Supporting Methods

Andrew D Smith Pavel Sumazin Michael Q Zhang

Using divergence to measure distance between motifs

The divergence (also called the Kullback–Leibler divergence) between two width w position-weight matrices M and M' is

$$\sum_{i=1}^{w}\sum_{j=A}^{T}(M_{ij} - M'_{ij})\log(M_{ij}/M'_{ij}),$$

where M_{ij} is the jth entry in the ith column of M, similarly for M'. When calculating divergences, if an entry in one matrix has a 0 value, and the corresponding entry in the other matrix does not, we replace the 0 entry with 0.005. As examples, the divergence between columns $[0.27, 0.27, 0.27, 0.19]$ and $[0.30, 0.30, 0.30, 0.10]$ is 0.07; between $[0.6, 0.4, 0.0, 0.0]$ and $[0.4, 0.6, 0.0, 0.0]$ is 0.16; between $[0.5, 0.4, 0.1, 0.0]$ and $[0.33, 0.33, 0.33, 0.00]$ is 0.36; between $[0.5, 0.5, 0.0, 0.0]$ and $[0.33, 0.33, 0.33, 0.00]$ is 1.52; and between $[0.5, 0.5, 0.0, 0.0]$ and $[0.25, 0.25, 0.25, 0.25]$ is 2.3.

Information content and the DME algorithm

The DME algorithm represents a shift in paradigm for motif discovery. The current prevalent paradigm is “occurrence centric.” Algorithms that fit the occurrence centric paradigm include Gibbs sampling-based methods, expectation maximization-based methods, and other algorithms such as CONSENSUS. These methods focus first on identifying the sites, and constructing position-frequency matrices from the sites. DME is “motif centric,” which means that DME first generates a motif matrix and then measures the quality of the motif with respect to the sequence data.

Occurrence centric algorithms must assume that the motif has a particular or expected number of occurrences. Some of these (e.g. CONSENSUS) assume each motif has one occurrence per input sequence. Others (e.g. GIBBS MOTIF SAMPLER) allow the user to provide an expected number of occurrences. The number of expected occurrences dictates the expected information content of the resulting motif, where more occurrences lead to lower expected information content. Therefore, by manipulating the number of occurrences (or its expectation), users can search for motifs with different levels of information content or degeneracy (low information content implies higher degeneracy).

The analogous parameter in motif centric methods is the information content of the motif. We measure information content of a motif in bits per column, and the DME algorithm only considers motifs that have more than the user-specified information content. Setting bits per column is therefore analogous to setting frequency of occurrences when using GIBBS MOTIF SAMPLER or MDSCAN. It is important to set the bits per column sufficiently high that noise is not pooled to create a high scoring motif, and low enough to capture the degeneracy in true signals.

The relationship between the information content of motifs and the expected number of occurrences of the motif was first studied in ref. 1. More information on this relationship can be found on the web at:

`http://www.lecb.ncifcrf.gov/~toms/paper/`

Unfortunately, we were not able to generate a perfect mapping between bits per column and frequency of occurrence. Even if this relationship were known for random data, we expect that it would not apply to real data. We are able to provide general guidelines based on parameter values that we have found successful

Width	Column Type Set	bits per column	min g (local search)
6	A	1.900	0.01
7	A	1.900	0.01
8	A	1.800	0.01
9	A	1.650	0.05
10	A	1.600	0.05
11	B	1.550	0.10
12	B	1.500	0.10

$$\text{A} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} .5 \\ .5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} .5 \\ 0 \\ .5 \\ 0 \end{bmatrix}, \begin{bmatrix} .5 \\ 0 \\ 0 \\ .5 \end{bmatrix}, \begin{bmatrix} 0 \\ .5 \\ .5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ .5 \\ 0 \\ .5 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ .5 \\ .5 \end{bmatrix}, \begin{bmatrix} .33 \\ .33 \\ .33 \\ 0 \end{bmatrix}, \begin{bmatrix} .33 \\ .33 \\ 0 \\ .33 \end{bmatrix}, \begin{bmatrix} .33 \\ 0 \\ .33 \\ .33 \end{bmatrix}, \begin{bmatrix} 0 \\ .33 \\ .33 \\ .33 \end{bmatrix}, \right\}$$

$$\text{B} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} .5 \\ .5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} .5 \\ 0 \\ .5 \\ 0 \end{bmatrix}, \begin{bmatrix} .5 \\ 0 \\ 0 \\ .5 \end{bmatrix}, \begin{bmatrix} 0 \\ .5 \\ .5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ .5 \\ 0 \\ .5 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ .5 \\ .5 \end{bmatrix} \right\}$$

Table 7: Values for parameters of the DME algorithm that we have found to produce good results. For each width, the table gives the set of column types, minimum bits per column and smallest value of g in local search. For the column type sets, we usually replace zero entries with 10^{-10} to prevent taking logarithms of 0. These are the values we use most often for identifying transcription factor binding sites.

(see Table 7). Bits per column values depend on column type used to generate motifs, motif width, and input size. Table 7 assumes input foreground set total size that is comparable with the data sets used in the paper "Identifying tissue-specific transcription factor binding sites in vertebrate promoters", which range from 20 to 50 Kb. Smaller foreground sets may permit reduction in the bits per column setting, and larger foreground sets may require an increase in the bits per column setting.

Remember, lowering the minimum bits per column of motifs in the search space has the effect of making the search space larger, and an exhaustive traversal of that space more time consuming. Good bits per column settings will result in imperfect motifs with high information content and a reasonable number of occurrences. We recommend that users try a few different settings for the bits per column parameter to identify the setting most appropriate to their particular data sets.

Other programs applied to tissue-selective data sets

We applied MDSCAN and GIBBS MOTIF SAMPLER to identify motifs in the same tissue-selective promoter sets used in the paper. These include the Liver Specific Promoter Set (LSPS), the subset of EPD vertebrate promoters associated with liver (EPD-Liver), the Wasserman–Fickett muscle selective promoters (WF) and the subset of EPD vertebrate promoters associated with muscle (EPD-Muscle). For background we used the vertebrate subsets of EPD, with promoters associated with liver and muscle removed for analysis of liver and muscle, respectively.

Motifs identified by MDSCAN The parameters for MDSCAN were set to: scan and refine 100 motifs, report the top 10 motifs, refine for 100 iterations, and look for candidate motifs in the top 30 sequences. Performance of MDSCAN on these data sets was poor. Regardless of how many motifs were requested, MDSCAN could only identify one distinct motif for each data set. These motifs, and the TRANSFAC motifs most similar to them, are given in Table 8.

To improve the performance of MDSCAN, we masked out simple repeats in the sequences. We define simple repeats as subsequences of length equal to or greater than the motif width that are composed of a single nucleotide or two alternating nucleotides such as "AGAGAGA." Results of MDSCAN applied to the

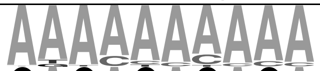
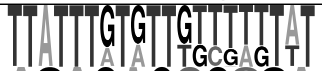
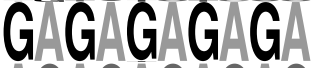
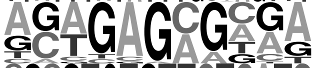
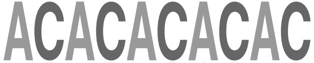
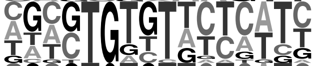
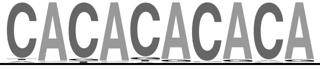
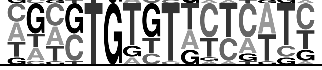
		Best TRANSFAC Match			
SeqSet	Sequence Logo	Sequence Logo	Accession	Factor	Divergence
EPD-liver			M00987	FOXP1	0.74
LSPS			M00723	GAGA	1.26
WF			M00317	Poly-A	1.55
EPD-muscle			M00317	Poly-A	1.27

Table 8: Motifs identified by MDSCAN

masked data sets are presented in Table 9.

For the modified liver sets, MDSCAN identified many near-identical motifs that resemble the motif for HNF1. For the EPD-Muscle set, MDSCAN identified motifs that seem to represent essentially two distinct signals, one resembling a poly-C signal, and another resembling SRF. MDSCAN did not perform as well on the Wasserman–Fickett muscle set; MDSCAN identified essentially one motif, which resembles the motif for Sp1, but no motifs resembling those associated with SRF, MEF2, or MyoD.

Motifs identified by GIBBS MOTIF SAMPLER We use GIBBS MOTIF SAMPLER to identify overrepresented motifs in the foreground without a background. The version we used was 2.04.014 (December 1, 2003), and we set the expected number of motif elements to 40, and requested 10 motifs, each of width 10. We used 20 seeds, and set the flag for using the nucleic acid alphabet.

GIBBS MOTIF SAMPLER identified a richer set of motifs than MDSCAN, and did not suffer from the problem of being confused by simple repeats. However, the Gibbs Motif Sampler did not identify any of the major motifs associated with the tissues. GIBBS MOTIF SAMPLER did not identify motifs that are similar to known HNF1 motifs in either of the liver promoter sets, nor did it identify motifs similar to known SRF or MEF-2 motifs in either of the muscle sets. Both DME and MDSCAN were able to identify some of these. This performance by GIBBS MOTIF SAMPLER, along with the extremely poor performance on simulated data sets (see the paper) suggests that these may essentially be random and not representative of any strong motifs in the data.

References

[1] Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188**, 415–431.

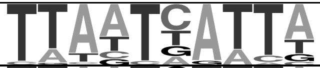
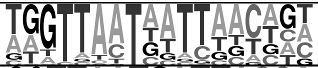
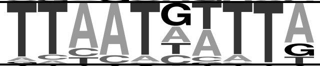
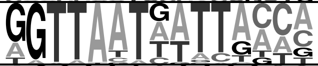
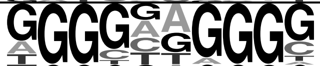
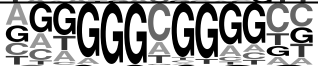
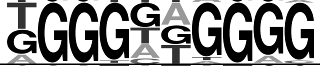
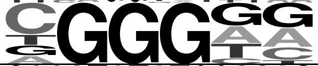
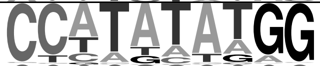
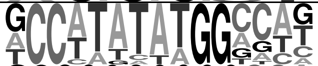
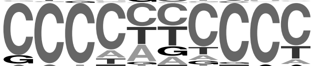
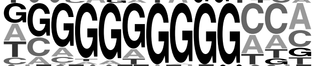
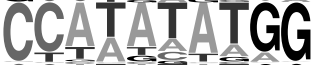
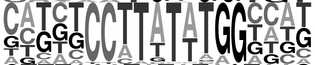
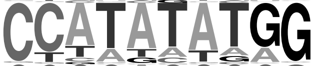
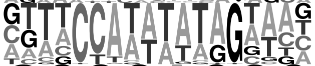
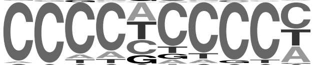
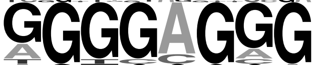
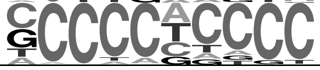
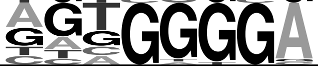
		Best TRANSFAC Match			
SeqSet	Sequence Logo	Sequence Logo	Accession	Factor	Divergence
EPD-liver			M00790	HNF1	0.32
LSPS			M00132	HNF1	0.24
WF			M00932	Sp-1	0.78
WF			M00986	Churchill	0.60
EPD-muscle			M00215	SRF	0.20
EPD-muscle			M00491	MAZR	0.72
EPD-muscle			M01007	SRF	0.25
EPD-muscle			M00392	AGL3	0.27
EPD-muscle			M00649	MAZ	0.69
EPD-muscle			M00083	MZF1	0.60

Table 9: Top motifs identified by MDSCAN after masking simple repeats

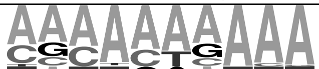
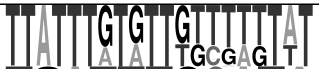
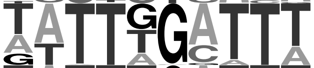
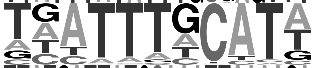
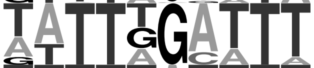
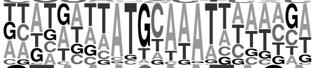
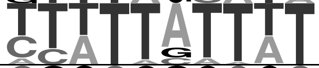
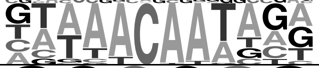
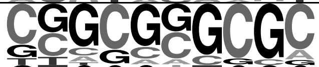
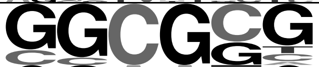
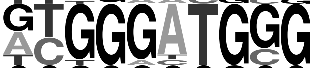
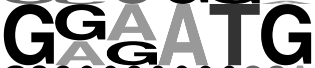
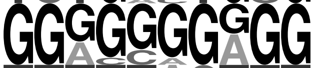
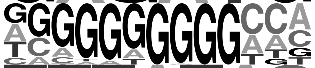
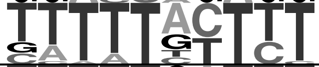
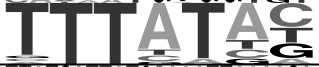
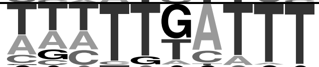
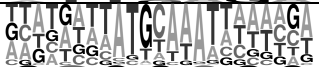
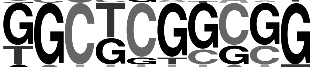
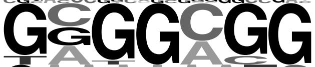
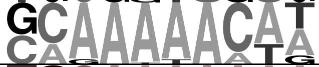
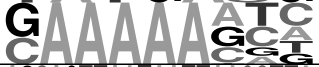
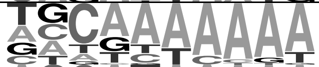
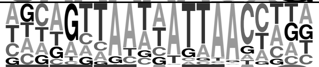
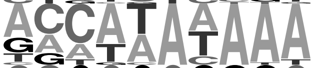
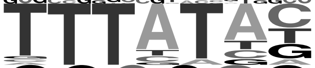
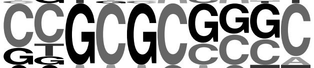
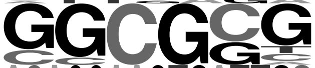
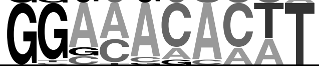
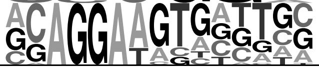
		Best TRANSFAC Match			
SeqSet	Sequence Logo	Sequence Logo	Accession	Factor	Divergence
EPD-liver			M00987	FOXP1	1.21
EPD-liver			M00930	Oct-1	1.38
EPD-liver			M00138	Oct-1	1.38
EPD-liver			M00160	SRY	1.16
LSPS			M00803	E2F	0.70
LSPS			M00704	TEF-1	0.69
LSPS			M00491	MAZR	0.97
LSPS			M00980	TBP	0.94
WF			M00138	Oct-1	0.97
WF			M00695	ETF	0.97
WF			M00734	CIZ	1.31
EPD-muscle			M01011	HNF1	1.01
EPD-muscle			M00980	TBP	0.72
EPD-muscle			M00803	E2F	0.57
EPD-muscle			M00339	c-Ets-1	1.41

Table 10: Top motifs identified by GIBBS MOTIF SAMPLER