

Systematic Detection of Statistically Overrepresented DNA Motif Association Rules

Jane Marie Lin¹ Zhiping Weng^{1,2}
janemlin@bu.edu zhiping@bu.edu

¹ Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

² Program in Bioinformatics and Systems Biology, Boston University, Boston, MA, 02215, USA

Abstract

DNA motifs, or cis-elements, are short nucleotide sequence patterns recognized by various transcription factors (TFs). In promoters, these TFs bind in a complex combinatorial manner in order to regulate the expression of a downstream gene. The combinatorial space is frequently large and difficult to manage since vertebrates have thousands of transcription factors and more than 20,000 genes. We introduce a computer program called CAYCE (Combinatorial Analysis of Cis-Elements) that systematically detects statistically overrepresented DNA motif association rules independent of Microarray information. CAYCE is an adaptation of the *a priori* algorithm traditionally used for association rule mining, but offers three significant advancements. (1) It analyzes multiple occurrences of an item, corresponding to multiple TF binding sites, (2) It compares results with a biologically relevant background, and (3), it provides p-values for straightforward statistical interpretation. CAYCE can be easily applied to any item-set data where the investigator is also interested in multiple occurrences of a single item, and/or overrepresentation of association rules compared with a background. Applying CAYCE to human promoters in 1% of the human genome, we discover that motif clusters containing five repetitions of SP1 are the most statistically significant.

Keywords: transcription factor, combination, association rule

1 Introduction

Computational prediction of transcription factor binding sites (TFBS) in the form of DNA motifs or cis-elements is common practice in computational biology. These algorithms come in several flavors: Expectation maximization (MEME [4], Improbizer [3]), Gibbs sampling (AlignACE [18], GLAM [11], MotifSampler [26]), simple word counting (YMF [24, 25], FindExplanator [5], MITRA [9]), position specific scoring matrix scanning (ANN-Spec [31]) by greedy algorithm (Consensus [16]), and multi-species comparisons [32]. The result of this analysis typically yields a list of motifs and their respective TFBS locations, although it is well known that these predictions have high false positive rates in eukaryotic genomes. One way to prune out these false positives is to rely on the biological insight that truly functional TFBS often appear in clusters, modules, or regulatory circuits [13, 29]. Various tools have also been developed to model these motif modules, again with an assortment of flavors: Hidden Markov models (MetaMEME [14], Cluster-Buster [12]), Gibbs sampling (Gibbs Module Sampler [27]), Monte Carlo motif screening (EMCModule [15]), and hierarchical mixture modeling (CisModule [33]). Because of the computational complexity of these algorithms, many module finders recommend a limit on the number of motifs inputted to the program. Many also require complex model parameters to be pre-specified, and it is often difficult to get simple statistics for the key motifs in those modules. Motivated by these existing shortcomings, we developed a tool called CAYCE: Combinatorial Analysis of Cis-Elements to extract significant combinatorial relationships among motifs, providing intuitive and

biologically familiar assessment statistics such as conditional probabilities and p-values. The main advantage of CAYCE is its simplicity in reporting, providing the investigator with an intuitive understanding of the motif distributions in her dataset. It helps the investigator answer three pertinent biological questions in module identification: What are common combinatorial rules governing associations among different motifs? Which motifs have repeated binding sites? And how many times is this motif repeated? While CAYCE is not a module finder per se, it does provide an informative compilation of relevant motif combinations that can be used as inputs to guide a module finding algorithm. In this sense, CAYCE bridges the gap between motif discovery and module discovery. In actuality, CAYCE is highly versatile. It can be used as an initial exploratory tool since it does not require the use of expression information. An investigator can use it to compare motif combinations by randomizing the current sequence set, versus randomizing a biologically relevant background sequence set. It can even be used *after* cluster/module discovery to determine directional relationships among motifs within those clusters/modules.

CAYCE is a biological adaptation of a well-known association rule mining algorithm called *apriori*, first developed by Agrawal *et al.* in 1993 [1, 2, 6, 7], and the current implementation is by Borgelt *et al.* [6, 7]. The *apriori* algorithm has been widely used in e-commerce and marketing, allowing a retailer to automatically generate personalized product recommendations based on one's transaction history and other transactions made with similar item selections. It employs breadth-first search and uses a hash-tree structure to efficiently enumerate item sets and discover important associations based on a support and confidence framework. In the classical example, the transaction histories of several customers are stored in a database. Customer A buys bread, milk, and cheese; customer B buys bread and milk; and customer C buys bread and cheese. Item sets can be formed by any subset of bread, milk and cheese, and association rules relate the occurrence of one item set with another. The association rule describing how likely a customer who buys cheese also buys bread is expressed as bread \leftarrow cheese, where cheese is the antecedent and bread is the consequent. This association rule is evaluated by its *support*, defined as the joint probability of observing the consequent and the antecedent together, in this case $P(\text{bread} \cap \text{cheese})$. It is also evaluated by its *confidence*, defined as the conditional probability of observing the consequent given the antecedent, in this case $P(\text{bread}|\text{cheese})$. From the transaction histories of customer A, B, and C, we can see that the bread \leftarrow cheese association rule is quite favorable since its support is $2/3$ (66%), and its confidence is $2/2$ (100%).

While it may not be very interesting to analyze how many loaves of bread a customer buys, how many times the same sequence motif appears is correlated with its binding in living cells. Analyses of TFBS verified by ChIP experiments have shown that TFBS frequently have many repetitions of the motif, although the precise number is unknown [20]. We have adapted our computer program CAYCE to handle these repetitions by appending a counter to the end of an item, and then filtering out trivial rules from the output. We have also adapted CAYCE to analyze association rules relative to a user-specified biologically relevant background because biological signals are frequently context dependent [10, 28]. For example, an investigator interested in discovering association rules among motifs in the promoters of housekeeping genes will observe different significance levels depending on the background, whether it is composed of randomly shuffled item sets or a biologically dissimilar sequence set (i.e. promoters of tissue specific genes). Our computer program CAYCE is equipped to perform both of these analyses.

We were also dissatisfied by the interpretability of the support and confidence measures, and wanted to obtain a more statistically and biologically familiar metric of evaluation. We insisted on obtaining a p-value for an association rule despite the computational cost. By randomly shuffling the item sets, re-distributing them according to the size of each transaction, and performing this randomization 1000 times, we obtain a distribution of confidence scores for a given association rule. The central limit theorem dictates that the resultant distribution is normal, and the p-value is obtained from the Z-score of the confidence of a rule.

2 Methodology

An investigator can discover common combinatorial rules among different motifs by examining output from the unique item set analysis, find out which motifs are repeated, and how many times by examining output from the multiple-occurrence item set analysis. We also remove redundant association rules that can easily confuse the investigator. For example, examining the rules $A \leftarrow B.1$ and $A \leftarrow B.2$ individually does not reveal the relationship between A and B in general. A more appropriate formulation would first examine the relationship between A and B without repetition ($A \leftarrow B$), and separately examine any repeating occurrences of A or B ($B.i \leftarrow B.j$ and $A.i \leftarrow A.j$). The drawback of this separation, however, is that the probabilities of observing a single item (i.e. $P(A)$ and $P(B)$) are underestimated in the unique item set analysis because repetitions are removed. They are also underestimated in the multiple item set analyses because sequences that contain only singly occurring motifs are removed from the calculation. We are developing new ways to circumvent this drawback and they will be included in newer versions of CAYCE.

2.1 Handling Multiple-Occurrence Item Sets

A list of motifs in promoters is separated into two files containing: (1) unique item sets and (2) multiple-occurrence item sets. They are separately analyzed using the *apriori* algorithm, but output from the multiple-occurrence item set list is further filtered to remove trivial and non-essential association rules. For example, four promoters contain motifs A, B, C, D and E.

Foreground:

P₁: A, A
 P₂: A, B
 P₃: B, C, B, C
 P₄: C, D, E

2.1.1 Unique Item Sets

A file that contains only unique item sets is created, and the *apriori* algorithm is run with this input:

P₁: A
 P₂: A, B
 P₃: B, C
 P₄: C, D, E

2.1.2 Multiple Item Sets

A file that contains only multiple-occurrence item sets is created, and the *apriori* algorithm is run with this input:

P₁: A.1, A.2
 P₂: B.1, B.2, C.1, C.2

Output from the *apriori* algorithm is passed through two filters. (1) Rules involving different motifs of the form $X_i \leftarrow Y_j$ where $X \neq Y$ are removed because these relationships are captured in the unique item set analysis. (2) Trivial rules of the form $X.i \leftarrow X.j$ where $i < j$ are removed. These rules are pruned out if the maximum repetition in the consequent is less than the maximum repetition in the antecedent of the rule ($MaxRep(i) < MaxRep(j)$) such that rules with more than one repetition in either the antecedent or consequent is first collapsed to a simpler form. For example, the rule $X.1, X.2 \leftarrow X.3, X.4$ is first collapsed into $X.2 \leftarrow X.4$ and since $2 < 4$, it is eliminated from the output set of rules.

2.2 Statistical Overrepresentation

2.2.1 Foreground Randomization

The aggregate list of items from promoters in the foreground is shuffled, and re-distributed according to the number of motifs in each promoter.

Aggregate list: {A, A, B, B, C, B, C, C, D, E}

Shuffled list: {C, B, E, A, B, C, C, D, A, A, B}

Re-distribute:

P₁: C, B

P₂: E, A

P₃: B, C, C, D

P₄: A, A, B

One thousand randomizations produce new lists that are then split into unique and multiple-occurrence item sets for efficient analysis by *apriori*. The effect of this shuffling is to randomize the joint probabilities of observing more than one item $P(X_1 \cap X_2 \dots \cap X_n)$, but fixing the probabilities of observing a single item $P(X_n)$.

2.2.2 Background Randomization

A different aggregate list of items is obtained from biologically relevant promoter sequences forming the background set. It may contain motifs not identified in the foreground set (although only motifs in common will be evaluated), and may be distributed differently:

Background:

P₁: B, C

P₂: E

P₃: A, B, C

P₄: A, A, F

Aggregate list: {B, C, E, A, B, C, A, A, F}

Shuffled list: {C, F, A, B, B, C, E, A, A}

Re-distribute:

P₁: C, F

P₂: A

P₃: B, B, C

P₄: E, A, A

One thousand randomizations are also performed, but in this case, the probabilities of observing a single item $P(X_n)$ is no longer fixed by the foreground.

2.3 P-Value Estimation

Final evaluation of the *apriori* output is performed using statistical overrepresentation. The default support and confidence cutoffs for running the *apriori* algorithm are set purposely low (10% and 30% respectively), since this produces many candidate rules. Association rules that have low confidence may actually be statistically overrepresented relative to the background, and rules with high confidence in the foreground may not be statistically significant if it occurs frequently and with high confidence in the background.

For each association rule discovered in the foreground set, a list of confidences is collected from occurrences of that rule in the 1000X shuffled foreground and 1000X shuffled background. Recall that the confidence of a rule is simply the conditional probability of observing the consequent given the antecedent, which is in turn determined by enumerating item set occurrences. According to the central

limit theorem, any sum of independent identically distributed random variables will approximately tend to the normal distribution. We can safely assume here, that the distribution of association rule confidences derived from randomized shuffling of aggregate item lists will be approximately normally distributed, since each randomization is independently generated and contains identical distributions of the items.

Two p-values are obtained for an association rule given the mean and standard deviation of confidence scores derived from the randomizations. The first one is based on randomizations of the foreground aggregate item list, and the second one is based on randomizations of the background aggregate item list. If the investigator queries with large sets of motifs, we find many association rules with p-values of dead zero, which means that the rule was never detected in the 1000 randomizations, and these rules are highly unlikely to occur by chance. A schematic of the full CAYCE methodology is detailed in Figure 1 below.

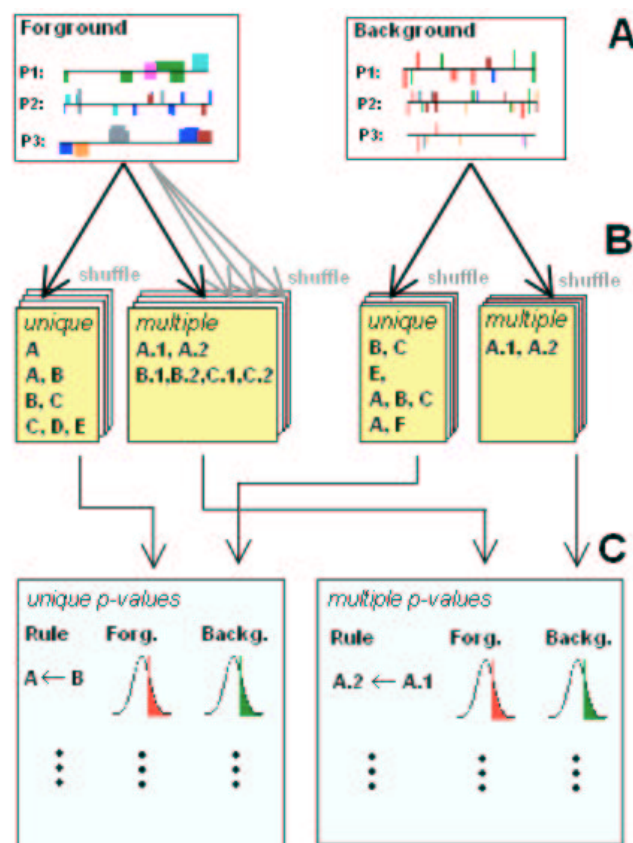


Figure 1: Schematic of CAYCE methodology. (A) DNA sequence motifs (colored boxes) form clusters on promoters (black lines). If the box is above the black line, then the motif is oriented in the forward direction, and vice versa. (B) The motif occurrences are collected into item sets, separated into unique and multiple-occurrence patterns and aggregate lists are shuffled 1000X. (C) Two p-values are calculated for an association rule. One is relative to the confidence scores of that association rule in randomizations generated from the foreground aggregate list, and another is relative to randomizations generated from the background aggregate list.

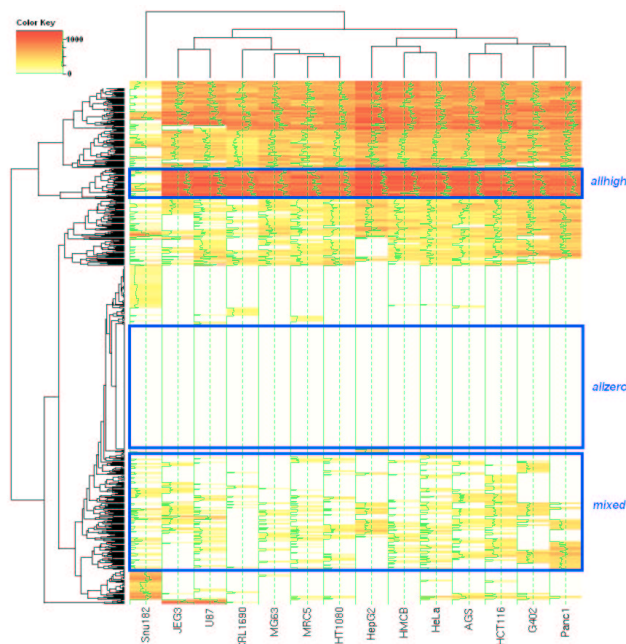


Figure 2: Promoter activities of 642 promoters measured across 14 cell lines. After hierarchical clustering, we focused on three regions of interest.

3 Results in Human Promoters

We applied our method to discover statistically overrepresented motif association rules in human promoters. This is an example of applying *CAYCE* after module discovery to obtain simple statistics and combinatorial relationships for motifs *within* the discovered modules. We obtained promoter activity profiles of 642 promoters tested across 14 cell lines [8], available for download on the UCSC Genome Browser [17, 19]. This dataset is different from Microarray expression data in that it does not measure mRNA transcript level. Promoter activity measures the regulatory potential of a promoter based on DNA sequence composition alone since epigenetic markers like methylation and chromatin are stripped in the assay. Therefore, it is the most appropriate experimental platform to base motif/sequence models of promoters. We applied hierarchical clustering to the dataset, and from the heatmap, we noticed a small region of ubiquitous and high promoter activity, a region of mixed activity, and a region of low promoter activity across all cell lines (Figure 2). We named them *allhigh*, *mixed*, and *allzero* respectively. Using YMF [24, 25, 28] followed by FindExplanators [5], we discovered a total of 77 overrepresented motifs in the 642 promoters.

Many of the 77 discovered motifs were redundant or reverse compliments of each other, and had significant overlaps with promoter motifs reported by Xie *et al.* [32]. We decided to investigate the top 8 known motifs that overlapped our list and Xie’s list. These are Nrf-1, Myc, Elk-1, Nf-Y, Sp1, AP-1, YY1, and GABP. These motifs were used as input to Cluster-Buster [13], which finds dense clusters of motifs in DNA sequences. Cluster-Buster outputs motif clusters that score above a log-likelihood threshold (we used 5), however these clusters do not convey any working relationships between motifs. In order to detect these relationships, we formatted Cluster-Buster output and analyzed the motif content of each cluster using *CAYCE*. We ran *CAYCE* twice to compare association rules discovered in the *allhigh* promoter set (foreground) with those in the *mixed* and *allzero* promoter set (two backgrounds). Comparison between the *allhigh* and *allzero* promoter sets emphasizes motif association rules that are important in functional promoters versus non-functional promoters. Comparison between the *allhigh* and *mixed* promoter sets emphasizes motif association rules that are important in

ubiquitously active promoters versus promoters of tissue specific genes.

3.1 Run 1: *Allhigh* versus *Allzero*

The significant ($p < 0.1$) association rules are indicated by bold print in the second and third column of Table 1. We observe several well-known transcription factor associations among Nrf-1, GABP and SP-1 [22, 23], indicating that the statistically significant interactions detected by CAYCE are also biologically relevant. We also discover novel associations with Elk-1, indicating that CAYCE is capable of discovering new associations. CAYCE can also pick up motifs that occur multiple times in a cluster, and output the repetition number that is statistically overrepresented. For promoters in the *allhigh* set, we find that single, double, and triple SP1 repeats are not statistically significant, but clusters containing 5 SP1 repetitions are significant. The remaining multiple-occurrence motifs do not appear to be statistically significant relative to randomizations of the foreground, but are still overrepresented relative to the *allzero* promoter set. This underscores the importance of appropriate background selection when analyzing biological data. While repetitions of Elk-1, AP-1, GABP, and Myc motifs may not be statistically significant among randomizations of ubiquitous and highly expressed genes (*allhigh* foreground), they are significant when compared to non-functional promoters. This is a way of using statistical significance to harness biological relevance in a context dependent manner.

3.2 Run 2: *Allhigh* versus *Mixed*

The significant ($p < 0.05$) association rules are indicated by bold print in the second and fourth column of Table 1. In this analysis the top association rules are the same as those in the first analysis, since we are using the same *allhigh* foreground. However, looking at the background 2 p-value column, we notice that most of them are no longer statistically overrepresented compared to the *mixed* background. In addition, triple repetitions of Elk1, triple repetitions of AP-1, and double repetitions of GABP are not significant at all, whereas they were significant relative to the *allzero* background. This suggests that promoters with mixed activities or tissue specific expression may have more repetitive occurrences of GABP and Elk-1, which is a possibility given that the transcriptional activity of Elk-1 is highly tissue specific [21]. In the previous analysis, association rules were added because of their significance over a biologically relevant background. In this example, association rules are eliminated because of their non-significance over a biologically relevant background.

4 Conclusions and Discussion

CAYCE is customized to answer biologically relevant questions about motif co-occurrence relationships. In particular, it takes into account repeated occurrences of a motif and determines the most significant repetition. Despite computational cost, we took pains to obtain p-values for association rules because it is the standard evaluation statistic employed by the biological community. This p-value is estimated relative to randomization of the foreground, and randomization of a biological contrast background.

We tested CAYCE on Human promoters and several known associations Nrf-1, GABP and Sp1 were discovered. The unique item set analysis suggests a potentially novel interaction between Elk-1 and the three motifs Nrf-1, GABP and Sp1. Our major contribution is in the multiple item set analysis where we discovered that the most significant of repetition of Sp1 in motif clusters is five.

CAYCE is a work in progress, and the most up to date source code is available upon request. We are also developing a web server for CAYCE, and streamlining steps in the analysis pipeline.

Table 1: Summary of association rules detected by CAYCE. Significant p-values ($p < 0.05$) are indicated by bold print.

Unique Motif Rules	Allhigh p-value	Allzero p-value	Mixed p-value
Elk-1 \leftarrow NF-Y	0	0	0.921
Sp1 \leftarrow Elk-1, Nrf-1	7.70E-09	0	0
Sp1 \leftarrow Nrf-1	3.50E-03	0	0
Sp1 \leftarrow GABP	9.70E-03	0	0.24
Sp1 \leftarrow Elk-1, GABP	0.016	0	0.29
Multiple-Occurrence Motif Rules	Foreground p-value	Background 1 p-value	Background 2 p-value
Sp1.5 \leftarrow Sp1.4	0.0297	0	0
Elk-1.3 \leftarrow Elk-1.2	0.298	0	0.108
AP-1.3 \leftarrow AP-1.2	0.433	0	0.963
GABP.2 \leftarrow GABP.1	0.4532	0	0.466
Myc.2 \leftarrow Myc.1	0.5398	0	0

Acknowledgments

This work was funded by the ENCODE grant R01HG03110 from NHGRI, NIH to ZW, and partly funded by the Training Program in Quantitative Biology and Physiology T32GM008764-06 from NIH to JML. The authors would like to thank Ulas Karaoz, Sara J Cooper and Nathan D Trinklein for scientific discussions and clarification of the technical details relating to the promoter activity dataset.

References

- [1] Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in massive databases, *Proc. of the ACM-SIGMOD Int'l Conference on Management of Data*, 207–216, 1993.
- [2] Agrawal, R. and Srikant, R., Fast algorithms for mining association rules, *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, 487–499, 1994.
- [3] Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., and Mango, S.E., Environmentally induced foregut remodeling by pha-4/foxa and daf-12/nhr, *Science*, 305(5691):1743–1746, 2004.
- [4] Bailey, T.L. and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2:28–36, 1994.
- [5] Blanchette, M. and Sinha, S., Separating real motifs from their artifacts, *Bioinformatics*, 17 Suppl. 1:S30–S38, 2001.
- [6] Borgelt, C. and Kruse, R., Induction of association rules: Apriori implementation, *In Proceedings of the 15th Conference on Computational Statistics*, 2002.
- [7] Borgelt, C., Efficient implementations of apriori and eclat, *Workshop of Frequent Item Set Mining Implementations*, 2003.

- [8] Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M., Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome, *Genome Res.*, 16(1):1–10, 2006.
- [9] Eskin, E. and Pevzner, P.A., Finding composite regulatory patterns in DNA sequences, *Bioinformatics*, 18 Suppl. 1:S354–S363, 2002.
- [10] Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., and Weng, Z., Detection of functional DNA motifs via statistical overrepresentation, *Nucleic Acids Res.*, 32(4):1372–1381, 2004.
- [11] Frith, M.C., Hansen, U., Spouge, J.L., and Weng, Z., Finding functional sequence elements by multiple local alignment, *Nucleic Acids Res.*, 32(1):189–200, 2004.
- [12] Frith, M.C., Hansen, U., and Weng, Z., Detection of cis-element clusters in higher eukaryotic DNA, *Bioinformatics*, 17(10):878–889, 2001.
- [13] Frith, M.C., Li, M.C., and Weng, Z., Cluster-buster: Finding dense clusters of motifs in DNA sequences, *Nucleic Acids Res.*, 31(13):3666–3668, 2003.
- [14] Grundy, W.N., Bailey, T.L., Elkan, C.P., and Baker, M.E., Meta-meme: Motif-based hidden Markov models of protein families, *Comput. Appl. Biosci.*, 13(4):397–406, 1997.
- [15] Gupta, M. and Liu, J.S., De novo cis-regulatory module elicitation for eukaryotic genomes, *Proc. Natl. Acad. Sci. USA*, 102(20):7079–7084, 2005.
- [16] Hertz, G.Z. and Stormo, G.D., Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics*, 15(7-8):563–577, 1999.
- [17] Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., Hillman-Jackson, J., Kuhn, R.M., Pedersen, J.S., Pohl, A., Raney, B.J., Rosenbloom, K.R., Siepel, A., Smith, K.E., Sugnet, C.W., Sultan-Qurraie, A., Thomas, D.J., Trumbower, H., Weber, R.J., Weirauch, M., Zweig, A.S., Haussler, D., and Kent, W.J., The UCSC genome browser database: Update 2006, *Nucleic Acids Res.*, 34(Database issue):D590–D598, 2006.
- [18] Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M., Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, *J. Mol. Biol.*, 296(5):1205–1214, 2000.
- [19] Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J., The UCSC table browser data retrieval tool, *Nucleic Acids Res.*, 32(Database issue):D493–D496, 2004.
- [20] Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B., and Marchal, K., Inferring transcriptional modules from ChIP-chip, motif and microarray data, *Genome Biol.*, 7(5):R37, 2006.
- [21] Rao, V.N., Huebner, K., Isobe, M., Ar-Rushdi, A., Croce, C.M., and Reddy, E.S., Elk, tissue-specific ets-related genes on chromosomes x and 14 near translocation breakpoints, *Science*, 244(4900):66–70, 1989.
- [22] Rosmarin, A.G., Resendes, K.K., Yang, Z., Mcmillan, J.N., and Fleming, S.L., Ga-binding protein transcription factor: A review of gabp as an integrator of intracellular signaling and protein-protein interactions, *Blood Cells Mol. Dis.*, 32(1):143–154, 2004.

- [23] Scarpulla, R.C., Nuclear activators and coactivators in mammalian mitochondrial biogenesis, *Biochim. Biophys. Acta*, 1576(1-2):1–14, 2002.
- [24] Sinha, S. and Tompa, M., A statistical method for finding transcription factor binding sites, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:344–354, 2000.
- [25] Sinha, S. and Tompa, M., Ymf: A program for discovery of novel transcription factor binding sites by statistical overrepresentation, *Nucleic Acids Res.*, 31(13):3586–3588, 2003.
- [26] Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y., A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling, *Bioinformatics*, 17(12):1113–1122, 2001.
- [27] Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S., and Lawrence, C.E., Decoding human regulatory circuits, *Genome Res.*, 14(10A):1967–1974, 2004.
- [28] Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., Van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z., Assessing computational tools for the discovery of transcription factor binding sites, *Nat. Biotechnol.*, 23(1):137–144, 2005.
- [29] Vavouri, T. and Elgar, G., Prediction of cis-regulatory elements using binding site matrices—the successes, the failures and the reasons for both, *Curr. Opin. Genet. Dev.*, 15(4):395–402, 2005.
- [30] Workman, C.T. and Stormo, G.D., Ann-spec: A method for discovering transcription factor binding sites with improved specificity, *Pac. Symp. Biocomput.*, 5:467–478, 2000.
- [31] Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M., Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals, *Nature*, 434(7031):338–345, 2005.
- [32] Zhu, Z., Shendure, J., and Church, G.M., Discovering functional transcription-factor combinations in the human cell cycle, *Genome Res.*, 15(6):848–855, 2005.