759	4.48e-12	502	30
cgtgtgt	gaggggac	24.465	9.23e-11	246	20
cccccgc	ggcagcgc	23.4131	2.66e-10	708	33
gcgggct	gtgcagat	24.2406	2.92e-10	158	16
ccaccgc	ggcagcgc	23.5676	7.53e-09	767	32
ccaccgc	cggagcag	24.3098	9.97e-09	389	22
ccgccgc	cggagcag	23.767	1.20e-08	393	22
ggagggg	gtgcagat	23.9499	3.30e-08	607	27
ccgcccc	ccagcggc	25.838	3.67e-07	728	28
gctctcc	tgtgtact	26.4662	1.28e-05	199	12
cgactcg	tagaaacc	24.7674	2.29e-05	146	10
ctccggg	ccagcggc	21.7382	2.17e-11	1355	50

TABLE A-5 (Continued)

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
ccaccgc	acccccgc	22.7254	2.95e-11	1054	43
ccgcccc	tagaaacc	24.2154	1.59e-10	342	23
gctcgcc	gcaggggc	25.3685	1.89e-10	477	27
gtagctg	gcaggggc	23.2156	2.97e-10	353	23
gcagcgc	acccccgc	23.9778	4.86e-10	362	23
cgcctgc	ttcagtgt	23.9173	6.53e-10	469	26
gtagctg	aacggttc	23.3883	8.22e-10	123	14
ggggcgg	ccagcggc	22.2725	1.23e-09	920	37
gctcgcc	gcatctcc	22.6317	2.33e-09	535	27
cgcctgc	tgcatctc	23.4268	4.62e-09	218	17
gctctcc	aacggttc	23.9362	8.28e-09	256	18
ccgcccc	ttcagtgt	25.7186	5.59e-08	706	29
cgactcg	ccagcgac	24.848	8.02e-08	79	10
gccagcg	cggagcag	24.8535	3.21e-07	196	14
ccaccgc	agaggctg	24.2132	4.46e-07	1059	35
gccagcg	gtgcagat	26.5545	1.42e-06	107	10
gactcgg	gtgcagat	25.7111	2.57e-06	67	8
tgcgaca	tttcacgt	23.0469	6.56e-06	100	9
gactcgg	agaggctg	25.3742	5.70e-05	657	22

Table A-6
Attributes of the detected motif pairs of class 8-8 as observed in the mouse genome

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
acccccgc	ccagcggc	21.6204	5.09E-14	392	29
acccccgc	ccagcgac	21.5338	2.94E-13	203	21
aacggttc	gcgcgagg	23.2809	1.78E-09	54	10
acccccgc	tgcatctc	22.1933	5.18E-09	97	12
gcatcatg	gcgcagtg	23.8556	1.05E-08	48	9
acccccgc	gcatcatg	23.6776	1.11E-06	27	6
cagcggcc	ccaccgcc	27.7737	2.48E-06	89	9
aacggttc	agaggctg	27.1444	1.29E-05	60	7

TABLE A-6 (Continued)

Motif 1	Motif 2	Information Content	P-value	Frequency In All Promoters	Frequency In Co-Regulated Promoters
ccccagct	tttcacgt	25.5314	1.78E-4	90	7
gcgcgagg	ggcagcgc	21.6866	6.68E-13	145	18
actcgggt	tgctctcg	21.8695	6.78E-12	23	9
gaggggac	tgtgtact	24.0518	2.47E-11	156	17
cgtgtgta	gaggggac	25.5001	1.42E-10	108	14
gcaggggc	ggggcggg	24.0481	1.98E-09	531	27
ccccagct	ggcagcgc	25.3908	2.19E-09	156	15
aacggttc	ggcagcgc	23.7515	2.59E-09	29	8
cagcggcc	ccagcggc	26.5778	5.72E-09	45	9
gcgcagtg	ttcagtgt	23.6768	2.61E-08	53	9
cgtgtgta	ccccagct	25.6502	5.12E-08	96	11
aggaaccc	gaggggac	23.718	1.07E-07	208	15
aacggttc	tagaaacc	25.2433	3.46E-07	92	10
aggaaccc	gcatctcc	24.5256	3.65E-07	169	13
tctgctaa	tgtgtact	24.6065	5.17E-07	96	10
ctggtgcg	gtgatctc	24.8686	2.08E-06	46	7
aggaaccc	gcgcagtg	27.25	6.70E-06	76	8
cttaggaa	tgctctcg	24.8572	2.84E-05	46	6