EPiC Series in Computing

Volume 60, 2019, Pages 27–35

Proceedings of 11th International Conference on Bioinformatics and Computational Biology

Identifying DNA Sequence Motifs of Pdx-1 and NeuroD1 Transcription Factors

Hassan Aldarwish, David R. Keller and Elena Y. Harris
California State University, Chico
Chico, California, 95929, USA
eyharris@csuchico.edu

Abstract

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. Two known transcription factors, Pdx-1 and NeuroD1, are shown to regulate gene expression in beta cells. Recently gene targets that are regulated by both Pdx-1 and NeuroD1 have been identified experimentally [7]. However, the motifs for this set of genes have not been found yet. Here we undertake the task of finding statistically overrepresented motifs in genes 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.

keywords: sequence analysis, motif discovery

1 Introduction

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 unveil regulatory networks and provide insights into the relations between transcription factors and their binding sites [4].

Several models are 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 [11], reflect the probability of the nucleotides occurring at a particular position in a motif. 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 [4]. 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 [10].

Motif identification is a complex process that consists of several stages: (1) Preprocessing (finding co-regulated genes sharing motifs in their promoters); (2) Motif detection; and (3) Postprocessing (clustering and scoring of the detected motif candidates).

Co-regulated genes share motifs in their promoters [9], where a promoter is usually defined as a sequence of 2000 base pairs upstream of the start of a gene. To identify motifs, the first step is to find a set of co-regulated genes. ChIP-seq (Chromatin immunoprecipitation sequencing) is a well-known technique to accomplish this task [1]. Using ChIP-seq, biologists identify and select the regions to which specific transcription factors bind, and 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. Once a set of co-regulated genes has been identified, their promoter sequences are extracted from the genome reference for further motif analysis.

After obtaining a set of sequences of co-regulated genes, the next step is to identify common motifs statistically overrepresented in the given set of sequences compared to a background sequences (usually promoters of all genes of an organism). Following is a description of two widely adopted strategies to identify motifs. 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 [3]. 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 [3].

The postprocessing step evaluates detected motifs using clustering and scoring. Grouping of similar motifs into clusters is called clustering. Clustering improves the significance of similar patterns and filters out spurious motifs [9]. Scoring involves estimating the statistical significance of motifs measured by the *p*-value [2] and calculating the false discovery rate, FDR.

Here we undertake the task of finding statistically overrepresented motifs in genes regulated by Pdx-1 and NeuroD1 given a set of their gene targets. 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. Here, we present a tool for finding overrepresented pairs of motifs given a set of sequences of interest and background sequences. Our method pipelines identification of pairs of motifs, clustering of pairs, and estimation of p-values and FDR for each cluster representative. Our tool supports multithreading to speed up calculations. We hope our tool might be useful for other biologists to facilitate their research. In addition to finding statistically overrepresented pairs, we further evaluate the top scored pairs by using phylogenetic conservation analysis, investigating positional bias relative to TSS (transcription start site) and analyzing information content of detected candidate motif pairs.

2 Methodology

To detect motif pairs, we enumerate all distinct 6-, 7-, 8-mers in the given set of promoters of co-regulated genes in the mouse genome [7]. The *k*-mers are, then, filtered according to a scoring criterion, which is described in section 2.1. Next, we evaluate every possible pair formed out of the selected *k*-mers by scoring each motif pair. Since our approach is enumerative, it is very likely to find very similar motif pairs. To avoid duplicity, we cluster highly similar motif pairs according to the Tanimoto distance. To further assess the biological significance of motif pairs, we analyze their conservation in two different genomes, namely rat and human. Lastly, to estimate 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.

2.1 Motif Pairs Identification and Scoring

The primary goal of this analysis is to measure the statistical significance of occurrences of motif pairs in the promoters of the given set of co-regulated 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:

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

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.

The set of mutants that minimizes the *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 the mouse genome. Then, the ten most significant mutants, having *p*-value less than 0.01, are processed by a heuristic from [6] to choose mutants that minimize the *p*-value of *MP*. Briefly, the heuristic uses dynamic programming to compute the *p*-value of combinations of the 10 lowest *p*-value mutants and choses the combination that has the lowest *p*-value.

Due to large search space, we restricted our search to the motif pairs that were formed from single statistically significant motifs whose *p*-value was less than or equal to 0.01 (*p*-value is calculated by the same formula). Then we formed all possible motif pairs from the set of statistically significant single motifs (a valid motif pair consists of two non-overlapping single motifs that occur in a single promoter), 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. We considered different combinations 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.

2.2 False Discovery Rate Estimation

Here we discuss our randomized analysis to identify the 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 in 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 shown to be efficient in controlling the false discovery rate [13]; 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 randomly choosing a set of co-regulated genes of the same size as 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, 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.

2.3 Motif Pairs Clustering

In order to weed out highly similar motif pairs, we used a clustering technique. We used the Tanimoto distance [6] as a metric to identify similar motif pairs. For two aligned motif pairs, *M1* (*a*, *b*) and *M2* (*c*, *d*), the Tanimoto distance *T*(*M1*, *M2*) 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.

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

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 beginning of the process. Complete linkage dictates that the maximum pairwise distance of motif pairs in a cluster is less than the threshold (0.5).

2.4 Phylogenetic Conservation Analysis

Here we discuss orthologous analysis of the selected statistically significant motif pairs. 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 [14]. To examine the biological significance of the identified statistically significant motif pairs, we calculated their *p*-values 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 the one described to identify 5% FDR threshold to distinguish statistically significant motif pairs.

2.5 Positional Analysis

To study positional preferences of a selected 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 pair in a window of a given size. In particular, we counted occurrences 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. Initially we subdivided 2000bp promoter into 2000 bins, and 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 in finding the occurrences.

2.6 Information Content Analysis

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* at position *i* and $p_b(x)$ is its probability in the background (all promoters) [5]:

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

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. Information content, *IC*, reflects the certainty of the content of motif pairs at each base. 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 of *IC* is 2; it indicates absolute certainty that the character at that position occurs not by a random chance.

3 Results

We used a set of co-regulated genes from the recent study that identified 277 genes with binding sites of Pdx-1 and NeuroD1 transcription factors in the mouse genome [7]. 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 genomes [8]. Total promoters used for the background sets included 21,204 mouse promoters, 14,205 rat promoters and 39,275 human promoters. Total promoters of co-regulated genes included 277, 201 and 214 mouse, rat and human promoters respectively.

After extracting all distinct single *k*-mers of sizes 6, 7, and 8 characters from the set of co-regulated promoters in the mouse genome and calculating their *p*-value, we selected *k*-mers whose *p*-value was less than or equal to 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 selecting mutants of each motif pair and computing the *p*-value for each motif. The ten mutants with the lowest *p*-values of each motif pair were processed by the dynamic programming heuristic to choose those that minimize the *p*-value of a motif pair. 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.

We further investigated biological evidence for the significant motif pairs. First, motif pairs of Tanimoto distance less than 0.5 were filtered out. 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. Table 1 summarizes our results.

Pair Size	Total Pairs	Statistically Significant Pairs	Pairs After Clustering	Pairs Statistically Significant In Rat or Human
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

Table 1: Total Motif Pairs After Each Step of Analysis

We have evaluated a total of 4465 motif pairs, which were formed using all single significant distinct *k*-mers in the co-regulated promoters of mouse genome. As a result, we have detected 178 motif pairs that are statically significant in at least two of three different genomes.

The *p*-values of the resulting motif pairs range from 1.89e-15 to 1.78e-4, as observed in the mouse genome. Distribution of *p*-value thresholds obtained from randomized analysis in mouse, rat and human genomes is provided in Table 2. These thresholds ensure that false discovery rate does not exceed 5%. The thresholds are shown for all combinations of lengths of two motifs in a pair.

	6-6	6-7	6-8	7-7	7-8	8-8
Mouse	0.06	0.06	0.07	0.08	0.08	0.08
Rat	0.05	0.06	0.08	0.07	0.08	0.1
Human	0.06	0.06	0.07	0.06	0.08	0.1

Table 2: *p*-value Thresholds Ensuring FDR of 5%

To investigate 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 genome and used these values to construct sequence logos. Figure 1 illustrates the conservation level of the motif pairs using sequence logos (created by *Seq2Logo*, [14]). The first column of the table shows the most conserved motif pairs in each class (by lengths of motifs in a pair), whereas the second column shows the least conserved. The scale of the vertical axis, *IC*, is in bits from 0 to 2. This example demonstrates that the detected candidates of motif pairs are fairly well conserved.

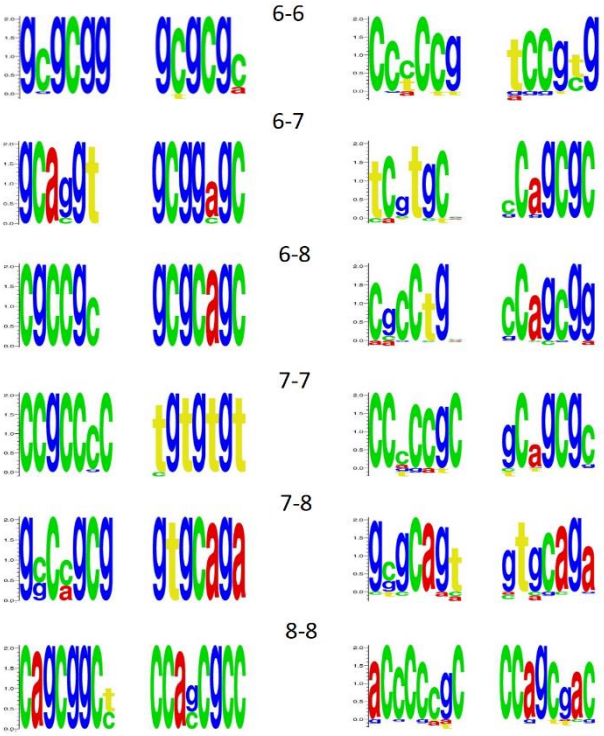

Figure 1: Sequence Logos of the Resulting Motif Pairs

The results of the positional analysis in promoters of mouse included all identified motif pairs and different window width. For the window width equal to 50, 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).

4 Conclusion

Using the set of co-regulated genes for Pdx1 and NeuroD1 transcription factors, we identified a total of 178 candidate 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 the rat or the human genomes. Though we considered mutated instances (or non-exact matches) of motif pairs in our evaluation, the information content of the detected motif pairs indicates a high level of specificity; most of the positions of motif pairs are well conserved. In addition, 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. Our study narrowed down the search of the statistically significant motif pairs from 4465 to 178, which we hope will help biologists to perform targeted experimental verification of our results.

Our tool is available upon request. To speed up calculations, our tool supports multithreading.

References

[1] Timothy Bailey, 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. *PLoS Computational Biology*, 9(11): e1003326, 2013. DOI: 10.1371/journal.pcbi.1003326

[2] Yoseph Barash, Gill Bejerano and Nir Friedman. 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. DOI: 10.1007/3-540-44696-6_22

[3] Modan K. Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8(7): S21, 2007. DOI: 10.1186/1471-2105-8-S7-S21

[4] Patrik D'haeseleer. What are DNA sequence motifs? *Nature Biotechnology* 24(4): 423-425, 2006. DOI: 10.1038/nbt0406-423

[5] Richard Durbin. *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press, 1998.

[6] Elena Y. Harris, 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): 601, 2011. DOI: 10.1186/1471-2164-12-601

[7] David M. Keller, Shannon McWeeney, Athanasios Arsenlis, Jacques Drouin, Christopher VE Wright, Haiyan Wang, Claes B. Wollheim, Peter White, Klaus H. Kaestner, and Richard 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): 32084-32092, 2007. DOI:10.1074/jbc.M700899200

[8] James W. Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996-1006, 2002. DOI: 10.1101/gr.229102

[9] Kenzie D. Macisaac and Ernest Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Computational Biology*, 2(4): e36, 2006. DOI: 10.1371/journal.pcbi.0020036

[10] Gary D. Stormo. DNA Binding Sites: Representation and discovery. Bioinformatics 16(1): 16-23, 2000. DOI: 10.1093/bioinformatics/16.1.16

[11] Paul Stothard. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28:1102-1104, 2000. DOI: 10.2144/00286ir01

[12] Martin Christen Frølund Thomsen and Morten Nielsen. 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. *Nucleic Acids Research*, 40(W1): W281-W287, 2012. DOI: 10.1093/nar/gks469

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

[14] Zhaolei Zhang and Mark Gerstein. Of mice and men: Phylogenetic footprinting aids the discovery of regulatory elements. *Journal of Biology*, 2(2): 11, 2003. DOI: 10.1186/1475-4924-2-11