

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

Zhaohui Qin^{1*}, Gordon Robertson², Misha Bilenky², Gang Su³, 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;

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
PAM:	Partition Around Medoids
ARI:	Adjusted Rand Index

When conserved non-coding elements (CNEs) and DNA sequence motifs are identified using the phylogenetic conservation and sequence specificity that are often associated with transcription factor binding sites¹⁻³, recognizing groups of similar motifs is a critical step in translating CNEs into putative regulatory information⁴. Model-based methods explicitly address uncertainty in clustering motifs^{5,6}; and, because they use no distance metric or 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 extended BMC, a Bayesian clustering method that previously had correctly identified prokaryote regulons, to allow motifs and motif clusters to have different widths. We then implemented the new algorithm in highly scalable two-stage clustering approach, motifOrganizer. Test showed favorable performance on TRANSFAC⁷ and JASPAR⁸ motifs, and on ~30k and then ~150k CNEs from the cisRED human v.2 database¹, indicating that MotifOrganizer can cluster whole genome sets of mammalian comparative genomics CNEs.

Identifying similar groups of motifs is a critical step in translating the conserved motifs identified by comparative genomics methods into putative model of regulatory elements^{1,2,9-11}. Motif clustering methods are typically distance-based^{12,13}. 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)¹ contain between 200 and 250K computationally identified conserved

genomic elements. The distance-based approaches that can process such datasets (e.g. OPTICS¹⁴, can require a computer system with tens of gigabytes of memory.

Model-based clustering methods^{5,6} 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. We previously developed a Bayesian Motif Clustering method, BMC, which correctly identified experimentally reported prokaryotic regulons and suggested novel ones.⁶

In this study, we explored the scalability of model-based clustering to address genome scale sets of conserved mammalian DNA sequence motifs. 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 approach 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, and MotifOrganizer can analyze motif datasets that are three orders of magnitude larger than BMC can handle. Given this, MotifOrganizer can use common computing systems to cluster collections of CNEs identified by whole genome, mammalian comparative genomics methods.

Results

We tested MotifOrganizer on three different datasets: a) ~1000 motifs from 59 TFBS models selected from the JASPAR 4 database⁷, b) ~5000 motifs from 319 TFBS models selected from the TRANSFAC 9.3 database⁸ and c) ~30,000 motifs from the cisRED human v.2 database. We used the first two tests 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 by tracking the average Adjusted Rand Index¹⁵ (ARI, Supplementary material) over 100 clustering runs. In the third dataset, we compared the clustering result to known motifs from TRANSFAC and JASPAR databases to assess the accuracy of the original and new algorithm.

1. JASPAR

We compiled a set of 1128 motifs that represented 59 mammalian TF binding models for which individual sequences were available from the JASPAR 4 database⁷ (Supplementary material). Motif widths ranged from 5 to 22 bp, and the number of sequences in each model ranged from 3 to 48, averaging ~19. We compared clustering the motifs with BMC2 and with the distance-based CLARA¹⁶. For CLARA, we used an edit distance similarity metric, and we 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 width. BMC2 achieved a higher ARI than CLARA (0.55 versus 0.44) and returned a number of clusters (54 vs. 34) that was closer to the number of input models. 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 an optimal solution.

We then assessed filtering BMC2's output clusters with two model-based cluster quality measures: a normalized Bayes ratio (NBR), which reflects a cluster's tightness; and a posterior assignment probability (PAP), which measures how well a motif fits the cluster to which it has been assigned (Supplementary material). Using filtering thresholds of $\text{NBR} = 0$ and $\text{PAP} = 0.5$, the average ARI improved from 0.55 to 0.71 over 100 BMC2 runs, while after 100 repeats CLARA showed a more modest ARI increase, from 0.44 to 0.51. These results suggested that motifs that fit clusters poorly can be removed from a clustering result set and that NBR can be used to prioritize clustering results for future experimental validation.

Figure 1 illustrates the clustering performance and relationships between motif models and structural classes. Its membership map shows two types of clustering errors: 'combining' errors, in which motifs that belong to different TFBS models were clustered together, and 'splitting' errors, in which motifs belonging to same TFBS model were assigned to different clusters. Combining errors were ~2.5 times more frequent than splitting errors.

Similarities between binding models for structurally related TFs have been summarized by 'familial binding profiles'¹⁷⁻¹⁹. Consistent with this, when BMC2's clusters contained motifs from more than one model, the models typically belonged to the same structural class (Supplementary materials). 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^{20,21} or enhanced PWM²² 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 (Supplementary material). The number of motifs in each model ranged from 1 to 169, averaging 18.5. Motif widths ranged from 6 to 30 bps. 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 (see Experimental Protocol). Compared to the one-stage approach BMC, the two-stage approach handled more motifs, yielded a higher average ARI (Fig. 2A and Supplementary Tables S2 and S3), estimated the number of clusters more accurately (Fig. 2B), completed runs in less than one third of the time and used about half of the memory. When we filtered clustering results using three thresholds for PAP (0, 0.5, 0.8) and NBR (0, 2, 5), cluster quality improved with increasing quality threshold values.

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 methods¹ (Supplementary material). We selected the subset of motifs that had p-values < 0.001 and widths between 6 and 20 bp. Applying the two-stage MotifOrganizer resulted in 8396 clusters, which we reduced to 4865 clusters that contained 15,330 motifs by filtering with thresholds of NBR = 0 and PAP = 0.5.

We compared the filtered clusters to 108 models from JASPAR CORE, and 398 models from TRANSFAC 9.3, using MatCompare²³ with PWMs and MACO²⁴ with sequences. We allowed a cluster to be similar to more than one JASPAR or TRANSFAC motif patterns. MatCompare identified 239 (5%) clusters as similar to at least one of the 506 known models, while MACO identified 510 (10%) as similar; 660 (14%) clusters were similar to a known motif using at least one method (Table 1 and Supplementary Material). Fifty eight percent of the known models matched at least one predicted cluster.

We then compared our clusters with JASPAR phyloFACTS motifs, which consists of 174 conserved motifs identified in a large scale mammalian comparative genomics study² (see Supplementary Table S2 of this ref). MatCompare identified 167 matches and MACO identified 166. Seventy matches were identified by both methods (Table 1 and Supplementary Material). Of the 174 phyloFACT motifs, 108 (62.1%) matched at least one predicted cluster. Overall, the results suggest that combining cisRED probabilistic motifs with MotifOrganizer's model-based motif clustering yields a genome-wide set of clusters that partially overlaps with but is largely distinct from clusters identified by the methods used in the other two studies. More information about the clustering result can be found at our website (<http://www.sph.umich.edu/csg/qin/motif>).

We also found that clusters with higher NBR were more likely to be similar to known motifs; of the top 1000 clusters, 162 matched to known motifs, while only 28, 25, 17 and 7 clusters matched in subsequent lower-ranked 1000 gene groups. MACO results were consistent with this (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 S4 in the Supplementary Material 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

MotifOrganizer's two-stage approach retained the clustering accuracy of the single-stage BMC, 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 takes 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 underestimated²⁵, 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 TFBS models, that more than half of known 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^{26,27}. As for ChIP-chip²⁸ and other types of ChIP-sequencing (references 1-10 in Robertson et al. 2007²⁷), 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. Computer programs and scripts written in C++ and Perl can be freely downloaded from <http://www.sph.umich.edu/csg/qin/motif>.

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.

Experimental Protocol

The basic units in the cisRED input data were CNEs or motifs, each of which was a stack of aligned short DNA sequences that had been identified by multiple, probabilistic, *de novo* comparative genomics motif discovery methods²⁹. cisRED motifs were identified in ~2kb promoter regions of sets of orthologous genes in, typically, 11 vertebrates species¹ (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.

BMC assumes that motifs that belong to a cluster follow the same product multinomial distribution³⁰, and implements a Gibbs sampler procedure to iteratively infer cluster membership^{31,32}. 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 variables³³. 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. Such a strategy allows us to generate clusters of different widths, and is able to group motifs of different widths in one cluster.

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. We devised a novel, two-stage, divide-conquer-combine approach. 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. We then applied the BMC2 algorithm 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 motif 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, and hence the two-stage strategy is 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.

To annotate the clustering results, we compared each cluster to models in JASPAR 4⁷ and TRANSFAC 9.3⁸ databases. We used MatCompare²³ and MACO²⁴ 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 scores from the empirical distribution function.

References:

- 1 Robertson G, Bilenky M, Lin K *et al*: cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* 2006; **34**: D68-73.
- 2 Xie X, Lu J, Kulbokas EJ *et al*: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 2005; **434**: 338-345.
- 3 McCue L, Thompson W, Carmack C *et al*: Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* 2001; **29**: 774-782.
- 4 van Nimwegen E: Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics* 2007; **8 Suppl 6**: S4.
- 5 Jensen ST, Shen L, Liu JS: Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* 2005; **21**: 3832-3839.
- 6 Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS: Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* 2003; **21**: 435-439.
- 7 Vlieghe D, Sandelin A, De Bleser PJ *et al*: A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 2006; **34**: D95-97.
- 8 Matys V, Kel-Margoulis OV, Fricke E *et al*: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006; **34**: D108-110.
- 9 Mwangi MM, Siggia ED: Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics* 2003; **4**: 18.
- 10 Bejerano G, Siepel AC, Kent WJ, Haussler D: Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat Methods* 2005; **2**: 535-545.
- 11 Conlan S, Lawrence C, McCue LA: Rhodospseudomonas palustris regulons detected by cross-species analysis of alphaproteobacterial genomes. *Appl Environ Microbiol* 2005; **71**: 7442-7452.
- 12 Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 2000; **296**: 1205-1214.
- 13 Pietrokovski S: Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 1996; **24**: 3836-3845.
- 14 Mihael A, Markus MB, Hans-Peter K, J, rg S. OPTICS: ordering points to identify the clustering structure, ACM, 1999, vol 28, pp 49-60.
- 15 Milligan GW, Cooper, M. C. : A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavior Research* 1986; **21**: 441-458.
- 16 Kaufman L, Rousseeuw PJ: Finding groups in data : an introduction to cluster analysis. New York, Wiley, 1990.
- 17 Suzuki M, Gerstein M, Yagi N: Stereochemical basis of DNA recognition by Zn fingers. *Nucleic Acids Res* 1994; **22**: 3397-3405.
- 18 Sandelin A, Wasserman WW: Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* 2004; **338**: 207-215.
- 19 Mahony S, Auron PE, Benos PV: DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol* 2007; **3**: e61.
- 20 Georgi B, Schliep A: Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics* 2006; **22**: e166-173.
- 21 Hannehalli S, Wang LS: Enhanced position weight matrices using mixture models. *Bioinformatics* 2005; **21 Suppl 1**: i204-212.
- 22 Gershenzon NI, Trifonov EN, Ioshikhes IP: The features of *Drosophila* core promoters revealed by statistical analysis. *BMC Genomics* 2006; **7**: 161.
- 23 Schones DE, Sumazin P, Zhang MQ: Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* 2005; **21**: 307-313.
- 24 Su G, Mao B, Wang J: MACO: a gapped-alignment scoring tool for comparing transcription factor binding sites. *In Silico Biol* 2006; **6**: 307-310.
- 25 ENCODE: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**: 799-816.
- 26 Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007; **316**: 1497-1502.

- 27 Robertson G, Hirst M, Bainbridge M *et al*: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007; **4**: 651-657.
- 28 Ren B, Robert F, Wyrick JJ *et al*: Genome-wide location and function of DNA binding proteins. *Science* 2000; **290**: 2306-2309.
- 29 D'Haeseleer P: How does DNA sequence motif discovery work? *Nat Biotechnol* 2006; **24**: 959-961.
- 30 Chen R, Liu JS: Predictive Updating Methods With Application to Bayesian Classification. *Journal of the Royal Statistical Society Series B-Methodological* 1996; **58**: 397-415.
- 31 Gelfand AE, Smith AFM: Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**: 398-409.
- 32 Liu J: Monte Carlo Strategies in scientific computing. New York, Springer-Verlag, 2001.
- 33 Jensen STL, J. S.: Bayesian Clustering of Transcription Factor Binding Motifs. *Journal of American Statistical Association* 2008; **To appear**.

Tables and figures

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 model. 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 cluster motif. The clusters are sorted by descending Normalized Bayes Ratio.





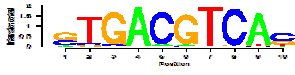



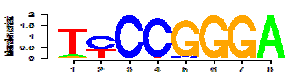
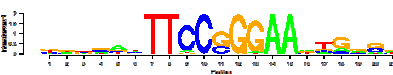

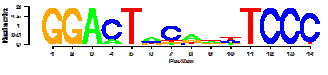



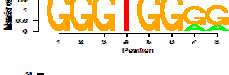
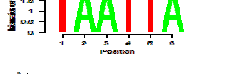
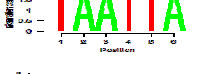
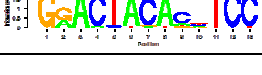
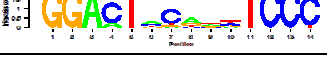
Cluster ID	Motif cluster logo	Bayes ratio	Transfac/Jaspar ID	TFBS profile logo	Mat-Compare divergen	MACO score
38		25.7	M00473 FOXO1		0.095	0.998
49		24.7	M00437 CHX10		0.072	0.999
95		20.9	M00179 CRE-BP1		0.542	0.942
157		17.9	MA0063 Nkx2-5		0.148	0.991
159		17.7	MA0137 STAT1		0.580	0.923
2		36.9	PF0074		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

Figure 1. A. Membership map that summarizes BMC2 clustering performance on 1152 motifs from 59 TFBS models from the JASPAR 4 database⁷. In 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 BMC but belonged to different JASPAR models, c) blue indicates cases in which motifs i and j were not clustered together by BMC2 cluster 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 the 1152 motifs. **A2.** Membership map for the 873 motifs filtered with thresholds NBR=0 and PAP=0.5.

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

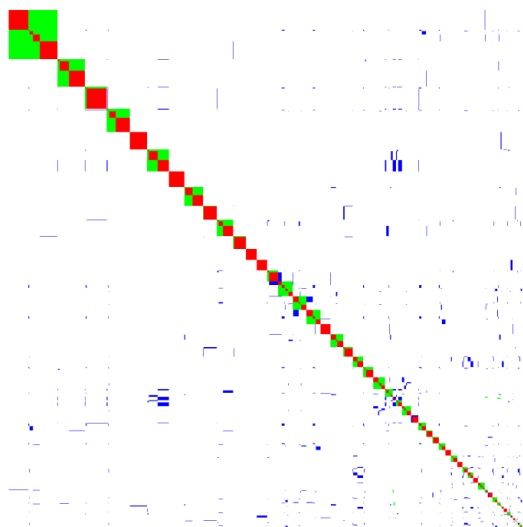
$$R = \frac{\text{number of motifs in both A and X}}{\text{total number of motifs in A}} \times \frac{\text{number of motifs in both A and X}}{\text{total number of motifs in X}}$$

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 TFBS model that are dispersed across more than one clusters, or both. **B1.** Membership map for all 1152 motifs. Each row represents one of 59 TFBS models, and each column corresponds to one of 56 BMC 2 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.

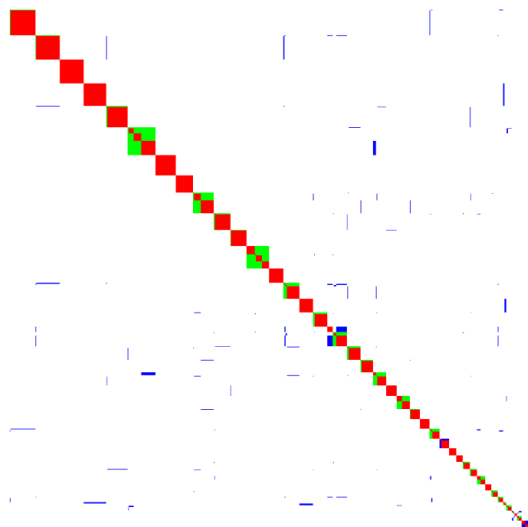
Figure 2. MotifOrganizer clustering results for TRANSFAC 9.3 data. This dataset contains 5452 sequence motifs that belonged to 295 TFBS models. MotifOrganizer's default settings were used, which permit motif widths to differ by no more than 2bp. To evaluate using quality metrics to filter clustering results, we tested nine different threshold settings, combining NBR thresholds of 0, 2, and 5 for clusters and posterior probability thresholds for motif assignments of 0, 0.5 and 0.8 (solid, dashed and dotted lines respectively). Red lines correspond to one-stage BMC 2 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.

Figure 1.

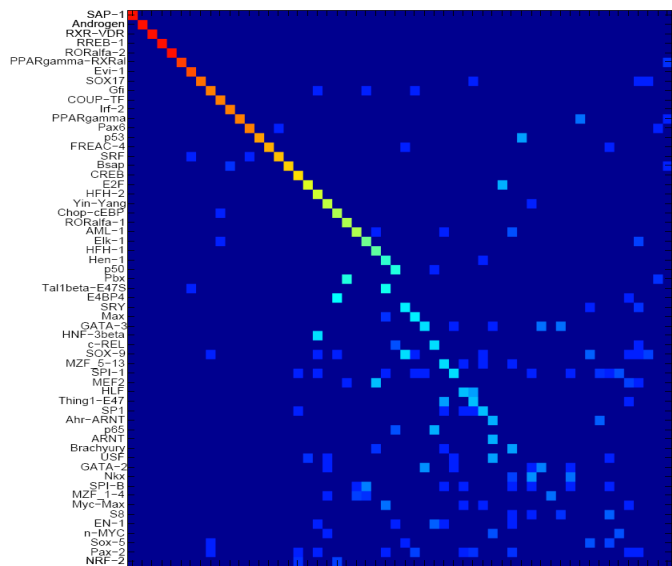
A1



A2



B1



B2

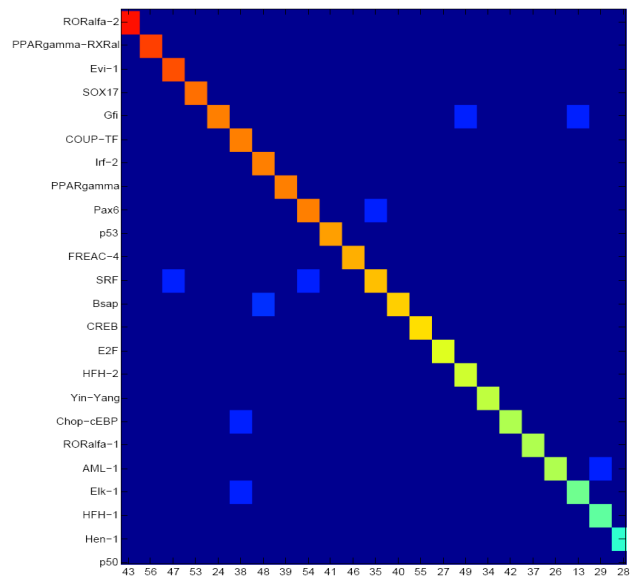
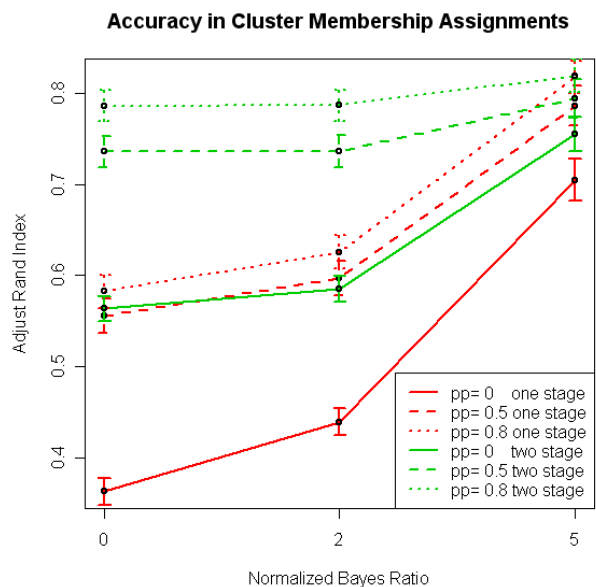


Figure 2

A



B

