MotifOrganizer: A scalable two-stage model-based clustering approach for grouping conserved non-coding elements in mammalian genomes

Zhaohui Qin[1*], Misha Bilenky[2], Gordon Robertson[2], Gang Su[3], Steven Jones[2*]

[1] Department of Biostatistics, University of Michigan, Ann Arbor MI 48109

[2] BC Cancer Agency Genome Sciences Centre, Vancouver BC Canada

[3] Bioinformatics Program, University of Michigan, Ann Arbor MI 48109

* Corresponding authors:

Steven Jones	Zhaohui Qin
BC Cancer Agency Genome Sciences	Center for Statistical Genetics
Centre,	Department of Biostatistics
675 West 10th Avenue,	University of Michigan
Vancouver BC V5Z 4S6	Ann Arbor, MI 48109
Canada	Email: qin@umich.edu
Email: sjones@umich.edu	

Running title: A scalable model-based motif clustering approach

Key words: model-based, clustering, conserved non-coding elements, motif, scalability

Abbreviations

CNE:	Conserved Non-coding Elements
TF:	Transcription Factor
TFBS:	Transcription Factor Binding Site
BMC:	Bayesian Motif Clustering
PWM:	Position specific Weight Matrix
NBR:	Normalized Bayes Ratio
PAP:	Posterior Assignment Probability
KL:	Kulback-Leibler
ARI:	Adjusted Rand Index

Abstract

After conserved non-coding elements (CNEs) and DNA sequence motifs are identified using the phylogenetic conservation and sequence specificity, recognizing groups of similar motifs is a critical step in translating CNEs into putative regulatory information. Model-based methods explicitly address uncertainty in clustering motifs. Because they use no distance metric, no need to calculate cumbersome pairwise distance matrix, and base inference on relationships between motifs and clusters rather than between motifs, they should be highly scalable. In the work reported here we describe MotifOrganizer, a scalable model-based clustering tool for grouping CNEs identified from large scale comparative genomics studies. It extended BMC, a Bayesian clustering method that previously had correctly identified prokaryote regulons. The new algorithm allows grouping of motifs and motif clusters with different widths and the novel two-stage operation scheme further increases the scalability of model-based clustering algorithms. Comparing to distance-based and one-stage model-based clustering tools, MotifOrgainzer showed favorable performance on motif sets selected from TRANSFAC and JASPAR. Test on ~30k and then ~150k CNEs from the cisRED human v.2 database demonstrated that MotifOrganizer can effectively cluster whole genome sets of mammalian CNEs.

The MotifOrganizer program, data used in this study and Supplementary material are available at http://www.sph,umich.edu/csg/qin/motif.

Introduction

Assembling a comprehensive catalog of all transcription factors (TFs) and the genes that they regulate--regulon is important in clarifying how gene expression is regulated. An important step towards achieving this goal involves identifying the regulatory elements bound by TFs. Success in finding these elements contributes to our understanding of TF-regulon relationships and provides critical insight into the mechanism of transcription regulation. With the fast accumulation of completely sequenced genomes, the combination of *ab initio* motif searching algorithms (MacIsaac and Fraenkel, 2006; Tompa, et al., 2005; van Nimwegen, 2007) and comparative genomics strategies constitutes a powerful approach for identifying novel regulatory elements (Stone, et al., 2005). Several recent studies carried out genome-wide motif discovery using sequence specificity and phylogenetic conservation information to identify motifs that are conserved at a higher-than-expected rate in non-coding regions (Elemento and Tavazoie, 2005; Kellis, et al., 2003; Pritsker, et al., 2004; Robertson, et al., 2006; Xie, et al., 2005). As a result, hundreds of thousands of conserved non-coding elements (CNEs) have been predicted. Many of them are likely to be regulatory elements that play functional roles.

The large collection of CNEs constitutes a great resource for mining biological knowledge. Because many regulatory proteins bind to DNA in a sequence specific manner (Bulyk, 2003; Stormo, 2000), similarities in motif sequences discovered in promoters of different genes can be used to infer co-regulation of these genes (Terai, et al., 2001; van Nimwegen, et al., 2002). Given this, a natural follow up of the aforementioned large-scale motif discovery efforts is to identify similar groups of motifs in the database. This is a critical step in translating the conserved motifs identified by comparative genomics methods into putative model of regulatory elements (Bejerano, et al., 2005; Conlan, et al., 2005; Mwangi and Siggia, 2003; Robertson, et al., 2006; van Nimwegen, 2007; Xie, et al., 2005).

Motif clustering methods are typically distance-based (Hughes, et al., 2000; Pietrokovski, 1996). Such methods are conceptually simple, and are readily available as algorithms that are suitable for data sets of moderate size. However, applying such methods to mammalian genomes is technically challenging because genome-wide motif discovery result sets can contain hundreds of thousands of CNEs. As examples, the cisRED human and mouse databases (www.cisred.org) (Robertson, et al., 2006) contain between 200 and 250K computationally identified conserved genomic elements. Clustering such a large motif set requires a huge pairwise distance matrix which is extremely costly to compute and maintain. The distance-based approaches that can process such datasets (e.g. OPTICS (Mihael, et al., 1999)), can require a computer system with tens of gigabytes of memory.

Model-based clustering methods (Jensen and Liu, 2008; Jensen, et al., 2005; Qin, et al., 2003) assume that objects in a cluster share the same probability distribution. No explicit distance metric or matrix is involved, and, because inference is based on relationships between motifs and clusters rather than on relationships between motifs, such methods can be more scalable than distanced-based approaches in practice. More importantly, distance-based approaches treat the input motifs as if they were the *bona fide* patterns, thus ignoring uncertainty in the input. In contrast, model-based approaches explicitly model the uncertainty using probability distribution which is more effective in capturing highly-variable DNA regulatory elements using sequence data alone. We previously developed a Bayesian Motif Clustering method, BMC, which correctly identified experimentally reported prokaryotic regulons and suggested novel ones (Qin, et al., 2003).

In this study, we explored the scalability of model-based clustering to address genome scale sets of CNEs discovered in mammalian species. We developed an improved version of the BMC algorithm, BMC2, as well as a two-stage, divide-conquer-combine clustering scheme, MotifOrganizer. The new algorithm and scheme are better suited to mammalian applications than BMC because they allow clusters, as well as a motif and the cluster that it joins, to have different widths. The two-stage scheme adopted by MotifOrganizer can increase the scalability of BMC by three orders of magnitude under the same computation environment. Given this, MotifOrganizer is able to cluster collections of CNEs identified by whole genome, mammalian comparative genomics methods. The data structure and overall operation scheme of our methods are illustrated in Figure 1.

Results

We tested MotifOrganizer on three different datasets: a) 1128 motifs from 59 TF binding site (TFBS) models selected from the JASPAR 4 database (Vlieghe, et al., 2006), b) 5452 motifs from 319 TFBS models selected from the TRANSFAC 9.3 database (Matys, et al., 2006) and c) ~30,000 motifs from the cisRED human v.2 database. We used the first two motif sets to assess how accurately the model-based approach could recapitulate the partitioning in the input model sets, and to characterize the performance of BMC and BMC2 model-based clustering algorithms. With the third motif set, we compared the clustering result to known motif models from JASPAR and TRANSFAC databases. Since model-based clustering is not a deterministic procedure, results from different runs might be slightly different. Given this, we measured the performance of all model-based clustering algorithms by repeating the clustering procedure 100 times and taking the average. Clustering performance is measured by Adjusted Rand Index (ARI) (Milligan, 1986), more details about ARI can be found in the Method section and the Supplementary material.

1. JASPAR

We compiled a set of 1128 motifs that represented 59 mammalian TFBS models for which individual sequences were available from the JASPAR 4 database (Vlieghe, et al., 2006). Motif widths ranged from 5 to 22 bp, and the number of sequences in each model ranged from 3 to 48, averaging ~19. Clustering was performed using BMC2, the maximum allowable width difference between a motif and a cluster is set to be 2 bp (we used the same setting throughout this study). For comparison, we also tested an advanced distance-based clustering method CLARA (Kaufman and Rousseeuw, 1990). We used the edit distance similarity metric and determined the number of clusters by doing runs over a range of target cluster numbers and selecting the result set that had the maximum average silhouette widths as recommended by Kaufman and Rousseeuw 1990. We found that BMC2 achieved a higher ARI than CLARA (0.55 versus 0.44) and returned a number of clusters that was closer (54 vs. 34) to the number of input TFBS models--59. BMC2 also required less than a third of the memory, and was at least three times faster, depending on how many CLARA runs were used to identify the optimal solution,.

Next we took advantage of the quality measures produced by model-based clustering and performed a "cleanup" on the clustering result. To be specific, we used the normalized Bayes ratio (NBR), which reflects a cluster's tightness; and the posterior assignment probability (PAP), which measures how well a motif fits the cluster to which it has been assigned. Further details about these measures can be found in the Method section. Using filtering thresholds of NBR = 0 and PAP = 0.5, On average, 835 (74%) motifs were retained after cleanup, Using filtering thresholds of NBR = 0 and PAP = 0.5, the average ARI improved from 0.55 to 0.71, while under the same conditions, CLARA showed a more modest ARI increase, from 0.44 to 0.51. These results suggested that NBR and PAP are indeed effective quality indicators and can be used to prioritize clustering results for future experimental validation.

Figure 2 illustrates the clustering performance and relationships between TFBS models and structural classes. Its membership map shows two types of clustering errors:

‘combining’ errors (indicated by green cell), in which motifs that belong to different TFBS models were clustered together, and ‘splitting’ errors (indicated by blue cell), in which motifs belonging to same TFBS model were assigned to different clusters. Combining errors were ~2.5 times more frequent than splitting errors. It is also evident from the two plots that the result obtained after cleanup (right) indeed shows better clustering quality.

Similarities between binding models for structurally related TFs have been summarized by “familial binding profiles” (Mahony, et al., 2007; Sandelin and Wasserman, 2004; Suzuki, et al., 1994). Consistent with this, when BMC2’s clusters contained motifs from more than one model, the models typically belonged to the same structural class. For example, in the top 15 filtered clusters, which contained more than half of all motifs, three clusters contained motifs from different TF models. One of these clusters was mainly MA0101 (c-REL), MA0105 (p50) and MA01007 (p65) sequences from the REL structural class; another was mainly MA0040 (HFH-1), MA0041 (HFH-2) and MA0047 (HNF-3beta) sequences from the FORKHEAD structural class. At the same time, we expect that sequences for some models may be dispersed across clusters, given that a mixture model (Georgi and Schliep, 2006; Hannenhalli and Wang, 2005) or enhanced PWM (Gershenzon, et al., 2006) may better represent variability over binding sequences for a transcription factor than a single PWM.

2. TRANSFAC

The second dataset consisted of 5452 motifs from 319 TRANSFAC 9.3 mammalian TFBS models (Matys, et al., 2006). The number of motifs in each model ranged from 1 to 169, averaging 18.5. Motif widths ranged from 6 to 30 bps. Our primary goal was to compare performance between one-stage and two-stage model-based clustering approaches. Using MotifOrganizer, we first partitioned the whole motif set into disjoint motif subsets by width, then applied BMC2 in each subset separately, and concatenated all resulting clusters as input to the second stage clustering. More details about the two-stage scheme can be found in the Method section). Clustering performances measured by average ARI were plotted in Figure 3 and summarized in Supplementary Tables S2 and S3. Compared to the one-stage approach BMC, one can see that the two-stage approach yielded a higher average ARI (green lines are always on top of the corresponding red lines in Figure 3A), estimated the number of clusters more accurately (green lines are always closer to the x-axis than the corresponding red lines in Figure 3B). In addition, Figure 2 showed that cluster quality improved with increasing threshold values of the two quality measures--NBR and PAP. On the other hand, the memory consumption is about half when using the two-stage scheme, and the computing time of the two-stage clustering scheme is less than one third than that of the regular one-stage clustering approach. Running first stage clustering jobs in parallel will further reduce running time.

3. cisRED

We applied MotifOrganizer to a subset of 29,490 conserved DNA sequence motifs from the cisRED human v.2 database. These motifs had been identified using genome-wide comparative genomics approaches that involved combining results from multiple probabilistic *de novo* discovery methods (Robertson, et al., 2006). We selected the subset of motifs that had p-values < 0.001 and widths between 6 and 20 bp. Additional details about this dataset can be found in the Supplementary material. We applied MotifOrganizer to this dataset. The partition before the first stage is based on the motif width, such that each subset contains motifs of the same width which ranges from 6 to 20. At the end, a total of 8396 clusters were produced from MotifOrganizer. Among which, 4865 clusters consist of 15330 motifs passed the cleanup step with NBR cutoff of 0 and PAP cutoff of 0.5.

Unlike JASPAR or TRANSFAC, there is no “gold standard” partition of the cisRED motif set. In order to gauge the effectiveness of MotifOrganizer, we compared all 4865 filtered clusters to a large set of high-quality known motif models: 108 models from JASPAR CORE, and 398 models from TRANSFAC 9.3. We used two published comparison tools, MatCompare (Schones, et al., 2005) and MACO (Su, et al., 2006), to identify “match” between predicted motif clusters and those known motif models. For MatCompare, a “match” is called if the default distance measure, Kulback-Leibler (KL) distance between the PWMs of two motifs is less than 1. For MACO, we call two motifs “match” each other if their correlation coefficient score is less than 1. In the end, MatCompare identified 558 matches between predicted clusters and known motif models (one cluster may map to multiple motif models); MACO produced 753 matches. 171 matches were identified by both methods. Table 1 contains five such matches. The full list of all matches identified by both MatCompare and MACO, with motif logo plots (Schneider and Stephens, 1990), can be found at our website. Among all 4865 clusters identified by MotifOrganizer, 660 (14%) were similar to one of 506 known motif models using at least one method Among all the 506 known motif models, 294 (58.1%) matched to at least one predicted cluster.

We then compared our clusters with JASPAR PHYLOFACTS motifs. This database consists of 174 conserved motifs identified in a large scale mammalian comparative genomics study (Xie, et al., 2005) (see Supplementary Table S2 of this ref). MatCompare identified 167 matches and MACO identified 166. Seventy matches were identified by both methods. Table 1 contains five such matches. The full list of all matches identified by both MatCompare and MACO, with motif logo plots, can be found at our website. Of the 174 PHYLOFACT motifs, 108 (62.1%) matched at least one predicted cluster.

We are encouraged that many of the predicted motif clusters identified by MotifOrganizer match to known motif models. And more than half of the known motif models match to at least one of the clusters predicted. We also found that clusters with higher NBR are more likely to match to known motif models in JASPAR CORE, JASPAR PHYLOFACTS or TRANSFAC databases. Of the top 1000 clusters out of 4865 in total, 162 matched to known motif models according to MatCompare, while in clusters 1001 – 2000, 2001 – 3000, 3001 – 4000, 4001 – 4865, only 28, 25, 17 and 7 clusters matched to known motif models. MACO results were similar (data not shown). This suggests that clusters with higher quality scores were more likely to be *bona fide* functional elements. However, some highly ranked clusters, and overall more than 80% of all identified clusters, matched no motifs in JASPAR CORE, JASPAR

PHYLOFACTS or TRANSFAC databases. Table 2 contains motif patterns from 12 such clusters. Most positions in these motifs were highly conserved, and many of the motifs were palindromic, which is typical of homodimer DNA binding proteins. This suggests that these novel motifs may represent binding sites for uncharacterized TFs that mediate expression levels of the genes with which they are associated.

Discussion

Large amount of evolutionarily conserved DNA elements has been discovered with fast accumulation of sequenced genomes. Build on the hypothesis that sequence similarity implies binding by the same regulatory protein, clustering these elements--CNEs is able to translate putative elements into regulatory information. It has been shown that the ability of modeling uncertainties explicitly gave model-based clustering approaches advantages over distance-based approaches. However, existing model-based approaches such as BMC are unable to handle large scale motif sets collected from mammalian species. In this study, we proposed a novel two-stage model-based algorithm for clustering CNEs identified from mammalian species genome-wide. Our new algorithm allows motifs of different widths to be clustered together and is capable of handling large scale input motif set. Comparison studies indicated that our new approach retained and surpassed the clustering accuracy achieved by the single-stage model-based approaches, while reducing computation time and memory requirements to levels that permit clustering genome-wide sets of mammalian CNEs on commodity computer systems. To further demonstrate MotifOrganizer's scalability, we tested it on 150K motifs from the cisRED human v.2 database. Using default parameter settings, the entire two-stage clustering process took about four days to complete on a regular shared cluster computer server with 2.8GHz CPU nodes, and the peak memory consumption was only about 250MB. The scalability of MotifOrganizer demonstrated in this study is quite promising. As the number of *cis* regulatory regions may currently be underestimated (ENOCODE, 2007), we anticipate a persistent need for highly scalable clustering tools.

We were encouraged that many of the predicted motif clusters identified by MotifOrganizer were similar to known motif models that more than half of known motif models tested matched at least one of the predicted clusters, and that highly-ranked clusters were more likely to be similar to known motifs. Some of our filtered clusters appeared to differ from cluster patterns reported by previous large-scale studies, and some clusters with high NBR rank matched to no known motif pattern. These results suggest that regulatory motifs are highly diverse and that a substantial number of new regulatory elements have yet to be discovered and validated.

We seek to create a comprehensive catalog of mammalian *cis* regulatory motifs that, by facilitating dimension reduction and pattern discovery, and so functional annotation, will contribute to understanding modules and networks in mammalian transcriptional regulation. We anticipate improving MotifOrganizer's performance by extending it to include parameters that address more aspects of eukaryotic transcriptional regulation. For example, clustering may be more effective when it integrates additional data types like co-factors, DNA and chromatin structure, and histone modifications. A number of such data types, including mammalian TF binding regions, appear to be cost-effectively characterizable by ChIP-Seq technologies (Johnson, et al., 2007; Robertson, et al., 2007). As for ChIP-chip (Ren, et al., 2000) and other types of ChIP-sequencing (references 1-10 in Robertson et al. 2007), motifs can be identified in bound or enriched regions using only the target genome. However, approaches that seek to combine motifs from regions identified by ChIP-Seq with deep genome-wide comparative genomics methods are likely to continue to require scalable ways of identifying both conserved motifs and groups of similar motifs. We anticipate that MotifOrganizer and its extensions will serve as an important resource for such work. MotifOrganizer package written in C++ and Perl can be freely downloaded from http://www.sph.umich.edu/csg/qin/motif.

Methods

Input Data

The basic units in the input data for BMC2 or MotifOrganizer are CNEs or motifs, each of which is a stack of aligned short DNA sequences that had been identified by multiple, probabilistic, *de novo* comparative genomics motif discovery methods (D'Haeseleer, 2006; MacIsaac and Fraenkel, 2006). For example, cisRED motifs were identified in ~2kb promoter regions of sets of orthologous genes in, typically, 11 vertebrates species (Robertson, et al., 2006) (www.cisred.org). Each sequence in a discovered motif is assumed to be a phylogenetic counterpart of the other sequences. As a special case, we allowed a motif to contain as few as one sequence. These motifs are represented by position specific weight matrices (PWMs) in our model-based methods, which are used to calculate Bayes ratios for iterative cluster assignments. See Figure 1 for an illustration of motifs and motif clusters.

BMC2

BMC (Qin, et al., 2003) assumes that motifs that belong to a cluster follow the same product multinomial distribution (Chen and Liu, 1996), and implements a Gibbs sampler procedure to iteratively infer cluster membership (Gelfand and Smith, 1990; Liu, 2001). By using only the middle part of each motif, the original algorithm allows the width of input motifs to be different, but requires that all clusters have the same width. Because the widths of mammalian TFBSs vary substantially (from 6 bp to >30 bp in JASPAR and TRANSFAC), such a constraint is too restrictive for clustering mammalian CNEs. In a recent paper, Jensen and Liu proposed an extended Bayesian model that treats cluster widths as random variables (Jensen and Liu, 2008). We adopted an alternative strategy in BMC2 by allowing motifs and clusters with different widths to be grouped together. Our strategy allows flexible alignment between motifs and clusters. Specifically, during the process that reassigns motifs to clusters, if a cluster's width is larger than a motif's width, we use a sliding window whose width is equal to the motif width to determine which part of the cluster pattern best fits the motif, and use this best match to calculate the likelihood that the motif will join the cluster. Conversely, when the motif is wider than the cluster, we use a sliding window with a width equal to the cluster width to determine which part of the motif best fits the cluster, and use only this subset of the motif to calculate the fit likelihood for this cluster. After a cluster has been identified as accommodating a motif, if the cluster width is smaller than the motif width, then only the aligned part of the motif is added to the cluster, and the rest of the motif is removed from consideration. Conversely, if the cluster width is greater than the motif width, then the aligned cluster pattern will be trimmed to the motif width, and only the part that aligned to the added motif is kept. Such a strategy allows us to generate clusters of different widths, and is able to group motifs of different widths into a common cluster.

MotifOrganizer

Because genome wide collections of mammalian CNE contain too many motifs for BMC to cluster, we sought to increase BMC's scalability while retaining its favorable performance. To accomplish this, we devised a novel, two-stage, divide-conquer-combine scheme (Figure 1). For input to the first stage, we partitioned the motif set into subsets, either randomly or by assigning motifs of different widths to different subsets.

After partitioning, BMC2 algorithm is applied independently to each subset. We then combined all output clusters from this stage to form a new input motif set, in which each input motif was a first-stage cluster. Since both motifs and clusters are represented by PWMs, the BMC2 algorithm could again be applied to group these first-stage clusters into final clusters. Because the number of clusters output by the first stage clustering is typically much smaller than the total number of original motifs, the overall demand on computer resources is substantially reduced, making the two-stage strategy highly scalable. Since clustering runs in the first stage can be carried out independently, it is straightforward to implement MotifOrganizer to take advantage of the increasingly available parallel computing environment, which will further reduce the computation time. The details about BMC2 algorithm and the two-stage MotifOrganizer clustering procedure can be found in the Supplementary material.

Quality measure

Model-based clustering methods allow formal statistical inference to be performed. BMC2 and MotifOrganizer take advantage of this to produce a set of significance measures. One measure is the ratio of the probability that all motifs follow the same product multinomial distribution versus the probability that each motif follow its own. This serves as a measure of the cluster's compactness. A higher ratio indicates that individual motif sequences in a cluster more closely resemble each other. In BMC2 and MotifOrganizer, since the widths of motifs may differ between clusters, we define normalized Bayes ratio (NBR) for a cluster by dividing *B* by the width of the cluster. BMC also calculates and reports the posterior assignment probability (PAP) for each motif in a cluster. This is the probability that a motif belongs to its current cluster, conditional on the assignments of all other motifs. A higher PAP indicates a better motif-to-cluster fit. These quality measurements can be used to prioritize clustering results. By sorting clusters from the most significant to the least significant according to NBR, one can focus on the top ones for experimental validation. One can also remove loosely fit motifs from its cluster if its PAP is lower than a pre-specified threshold.

Clustering accuracy

To evaluate and compare the performance of different clustering approaches, we need to quantify the agreement between different clustering results. Among many statistics proposed, we adopted the Adjusted Rand Index (ARI) (Hubert, 1985; Milligan, 1986). Its values lie between 0 and 1, with a higher value indicating better agreement. Milligan and Cooper (1986) recommended it as the measure of agreement based on extensive empirical studies. The detailed formula on how to calculate ARI can be found in the Supplementary material.

Motif matching tools

To annotate the clustering results, we compared each cluster to motif models in JASPAR 4 (Vlieghe, et al., 2006) and TRANSFAC 9.3 (Matys, et al., 2006) databases. We used MatCompare (Schones, et al., 2005) and MACO (Su, et al., 2006) to assess the similarity of motif pairs. For MatCompare, we used the default distance measure, which is the minimum Kulback-Leibler (KL) divergence between matched fragments in two motifs; motifs with divergence per column less than 1.0 are regarded as very similar. For MACO, we determined the threshold value for the matching correlation coefficient scores based on the empirical distribution obtained from 3000 random matrices.

Acknowledgements

We thank Dr. Dustin Schones for assistance with MatCompare program. And we thank the cisRED team at BCGSC for their help with the cisRED dataset.

Figure legends

Figure 1. A. Illustration of phylogenetic footprinting technique and the motif clustering strategy, as well as examples of motifs and motif clusters. **B.** A diagram of the two-stage divide-conquer-combine scheme MotifOrganizer proposed to enable model-based clustering to be performed on a large motif collection.

Figure 2. A. Membership map that summarizes BMC2 clustering performance on 1152 motifs from 59 TFBS models from the JASPAR 4 database. Rows represent TFBS models in JASPAR, columns represent clusters generated from BMC2. For each cell (*i*,*j*), a) red indicates cases in which motifs *i* and *j* were clustered together by BMC2 and belonged to the same JASPAR model, b) green indicates cases in which motifs *i* and *j* were clustered together by BMC2 but belonged to different JASPAR models, c) blue indicates cases in which motifs *i* and *j* were not clustered together by BMC2 but belonged to the same JASPAR model, and d) white indicates cases in which motifs *i* and *j* were neither in the same BMC2 cluster nor belonged to the same JASPAR model . **A1.** Membership map for all 1152 motifs in the input dataset. **A2.** Membership map for remaining motifs after filtering with thresholds NBR=0 and PAP=0.5.

B. Membership map that summarizes BMC2 clustering performance for 1152 motifs from 59 TFBS models from the JASPAR 4 database. For each JASPAR model, A, and each BMC2 cluster, X, we define a product of two ratios:

$$R = \frac{\text{number of motifs in both A and X}}{\text{total number of motifs in A}} \times \frac{\text{number of motifs in both A and X}}{\text{total number of motifs in X}}$$

An ideal cluster (R=1), contains motifs from only one model. Values of R<1 indicate clusters that contain motifs from several models, or motifs from one model that are dispersed across more than one clusters, or both. **B1.** Membership map for all 1152 motifs. Each row represents one of 59 JASPAR TFBS models, and each column corresponds to one of 56 BMC2 clusters. Red indicates $R = 1$ and light blue $R < 1$. **B2.** Details of the upper left corner of the membership map, showing high quality ($R \approx 1$) clusters.

Figure 3. MotifOrganizer clustering results for TRANSFAC 9.3 data. This dataset contains 5452 motifs that belonged to 295 TRANSFAC TFBS models. To evaluate using quality measures to filter clustering results, we tested nine different threshold settings, combining NBR thresholds of 0, 2, and 5 for clusters and PAP thresholds of 0, 0.5 and 0.8 for cluster assignments (solid, dashed and dotted lines respectively). Red lines correspond to one-stage BMC2 clustering and green lines to two-stage motifOrganizer clustering. **A.** Clustering performance measured by ARI; **B.** Differences between numbers of clusters and number of TRANSFAC models in the remaining motif sets after filtering. See Supplementary Tables S2 and S3 for numerical values.

Figures

Figure 1.

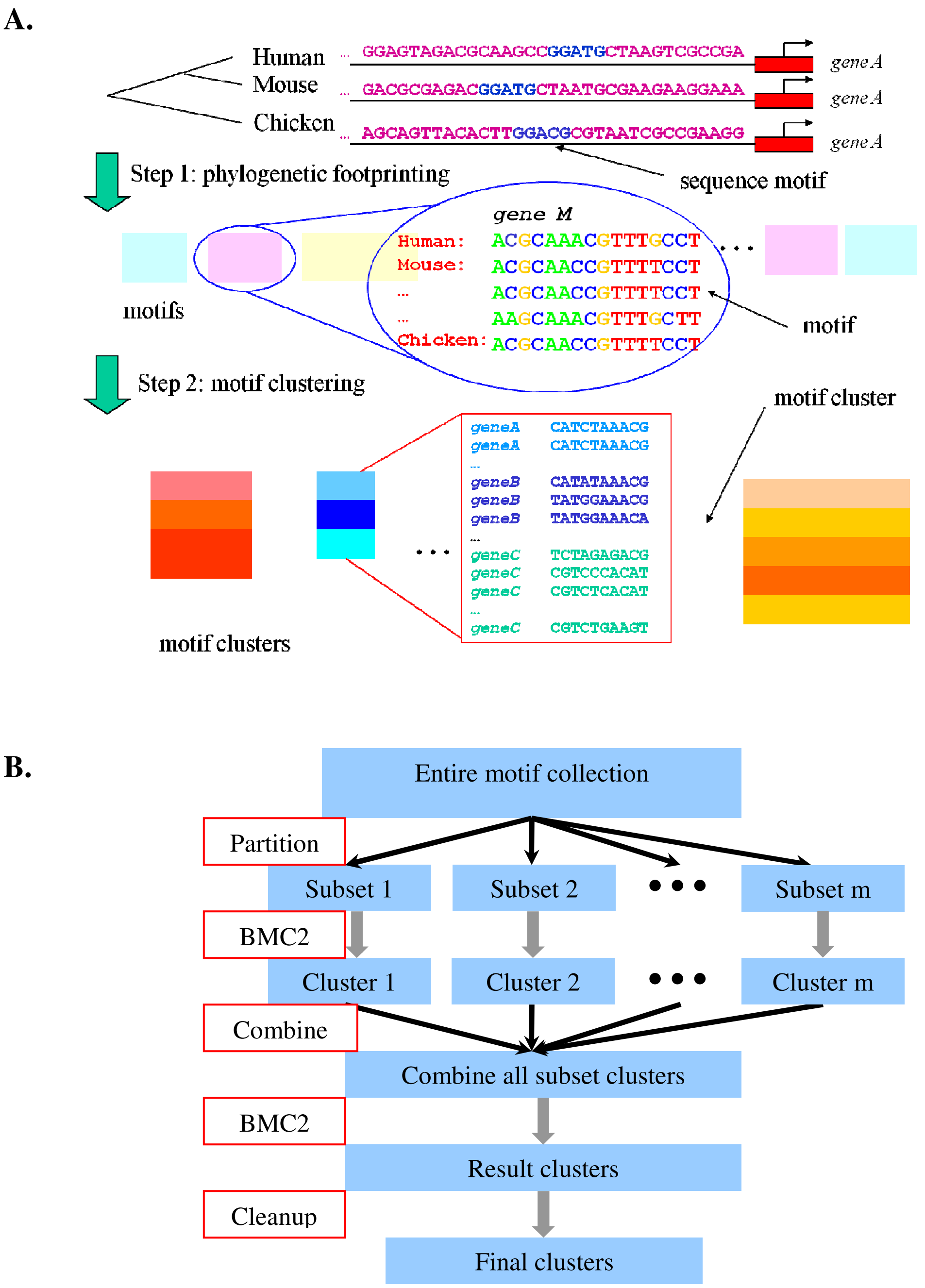


Figure 2.

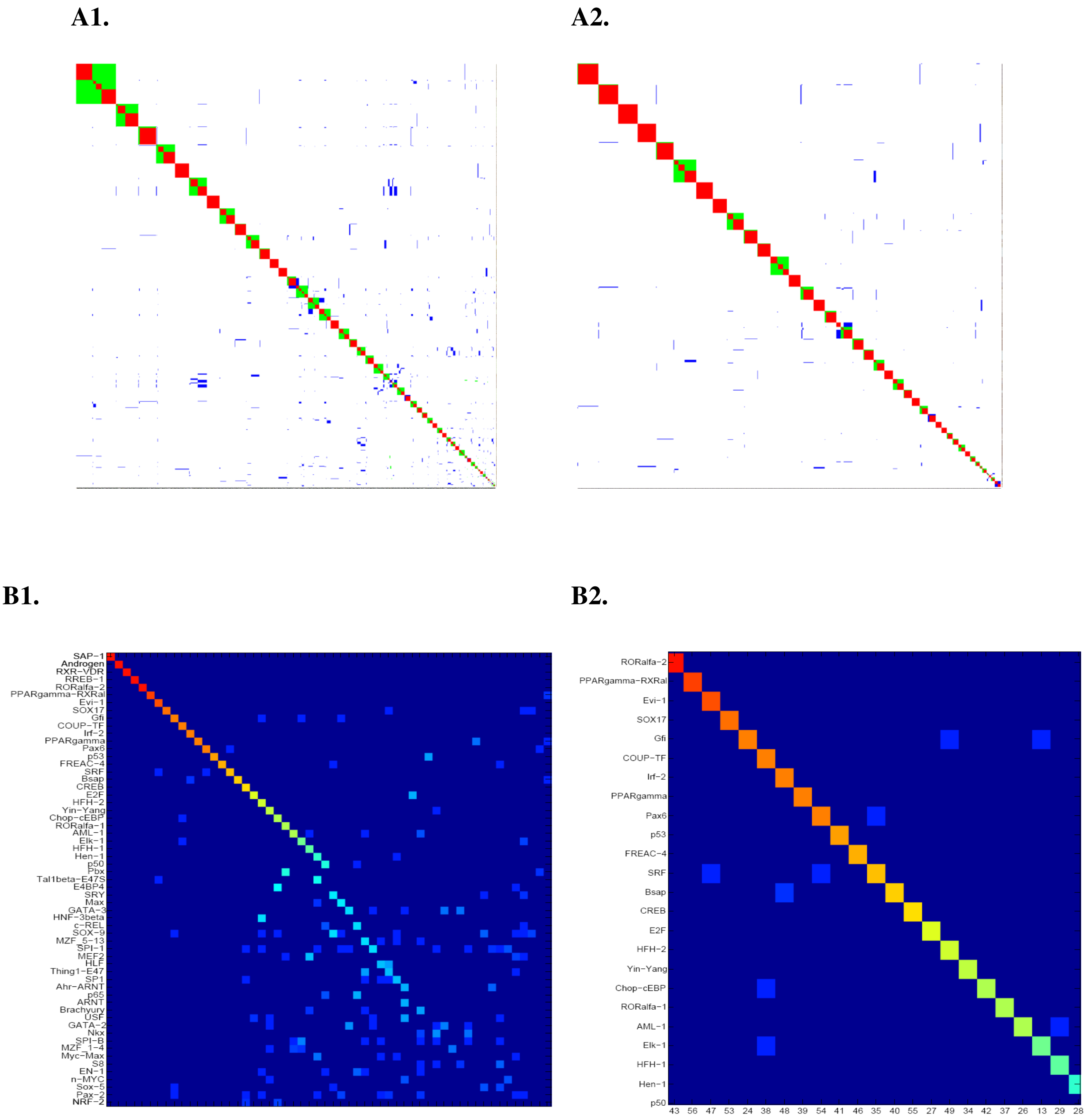
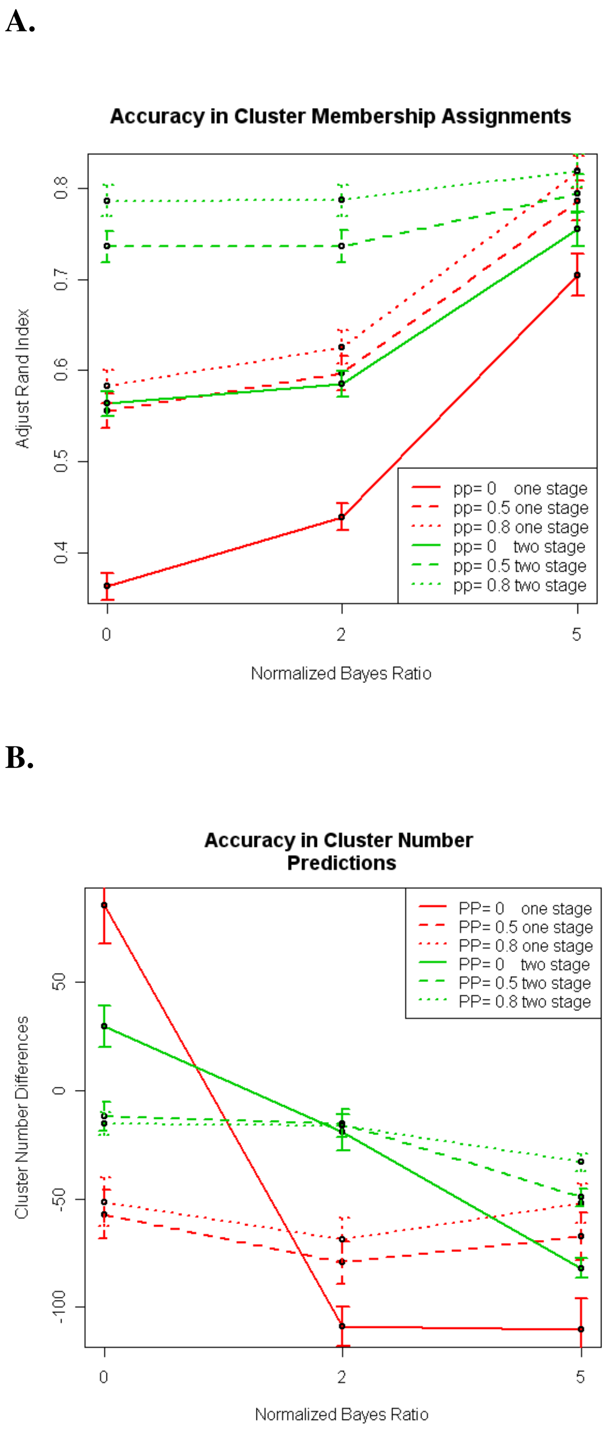

A1.
A2.
B1.
SAP−1
Androgen
RXR−VDR
RREB−1
RORalfa−2
PPARgamma−RXRal
Evi−1
SOX17
Gfi
COUP−TF
Irf−2
PPARgamma
Pax6
p53
FREAC−4
SRF
Bsap
CREB
E2F
HFH−2
Yin−Yang
Chop−cEBP
RORalfa−1
AML−1
Elk−1
HFH−1
Hen−1
p50
Pbx
Tal1beta−E47S
E4BP4
SRY
Max
GATA−3
HNF−3beta
c−REL
SOX−9
MZF_5−13
SPI−1
MEF2
HLF
Thing1−E47
SP1
Ahr−ARNT
p65
ARNT
Brachyury
USF
GATA−2
Nkx
SPI−B
MZF_1−4
Myc−Max
S8
EN−1
n−MYC
Sox−5
Pax−2
NRF−2
B2.
RORalfa−2
PPARgamma−RXRal
Evi−1
SOX17
Gfi
COUP−TF
Irf−2
PPARgamma
Pax6
p53
FREAC−4
SRF
Bsap
CREB
E2F
HFH−2
Yin−Yang
Chop−cEBP
RORalfa−1
AML−1
Elk−1
HFH−1
Hen−1
p50
43 56 47 53 24 38 48 39 54 41 46 35 40 55 27 49 34 42 37 26 13 29 28

Figure 3.


A.
Accuracy in Cluster Membership Assignments
Adjust Rand Index
0.4
0.5
0.6
0.7
0.8
0
2
5
Normalized Bayes Ratio
pp= 0 one stage
pp= 0.5 one stage
pp= 0.8 one stage
pp= 0 two stage
pp= 0.5 two stage
pp= 0.8 two stage
B.
Accuracy in Cluster Number Predictions
Cluster Number Differences
-100
-50
0
50
0
2
5
Normalized Bayes Ratio
PP= 0 one stage
PP= 0.5 one stage
PP= 0.8 one stage
PP= 0 two stage
PP= 0.5 two stage
PP= 0.8 two stage

Tables

Table 1. The top five rows contain selected cisRED motif clusters that were reported by both MatCompare and MACO as similar to at least one JASPAR CORE or TRANSFAC TFBS models. The bottom five rows contain selected cisRED motif clusters that were reported by both MatCompare and MACO as similar to at least one JASPAR PHYLOFACTS motifs. The clusters are sorted by descending Normalized Bayes Ratio (NBR).

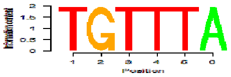
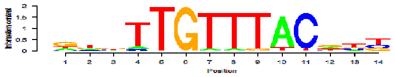
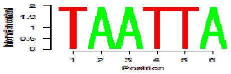
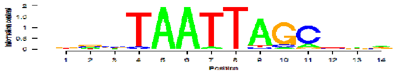
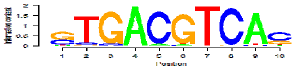
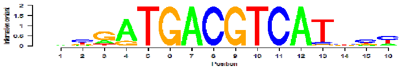
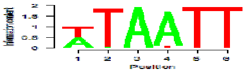
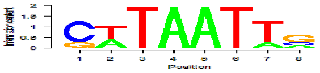
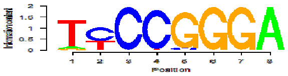
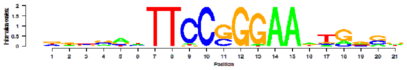
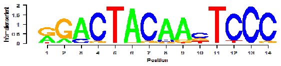
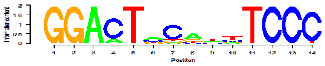
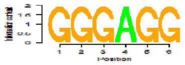
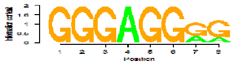
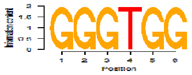
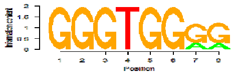
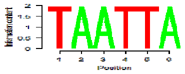
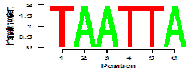
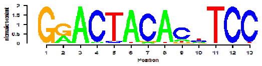
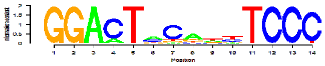
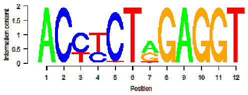
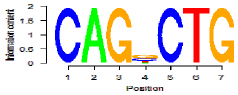
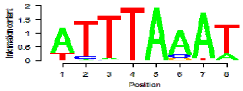
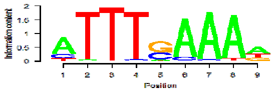
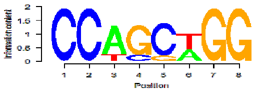
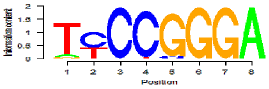
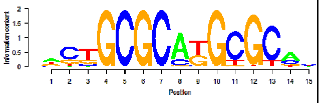
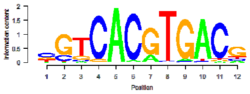
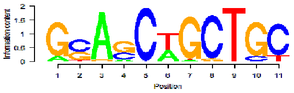
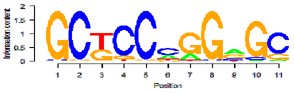
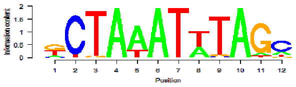
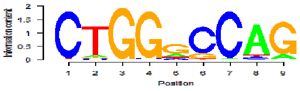
Cluster ID	Motif cluster logo	Bayes ratio	Transfac/ Jaspar ID	TFBS profile logo	Mat-Compare divergen	MACO score
JASPAR CORE + TRANSFAC						
38		25.7	M00473 FOXO1		0.095	0.998
49		24.7	M00437 CHX10		0.072	0.999
95		20.9	M00179 CRE-BP1		0.542	0.942
157		17.9	MA0063 Nkx2-5		0.148	0.991
159		17.7	MA0137 STAT1		0.580	0.923
JASPAR PHYLOFACTS						
2		36.9	PF0074		0.749	0.774
5		33.4	PF0024		0	1
15		29.8	PF0056		0	1
49		24.7	PF0023		0	1
101		20.5	PF0074		0.631	0.887

Table 2. Selected motif clusters from the result of clustering the 30K motif cisRED Human v.2 dataset using motifOrganizer. The motif pattern of these clusters does not match to any JASPAR CORE, PHYLOFACTS or TRANSFAC motif model according to both MatCompare and MACO software. The clusters are sorted by normalized Bayes ratio (NBR).

Cluster ID	Motif cluster logo	NBR	PHYLOFACTS model logo
71		22.9	DUSP8, NFKBIA
78		22.4	GGTLA1, DBH
83		22.1	C14orf103, NP_060229.2
94		20.9	MID1, CLDN17, VRK1
151		18.1	MC1R, DBH, COL2A1
159		17.7	NP_079167.1, C9orf82, HNRPR
187		16.3	SAFB, MYBBP1A, SDCCAG8, EIF2S2,ZC3H8
196		16.0	HSF2, RAB5A, LMTK2, BRDT, LATS2, OSBPL8, PHF20L1
218		15.3	Q8NBE0_HUMAN, BRUNOL4
228		15.1	GRM5, WNT3A, FAM49A, ANKRD5, AP3S2, HOXD11
232		14.9	TTN, PICALM, ESR1, KBTBD5
238		14.7	Q6ZT51_HUMAN, ZNF385, ARHGAP23, CNNM1

References

Bejerano, G., Siepel, A.C., Kent, W.J. and Haussler, D. (2005) Computational screening of conserved genomic DNA in search of functional noncoding elements, *Nature methods*, **2**, 535-545.

Bulyk, M.L. (2003) Computational prediction of transcription-factor binding site locations, *Genome biology*, **5**, 201.

Chen, R. and Liu, J.S. (1996) Predictive Updating Methods With Application to Bayesian Classification, *Journal of the Royal Statistical Society Series B-Methodological*, **58**, 397-415.

Conlan, S., Lawrence, C. and McCue, L.A. (2005) Rhodopseudomonas palustris regulons detected by cross-species analysis of alphaproteobacterial genomes, *Appl Environ Microbiol*, **71**, 7442-7452.

D'Haeseleer, P. (2006) How does DNA sequence motif discovery work?, *Nature biotechnology*, **24**, 959-961.

Elemento, O. and Tavazoie, S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach, *Genome biology*, **6**, R18.

ENOCODE (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, **447**, 799-816.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.

Georgi, B. and Schliep, A. (2006) Context-specific independence mixture modeling for positional weight matrices, *Bioinformatics (Oxford, England)*, **22**, e166-173.

Gershenzon, N.I., Trifonov, E.N. and Ioshikhes, I.P. (2006) The features of Drosophila core promoters revealed by statistical analysis, *BMC genomics*, **7**, 161.

Hannenhalli, S. and Wang, L.S. (2005) Enhanced position weight matrices using mixture models, *Bioinformatics (Oxford, England)*, **21 Suppl 1**, i204-212.

Hubert, L., Arable, P. (1985) Comparing partitions, *Journal of Classification*, **2**, 193-218.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae, *Journal of molecular biology*, **296**, 1205-1214.

Jensen, S.T. and Liu, J.S. (2008) Bayesian Clustering of Transcription Factor Binding Motifs, *Journal of American Statistical Association*, **103**, 188-200.

Jensen, S.T., Shen, L. and Liu, J.S. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes, *Bioinformatics (Oxford, England)*, **21**, 3832-3839.

Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions, *Science (New York, N.Y*, **316**, 1497-1502.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding groups in data : an introduction to cluster analysis*. Wiley, New York.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature*, **423**, 241-254.

Liu, J. (2001) *Monte Carlo Strategies in scientific computing*. Springer-Verlag, New York.

MacIsaac, K.D. and Fraenkel, E. (2006) Practical strategies for discovering regulatory DNA sequence motifs, *PLoS computational biology*, **2**, e36.

Mahony, S., Auron, P.E. and Benos, P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies, *PLoS computational biology*, **3**, e61.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E. and Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic acids research*, **34**, D108-110.

Mihael, A., Markus, M.B., Hans-Peter, K., J and rg, S. (1999) OPTICS: ordering points to identify the clustering structure. ACM, 49-60.

Milligan, G.W., Cooper, M. C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavior Research*, **21**, 441-458.

Mwangi, M.M. and Siggia, E.D. (2003) Genome wide identification of regulatory motifs in Bacillus subtilis, *BMC bioinformatics*, **4**, 18.

Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments, *Nucleic acids research*, **24**, 3836-3845.

Pritsker, M., Liu, Y.C., Beer, M.A. and Tavazoie, S. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation, *Genome research*, **14**, 99-108.

Qin, Z.S., McCue, L.A., Thompson, W., Mayerhofer, L., Lawrence, C.E. and Liu, J.S. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites, *Nature biotechnology*, **21**, 435-439.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins, *Science (New York, N.Y*, **290**, 2306-2309.

Robertson, G., Bilenky, M., Lin, K., He, A., Yuen, W., Dagpinar, M., Varhol, R., Teague, K., Griffith, O.L., Zhang, X., Pan, Y., Hassel, M., Sleumer, M.C., Pan, W., Pleasance, E.D., Chuang, M., Hao, H., Li, Y.Y., Robertson, N., Fjell, C., Li, B., Montgomery, S.B., Astakhova, T., Zhou, J., Sander, J., Siddiqui, A.S. and Jones, S.J. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements, *Nucleic acids research*, **34**, D68-73.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O.L., He, A., Marra, M., Snyder, M. and Jones, S. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nature methods*, **4**, 651-657.

Sandelin, A. and Wasserman, W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics, *Journal of molecular biology*, **338**, 207-215.

Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences, *Nucleic acids research*, **18**, 6097-6100.

Schones, D.E., Sumazin, P. and Zhang, M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites, *Bioinformatics (Oxford, England)*, **21**, 307-313.

Stone, E.A., Cooper, G.M. and Sidow, A. (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics, *Annual review of genomics and human genetics*, **6**, 143-164.

Stormo, G.D. (2000) DNA binding sites: representation and discovery, *Bioinformatics (Oxford, England)*, **16**, 16-23.

Su, G., Mao, B. and Wang, J. (2006) MACO: a gapped-alignment scoring tool for comparing transcription factor binding sites, *In Silico Biol*, **6**, 307-310.

Suzuki, M., Gerstein, M. and Yagi, N. (1994) Stereochemical basis of DNA recognition by Zn fingers, *Nucleic acids research*, **22**, 3397-3405.

Terai, G., Takagi, T. and Nakai, K. (2001) Prediction of co-regulated genes in Bacillus subtilis on the basis of upstream elements conserved across three closely related species, *Genome biology*, **2**, RESEARCH0048.

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. (2005) Assessing computational tools for the discovery of transcription factor binding sites, *Nature biotechnology*, **23**, 137-144.

van Nimwegen, E. (2007) Finding regulatory elements and regulatory motifs: a general probabilistic framework, *BMC bioinformatics*, **8 Suppl 6**, S4.

van Nimwegen, E., Zavolan, M., Rajewsky, N. and Siggia, E.D. (2002) Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7323-7328.

Vlieghe, D., Sandelin, A., De Bleser, P.J., Vleminckx, K., Wasserman, W.W., van Roy, F. and Lenhard, B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles, *Nucleic acids research*, **34**, D95-97.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals, *Nature*, **434**, 338-345.