BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i283–i291
doi:10.1093/bioinformatics/bti1025

Improved detection of DNA motifs using a self-organized clustering of familial binding profiles

Shaun Mahony[1,], Aaron Golden[1,2], Terry J. Smith[1] and Panayiotis V. Benos[3]*

[1]*National Centre for Biomedical Engineering Science, NUI Galway, Galway, Ireland,*
[2]*Department of Information Technology, NUI Galway, Galway, Ireland and*
[3]*Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh Cancer Institute and Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA*

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: One of the limiting factors in deciphering transcriptional regulatory networks is the effectiveness of motif-finding software. An emerging avenue for improving motif-finding accuracy aims to incorporate generalized binding constraints of related transcription factors (TFs), named familial binding profiles (FBPs), as priors in motif identification methods. A motif-finder can thus be 'biased' towards finding motifs from a particular TF family. However, current motif-finders allow only a single FBP to be used as a prior in a given motif-finding run. In addition, current FBP construction methods are based on manual clustering of position specific scoring matrices (PSSMs) according to the known structural properties of the TF proteins. Manual clustering assumes that the binding preferences of structurally similar TFs will also be similar. This assumption is not true, at least not for some TF families. Automatic PSSM clustering methods are thus required for augmenting the usefulness of FBPs.

Results: A novel method is developed for automatic clustering of PSSM models. The resulting FBPs are incorporated into the SOMBRERO motif-finder, significantly improving its performance when finding motifs related to those that have been incorporated. SOMBRERO is thus the only existing *de novo* motif-finder that can incorporate knowledge of all known PSSMs in a given motif-finding run.

Availability: The methods outlined will be incorporated into the next release of SOMBRERO, which is available from http://bioinf.nuigalway.ie/sombrero

Contact: shaun.mahony@nuigalway.ie.

1 INTRODUCTION

Finding *cis*-regulatory motifs in DNA sequences remains a fundamental problem in computational biology. Many approaches to the solution of the problem exist, with methods based on statistical learning theory being particularly popular. For example, maximum likelihood estimation (e.g. MEME (Bailey and Elkan, 1994)) and Gibbs sampling [e.g. AlignACE (Hughes *et al*., 2000), Co-Bind (GuhaThakurta and Stormo, 2001) and BioProspector (Liu *et al*., 2001)] are widely used. Alternative motif identification methods have also been proposed, including word enumeration, winnowing and dictionary construction-based methods (Bussemaker *et al*., 2000; Gupta and Liu, 2003; Pevzner and Sze, 2000; Rigoutsos and Floratos, 1998). However, a recent survey indicates that the effectiveness of existing methods is still in need of much improvement (Tompa *et al*., 2005).

Phylogenetic footprinting techniques are obviously one avenue for improving the accuracy and effectiveness of motif-finders (Blanchette and Tompa, 2003; Lenhard *et al*., 2003; Loots *et al*., 2002; McCue *et al*., 2002). However, these approaches only reduce the amount of sequence to be analysed, and do not improve the accuracy of the underlying motif-finding algorithm, which remains the core issue.

A recently proposed alternative avenue for improving motif detection aims to change the *de novo* motif detection problem from "an unsupervized learning problem into a semi-unsupervized learning problem that makes substantial use of existing biological knowledge" (Xing and Karp, 2004). The emerging methods are directed by the realization that great potential exists for improving motif recognition by modelling and exploiting the regularities that are shared by structurally related transcription factors. In their MotifPrototyper framework, Xing and Karp use hidden Markov–Dirichlet multinomial models to represent the structural features of a given family of related transcription factors (TFs). The authors show how a mixture model built on top of multiple structural models can facilitate a Bayesian estimation of the position specific scoring matrix (PSSM) of a novel motif, and therefore known

*To whom correspondence should be addressed.

biological information is used to improve the performance of a motif-finder.

While Xing and Karp use knowledge of structurally similar DNA-binding motifs to improve a motif-finder, Sandelin and Wasserman (2004) aim for the same goal by using alignments of similar DNA-binding motifs. The latter demonstrate an effective way of representing the constrained binding site diversity within a family of structurally related transcription factors by building familial binding profiles (FBPs) that are in effect the 'average' binding motif for a set of related transcription factors. Sandelin and Wasserman suggest two applications of their familial binding profiles; first, in improving motif-finding detection, and second, in allowing the structural classification of a newly discovered DNA-binding motif. They successfully demonstrate both applications in their study.

Sandelin and Wasserman's FBPs are manually constructed from PSSMs for which the structural class of the corresponding TF is known (11 TF families in their study). Lack of structural knowledge for a large number of TFs precludes the extended manual construction of FBPs, and therefore methods for automatic clustering of similar PSSMs are needed. Such methods should have the ability to construct FBPs without prior knowledge of the structural classes of constituent PSSMs. In the current study, we show how an unsupervized neural network method, the self-organizing map (SOM), can be effectively applied to the automatic clustering of PSSMs.

One application of FBPs is their use as priors in motif identification algorithms (Sandelin and Wasserman, 2004). Indeed, Sandelin and Wasserman demonstrated that the motif-finding performance of both the Gibbs Motif Sampler (Thompson *et al*., 2003) and ANN-Spec (Workman and Stormo, 2000) are dramatically improved when an appropriate FBP is incorporated as a prior bias in these methods. If binding sites that are similar to the FBP prior are present in the input sequences, the motif-finders are effectively biased intentionally towards finding the correct pattern. However, this enhancement to traditional motif-finders is critically dependant on the correct choice of biasing FBP. The inclusion of an incorrect prior can result in the failure of the motif-finder to detect sites for a different TF. Currently, traditional motif-finding methods can only incorporate one FBP as a prior during any given motif-finding run. In effect, the TF that is acting through sites in the input sequences (or at least its functional class) must be known in advance, and this is not usually possible in typical *de novo* motif-finding experiments. The application of FBPs as priors for motif-finding methods would therefore seem to have limited applicability.

However, our recently described motif-finding algorithm, named SOMBRERO (Mahony *et al*., 2005), has the potential to bypass the problems that other methods face when incorporating FBPs as priors. SOMBRERO, based on the self-organizing map (SOM) neural network, can find a set of multiple distinct motifs for an input dataset in a truly simultaneous manner (see Methods for details). As we demonstrate in this study, the nodes on SOMBRERO's neural grid can be initialized to correspond to a complete set of familial binding profiles through the use of a SOM clustering of PSSMs. This biases certain nodes towards finding particular TF binding sites. If a motif that is similar to a FBP present on the SOM grid is also present in the input sequences, the corresponding binding sites will be attracted to a particular node and the motif will remain conserved throughout SOM training. On the other hand, if a motif is not present in the input sequences, corresponding binding sites will not be available to reinforce the existence of the motif on the grid, and therefore the motif will 'fade out' during training until it is not present on the grid at completion. We demonstrate that the use of a complete set of FBPs as priors for SOMBRERO acts as an effective way to improve the performance of motif detection, without biasing against finding motifs that are not represented in the prior.

2 METHODS

2.1 Algorithm for profile comparison

In this study, Pietrokovski's (1996) methods for aligning two PSSMs are used, but are combined with Sandelin and Wasserman's method for calculating the P-value of an alignment. Specifically, column-to-column comparisons are made using Pearson's correlation coefficient:

$$r(C,D) = \frac{\sum_{b=A}^{T}(C_b - \bar{C}) \cdot (D_b - \bar{D})}{\sqrt{\sum_{b=A}^{T}(C_b - \bar{C})^2 \cdot \sum_{b=A}^{T}(D_b - \bar{D})^2}}$$

where C_b and D_b are the probability values of base b in columns C and D, respectively, and $\bar{C}$ and $\bar{D}$ are the means of the values in columns C and D, respectively. A modified Smith–Waterman algorithm (Smith and Waterman, 1981) is used to find optimal local alignments of PSSM pairs. In the current work, no gaps were allowed in the alignment.

In order to compare alignments of different widths, the method for the calculation of empirical P-values described by Sandelin and Wasserman is followed exactly. The method involves extensive analysis with simulated PSSMs to determine the likelihood of any score given the lengths of aligned matrices. The simulated PSSMs reflect the properties of the PSSMs in the JASPAR database (Sandelin and Wasserman, 2004). The construction of a dataset of 10 000 simulated matrices follows the instructions on Sandelin and Wasserman's website (http://forkhead2.cgb.ki.se/jaspar/additional/index.htm).

2.2 A SOM for clustering PSSMs

The SOM is an unsupervized neural network algorithm (Kohonen, 1995). The general structure of the SOM is a 2D lattice of interconnected nodes. The nodes contain models (typically n-dimensional vectors) that are adjusted during the

training process in order to represent features of the input space. At the end of each training iteration, new node models are generated by averaging the input datapoints clustered at that node, as well as including contributions from neighbouring nodes. The training algorithm results in similar input datapoints being clustered at a given node, and similar nodes will be located close to one another on the lattice. Because of this property of (local) topological conservation, the SOM is often used to visualize high-dimensional data on a 2D display, but the clustering mechanism also makes the SOM an effective means of finding common features in a dataset. Typical applications of the SOM in biosequence analysis include the study of codon usage (Abe *et al*., 2003; Mahony *et al*., 2004; Wang *et al*., 2001) and the clustering of similar protein sequences (Kohonen and Somervuo, 2002).

In the current application of the SOM, the data-space to be explored will be a set of PSSMs. Therefore, PSSMs are used as models at each SOM node. The training algorithm for the SOM of PSSMs, henceforth referred to as the 'binding profile SOM' (BP-SOM), proceeds as follows:

(1) The BP-SOM lattice size is chosen. In general, the choice of BP-SOM lattice size is arbitrary. However, when the BP-SOM is used in conjunction with SOMBRERO (see below), the BP-SOM lattice size must be equal to the SOMBRERO lattice size. The choice of SOMBRERO's lattice size is dependant on the size of the dataset being analysed and will be discussed in Section 2.3. Each node model m_j is initialized as a length 12 PSSM with random values.

(2) For each PSSM, $x_i, (i = 1, \ldots, N)$:
   2.1 x_i is aligned to every SOM node model m_j using the alignment method described in Section 2.1 above.
   2.2 The node w whose model m_w has the best P-value score to x_i is selected and x_i is clustered at node w.

(3) Update step:
   3.1 At each node j, all clustered members are aligned to give the new weighted alignment matrix A_j. The weight (Z_v) of each member is calculated by the average P-value (P_v) obtained from comparisons of profile v with all other members of the same node: $Z_v = 1 - P_v$. Following the FBP construction procedure outlined by Sandelin & Wasserman (2004), the node member with the highest Z_v is designated as the alignment positioning template.
   3.2 New models are generated according to the equation:

$$m_j(t+1) = \sum_i^N align(x_{i,k} \times Z_{i,k} \times e^{-|j-k|^2/\gamma})$$

where *align( )* is an alignment function that aligns the columns of each x_i (clustered at node k) to the alignment positioning template at node j, and $|j - k|$ is the distance on the SOM grid between nodes j and k. The measure of sharpness γ of the Gaussian factor is defined as $\gamma = 1/(\log(\delta))$, where δ ranges from 4 to 30 during training. This Gaussian neighbourhood factor ensures that adjacent nodes will strongly contribute to each other initially, but end up contributing little to each other at the end of training. The length of the new model depends on the quality of the alignment. Flanking columns with low information content (<0.4 bit) or insufficient sequence depth (at least half of the PSSMs clustered at node j have to contribute) are excluded from the new model, up to a minimum model length of eight. The length of each node's FBP model therefore changes during the training process, and this is possible because of the length independent nature of the P-value score. Finally, each m_j is normalized.

(4) The training process repeats from step 2 for 100 training cycles.

2.3 SOMBRERO

In SOMBRERO's algorithm, the nodes are PSSM models, but as the input space in the motif-finding domain is typically a set of DNA promoter sequences, the input datapoints to be clustered on the SOM are ℓ-mer sequence strings. Training proceeds by assigning ℓ-mers to various nodes on the lattice using a log-likelihood similarity measure, and model updates are carried out that incorporate contributions from neighbouring nodes. At the completion of the training process, SOMBRERO's nodes represent a complete set of motif features present in the input dataset. Training can be repeated over various values of ℓ in order to find motifs of different lengths.

In order to identify potential transcription factor binding sites (TFBS) motifs in the complete set of motif features, a third-order Markov chain-based model (of the intergenic DNA in the genome being studied) is used to generate a set of random datasets. Using these simulated datasets, the expected number of matches to each node's motif is found, and thus the motifs are ranked in terms of significance in a manner similar to that used by Sinha and Tompa (2002). Motifs that represent repetitive chains of DNA are filtered out at this stage using a suitable motif complexity threshold. Complete details of the SOMBRERO algorithm, including the simulation procedure and the complexity score, are described in our recent publication (Mahony *et al*., 2005).

The parameters used by SOMBRERO in the current study are mostly default settings. Specifically, SOMBRERO is run for 100 training iterations, using a third-order Markov background model of the organism under examination, and

training repeats to find motifs between lengths 8 and 22 bp. The size of the SOM lattice is scaled automatically with the input dataset size. According to previous tests, the optimum motif-finding performance of SOMBRERO can be obtained by keeping the ratio of lattice nodes to input dataset size roughly in the order of one node to 10 bp (Mahony *et al*., 2005). Applying this approximate ratio, the following SOM sizes were used in this study; 10×10 nodes for datasets in the interval 0–1999 bp, 20×10 nodes for the interval 2000–3999 bp, 30×15 nodes for the interval 4000–7999 bp, 40×20 nodes for the interval 8000–12499 bp and 50×25 nodes for datasets >12500 bp.

In the current study, three lattice initialization strategies were tested in order to determine how SOMBRERO's motif-finding performance is affected by such choices. The first method is the gradient random initialization (Mahony *et al*., 2005), which results in a grid that is biased towards mononucleotide distributions at each corner, with gradients of preference in the intervening nodes. The second method evaluated in this study is a standard random initialization of the nodes. The third initialization method is named the 'prior initialization', and is a strategy that uses the final models that result from clustering known PSSMs using a BP-SOM as described in Section 2.2. In the latter case, the SOMBRERO lattice begins the training process with various nodes biased towards finding particular TF binding motifs. When using a BP-SOM to bias a SOMBRERO lattice, the lattice sizes obviously need to be equivalent. In order to make the BP-SOM models compatible with the length ℓ SOMBRERO models, only the ℓ most informative concurrent columns in each BP-SOM node model are used. Padding, using columns of neutral bias, is used if the BP-SOM model is of a length less than ℓ.

2.4 Training datasets

JASPAR (Sandelin *et al*., 2004), the non-redundant set of high-quality transcription factor binding matrices (TFBM), was used in this study. A subset of 71 JASPAR PSSMs, defined by Sandelin and Wasserman as those PSSMs belonging to families represented by five or more members in JASPAR (and excluding zinc finger motifs), is used in various contexts. In this study, three different PSSM datasets are used to train BP-SOMs as priors for SOMBRERO. The three collections are: (1) a selection of 257 mammalian-specific PSSMs taken from the JASPAR and TRANSFAC databases, (2) the entire set of yeast specific PSSMs contained in the *Saccharomyces cerevisiae* Promoter Database (SCPD; http://cgsigma.cshl.org/jian/) and (3) a set of 75 *Drosophila*-specific PSSMs constructed by Dan Pollard from a DNAse I footprint database (Bergman *et al*., 2005), and available from http://rana.lbl.gov/∼dan/matrices.html.

2.5 Evaluation datasets

Both artificial and real sequence datasets are used to evaluate the performance of various motif-finders in this study. The artificial datasets were constructed using a third-order Markov model of yeast intergenic sequence. The datasets contained various instances of one of four different motifs; *CREB*, *E4BP4*, *MIG1* and *GAL4* (each PSSM was procured from TRANSFAC). For each motif, 80 datasets were constructed. Each dataset contained sequences to a total sequence length between 1 and 8 kb per dataset. A random number of the relevant motif instances (mean occurrences for each motif was ∼15) were placed at random intervals in each dataset.

Ten yeast genomic sequence sets were collected from the SCPD database. The selection of sequence sets is based on there being at least a total of four motif instances annotated in each set. Each sequence set consists of multiple yeast promoter regions, each region at least 500 bp long and containing on either strand a number of occurrences of a predominant motif (and also possibly other minor motifs) as specified by the name of the dataset (Table 1, where the 'bp' column gives the size of each sequence set, and the 'sites' column gives the number of known motif instances in each set).

A dataset of *Drosophila* regulatory regions is also used in the present study. The set contains 19 regulatory regions of 9 *Drosophila* genes (Berman *et al*., 2002) that harbour binding sites for the TFs *Bicoid* (*bcd*), *Caudal* (*cad*), *Hunchback* (*hb*), *Knirps* (*kni*) and *Krüppel* (*Kr*). The total size of the *Drosophila* dataset is 22 535 bp.

The (real) yeast and *Drosophila* genomic sequence datasets have been used previously by ourselves and others to test the accuracy of motif-finder algorithms (Mahony *et al*., 2005; Xing *et al*., 2004).

2.6 MEME and AlignACE

We compared the performance of SOMBRERO (with and without priors) to MEME (Bailey and Elkan, 1994) and AlignACE (Hughes *et al*., 2000). Every effort has been made to ensure that the comparison between the various motif-finding programs is fair. However, we understand that the conclusions from such comparisons should always be interpreted with caution since the very nature of the algorithms and the input parameters make such comparisons somewhat 'biased'. Furthermore, it should be noted that neither MEME nor AlignACE allow priors to be used in the sense that SOMBRERO allows. Therefore, the results of the comparisons should serve only as a frame of reference.

MEME is run using its default parameters, except for allowing the search of both strands for up to 20 motifs, each of which can occur zero or more times in a sequence. AlignACE is run online (http://atlas.med.harvard.edu/cgi-bin/alignace.pl) with default parameters, which are nearly identical to those of MEME, except that the motif length needs to be specified. In all AlignACE runs, the correct motif length was specified.

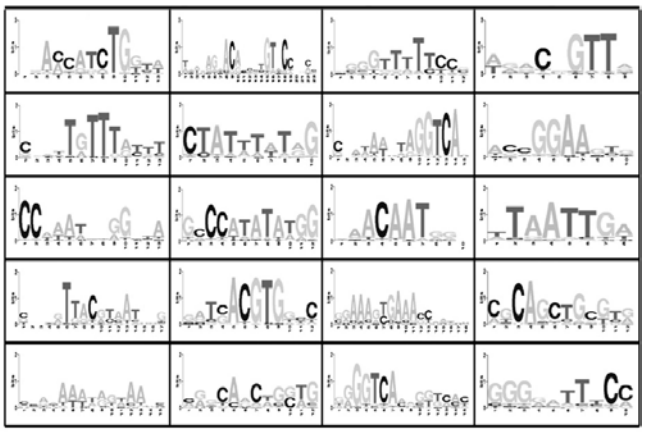
In every case, accuracy of motif-finding is judged in relation to the best matching motif found in the top 20 results from each method. We chose to examine the top 20 motifs instead of only the absolute top motif as in other studies (Tompa *et al*.,

2005), because sometimes many of the top scoring motifs consist of slight variations of the same 'dominant' motif in the set. Examining the top 20 motifs is also consistent with the biological usefulness of such algorithms (biologists almost never limit their testing to the absolute topmost prediction).

3 RESULTS

3.1 Clustering PSSMs using the SOM

As a demonstration of the clustering power of the SOM, a 5×4 node BP-SOM was trained on the 71 member JASPAR subset using the algorithm outlined in Section 2.2. The final states of the SOM nodes are shown in Figure 1 (high quality figure available as supporting information; http://bioinf.nuigalway.ie/ISMB2005). Figure 1(b) shows the names of the PSSMs (including family names) that are clustered at each node at the end of training, and therefore contribute to the corresponding final motifs. A high degree of clustering according to familial membership can be observed in the nodes. Figure 1(a) shows the motifs that correspond to the final PSSM clustering on the SOM, and should be compared with the 11 familial binding profiles described by Sandelin and Wasserman (http://forkhead2.cgb.ki.se/jaspar/additional/fbs.htm). A high degree of similarity exists between certain nodes and particular FBPs, for example, the Sandelin and Wasserman REL family FBP and node (3, 0), or the Sandelin and Wasserman HMG family FBP and node (2, 2).



(a)

bHLH Tal1beta-E47S bHLH Thing1-E47	NUCLEAR Androgen HMG HMG-1	REL Dorsal_1	TRP-CLUSTER MYB.ph3 TRP-CLUSTER c-MYB_1 TRP-CLUSTER GAMYB
FORKHEAD HFH-2 FORKHEAD HNF-3b FORKHEAD HFH-1 FORKHEAD HFH-3 FORKHEAD FREAC-2 FORKHEAD FREAC-4	MADS AGL3 MADS MEF2 FORKHEAD FREAC-7	NUCLEAR RORalfa-1 NUCLEAR RORalfa-2	ETS SAP-1 ETS SPI-B ETS NRF-2 ETS SPI-B ETS E74A ETS e-ETS ETS Elk-1
MADS Agamous	MADS SRF	HMG SOX-9 HMG Sox-5 HMG SOX17 HMG SRY	HOMEO Nkx HOMEO S8
bZIP HLF bZIP E4BP4 bZIP cEBP bZIP CREB	bHLH-ZIP Max bHLH-ZIP USF bHLH-ZIP n-MYC bHLH-ZIP Myc-Max bHLH ARNT bZIP bZIP911 bZIP TCF11 bZIP bZIP910	TRP CLUSTER Irf-1 TRP_CLUSTER Irf-2	bHLH Hen-1 bHLH Ahr-ARNT FORKHEAD FREAC-3
MADS SQUA HOMEO HNF-1 HMG HMG-IY	bHLH Myf	NUCLEAR CFI-USP NUCLEAR COUP-TF NUCLEAR RXR-VDR NUCLEAR PPARgamma NUCLEAR PPARgamma-rxr	REL p65 REL NF-kB REL p50 REL Dorsal_2 REL c-REL HOMEO EN-1 bZIP Chop-cEBP

(b)

Fig. 1. Clustering 71 JASPAR PSSMs with a 5×4 SOM.

3.2 Using familial binding profiles as priors for SOMBRERO: improved performance in artificial sequence data

Given that the BP-SOM has been demonstrated to automatically cluster PSSMs according to familial binding constraints, we are now interested in using the familial binding profiles thus generated to improve the discovery of transcription factor binding sites. This can be achieved by incorporating the end state of a BP-SOM that has been trained on a dataset of known PSSMs as the starting state for SOMBRERO (see Methods).

In this section, the performance of three SOMBRERO initialization strategies are compared when applied to artificial sequence data (described in Section 2.4). Three SOMBRERO SOMs (lattices sized according to the dataset size, as explained in Section 2.3) are trained on each of the artificial datasets, where each SOM uses a different initialization strategy (gradient random, standard random and BP-SOM priors). For the cases where SOMBRERO uses a prior, the prior refers to the end state of a BP-SOM that has been previously trained on a selection of 257 mammalian-specific PSSMs.

Figure 2 collates the results of this analysis. Each of the graphs shows the average performance of each initialization strategy in the datasets, measured with the *harmonic mean* (F) of sensitivity (S_N) and specificity (S_P):

$$F = 2(S_N \cdot S_P)/(S_N + S_P).$$

The value of F varies between 1, representing perfect recall of the true motif instances with no false positive predictions, and 0, representing no partially correct prediction found for the motif.

As can be observed from Figure 2, there is little variation between the average performance of the gradient random initialization and the random initialization in any of the four datasets, and this is in line with previous observations (Mahony *et al*., 2005). However, the use of the prior initialization improves motif-finding performance over the other initialization strategies for the two motifs that are present in the prior set (*CREB* and *E4BP4*), and the improvement lasts consistently as the dataset size increases. As expected, no improved performance is observed through the use of a mammalian-specific prior for either the *GAL4* or *MIG1* motif datasets, as neither motif is included in the prior dataset. However, neither does the use of the mammalian-specific prior initialization adversely affect performance when finding these yeast motifs, and the prior initialization performs similarly to the random initializations throughout these two datasets. This example therefore demonstrates that the use of

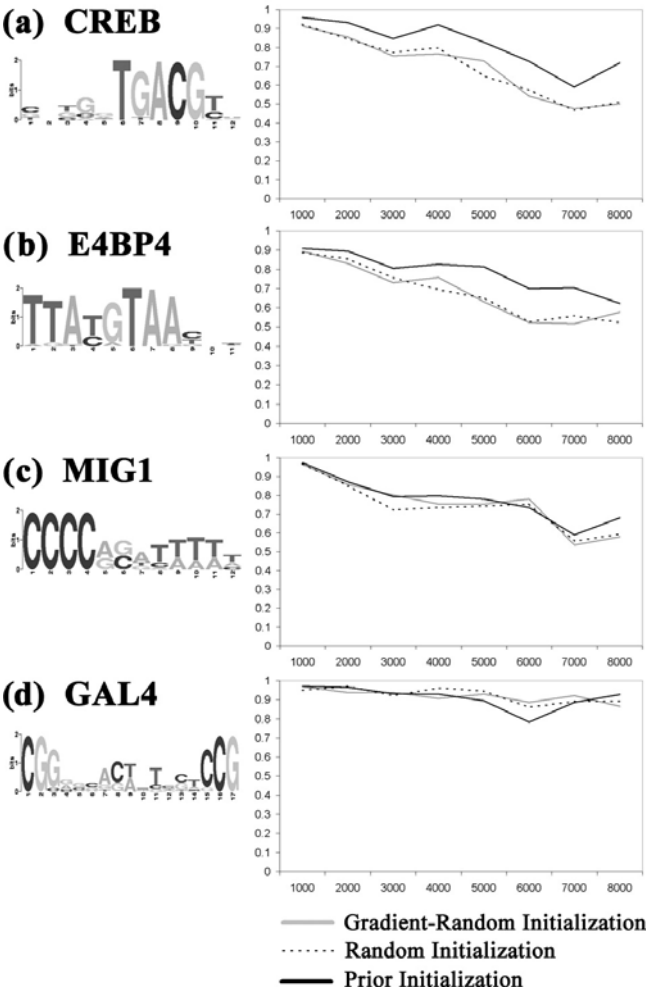


Fig. 2. SOMBRERO's performance in artificial sequence datasets using various initialization strategies. The graphs plot the average harmonic mean values (F) of each initialization strategy against dataset size (in bp). The motifs *CREB* and *E4BP4* are included in the PSSM dataset used in the prior initialization, but the motifs *MIG1* and *GAL4* are not.

prior biological knowledge with SOMBRERO gives the type of improved performance that is desired; motif-finding performance is improved if the relevant motif is present in the prior, and performance is not negatively affected for those motifs or structural classes that are not represented in the prior.

3.3 Improved SOMBERO performance in *S.cerevisiae* regulatory regions

The performance of a selection of motif-finders is evaluated using 10 real genomic datasets taken from *S.cerevisiae* (described in Section 2.4). SOMBRERO's performance using the original, gradient random initialization, and using the prior initialization (which uses an appropriately sized BP-SOM previously trained on yeast specific PSSMs) is compared with the performance of MEME and AlignACE in the datasets.

The results of the analysis are displayed in Table 1. The results in each dataset are described in terms of false-negative rates (FN), false-positive rates (FP) and harmonic mean (F). It can be seen from the table that the use of the prior initialization allows SOMBRERO to gain the best performance rate in 8 of the 10 datasets. An entire collection of yeast specific motifs has therefore been used as prior knowledge in order to improve the performance of a motif-finder in real genomic data.

3.4 Improved SOMBRERO performance in *Drosophila* regulatory regions

In order to demonstrate the advantages of using priors in a large genomic dataset, and also to show that SOMBRERO's ability to predict multiple distinct motifs is not affected by the use of a prior, a selection of motif-finders were tested on a large (22 535 bp) set of *Drosophila* regulatory regions. Again, the original SOMBRERO initialization, MEME and AlignACE were compared with SOMBRERO incorporating a set of priors. SOMBRERO was run using a 50×25 node SOM, and the prior used was a 50×25 node BP-SOM previously trained on a set of 75 *Drosophila*-specific PSSMs.

The results of the comparison are shown in Table 2. Again, the results are presented in terms of FN, FP and F. From the table, it can be seen that no method can be said to have effectively detected the *kni* binding motif. However, for the other four motifs, the use of a prior with SOMBRERO leads to the greatest performance rates. The use of a prior gives a significant increase in performance in finding the *bcd* and *Kr* motifs. Overall, the use of the prior initialization yields the same performance as the gradient random initialization in finding the *cad* and *hb* motifs. It should be noted, however, that the gradient random initialization of the original SOMBRERO program results in nodes that are biased towards mono-nucleotide motifs. Since *cad* and *hb* are both motifs that contain constraints for the binding of a run of 'T's, the original gradient random initialization may have contained an inadvertent 'prior' that is biased towards finding *cad* and *hb*.

4 DISCUSSION

The use of the automatic clustering method presented in this work can be viewed as a natural progression in the study of familial binding profiles. We have expanded the applicability of familial binding profiles as prior knowledge for motif identification in a significant way through the use of the SOM. We demonstrated that the SOM can be used to cluster PSSMs on a 2D grid. In general, the resulting grid of FBPs agreed with the evolutionary classification of the TFs, although differences were observed. Such differences are expected since members of the same TF family can have drastically different binding preferences, depending on the amino acids in 'key' amino acid positions (Benos *et al*., 2002).

Table 1. Comparison of motif detectors on 10 yeast promoter sequence datasets

	bp	sites	SOMBRERO (original initialization)			SOMBRERO (with prior)			MEME			AlignACE		
			FN	FP	*F*	FN	FP	*F*	FN	FP	*F*	FN	FP	*F*
abf1	8600	20	0.45	0.56	*0.489*	0.40	0.29	***0.649***	0.55	0.18	*0.581*	0.50	0.38	*0.556*
csre	2550	4	0.25	0.73	***0.400***	0.00	0.75	***0.400***	0.50	0.67	***0.400***	0.25	0.82	*0.286*
gal4	3100	14	0.07	0.24	*0.839*	0.07	0.07	***0.929***	0.29	0.17	*0.769*	0.21	0.08	*0.846*
gcn1	4500	25	0.60	0.29	*0.513*	0.44	0.33	***0.609***	0.92	0.80	*0.114*	0.60	0.44	*0.465*
gcr1	3350	9	0.22	0.69	*0.389*	0.00	0.41	***0.720***	0.44	0.44	*0.556*	0.33	0.63	*0.480*
hstf	3400	9	0.11	0.57	*0.552*	0.11	0.53	***0.615***	0.33	0.75	*0.364*	0.11	0.56	*0.593*
mat	3500	13	0.31	0.25	*0.720*	0.15	0.27	*0.801*	0.15	0.27	*0.786*	0.31	0.00	***0.818***
mcb	3150	12	0.08	0.65	*0.512*	0.08	0.31	*0.786*	0.25	0.25	*0.750*	0.08	0.08	***0.917***
mig1	4500	10	0.20	0.68	*0.457*	0.10	0.47	***0.667***	1.00	1.00	*0.000*	0.90	0.91	*0.095*
pho2	2350	6	0.50	0.91	*0.154*	0.33	0.80	***0.364***	1.00	1.00	*0.000*	1.00	1.00	*0.000*
Avg			0.32	0.61	*0.493*	0.22	0.42	***0.663***	0.56	0.45	*0.489*	0.43	0.47	*0.550*

The best *F*-score in each dataset is highlighted in bold.

Table 2. Comparison of motif detectors on 19 *Drosophila* regulatory sequences that contain instances of 5 regulatory binding sites

	sites	SOMBRERO (original initialisation)			SOMBRERO (with prior)			MEME			AlignACE		
		FN	FP	*F*	FN	FP	*F*	FN	FP	*F*	FN	FP	*F*
bcd	23	0.57	0.80	*0.274*	0.43	0.73	***0.366***	0.87	0.93	*0.094*	0.78	0.83	*0.189*
cad	63	0.43	0.46	*0.554*	0.43	0.45	***0.562***	0.75	0.43	*0.352*	0.78	0.67	*0.264*
hb	119	0.35	0.40	*0.621*	0.50	0.16	***0.624***	0.82	0.21	*0.299*	0.77	0.37	*0.333*
kni	24	0.76	0.94	*0.095*	0.76	0.89	***0.150***	0.88	0.82	*0.146*	0.88	0.93	*0.086*
Kr	61	0.61	0.59	*0.400*	0.30	0.33	***0.683***	0.64	0.46	*0.431*	0.52	0.25	*0.581*

The best *F*-score for each motif is highlighted in bold.

Using the BP-SOM as prior knowledge significantly improves SOMBRERO's motif-finding performance, and allows SOMBRERO to outperform other popular motif-finders on real genomic datasets. Currently, SOMBRERO is the only motif-finding algorithm that can use an entire set of PSSM models as priors, and therefore overcomes adverse biases introduced by the selection of an incorrect PSSM model, or FBP, as a prior.

The approach to motif-finding demonstrated by SOMBRERO may be further improved by integration with phylogenetic footprinting methods, or perhaps through the use of alternative representations of binding sites (e.g. those described by Barash *et al*. (2003)). Other improvements may come through the use of different PSSM scoring functions, such as the recently described ALLR measure (Wang and Stormo, 2003). However, one obstacle preventing the adoption of SOM-based motif-finders remains the computational cost of the SOM algorithm. While SOMBRERO takes only 97 s to completely analyse a 500 bp sequence on one processor (using settings described in Methods), the analysis of 15 Kbp using a suitably scaled size SOM takes on average 57 min using 8 processors (when deployed on the SGI Origin 3800, see Fig. 3 for further running time information). Indeed, further parallelization and deployment on distributed computing resources may bring down the time taken for SOMBRERO to run on very large datasets to more acceptable levels.

As mentioned in the introduction, Sandelin and Wasserman described two applications for familial binding profiles, the second of which suggests that a collection of FBPs can be employed to predict the structural class of TFs that are likely to act through a newly discovered motif. The classification of novel motifs is dependent on there being a relevant FBP in the collection. The manual construction of FBPs in the Sandelin and Wasserman study is limited to those families that are well represented in the JASPAR database, and it would therefore seem that an automatic clustering of every known PSSM into familial models would be more suitable for automatically classifying novel motifs.

We do not expect that the BP-SOM algorithm described above would be the most suitable algorithm available for automatically classifying novel motifs. The output lattice does not clearly delineate families or suggest the relationships between two nodes, the number of clusters (or FBPs) found by the SOM is dependent on the lattice size, and empty nodes representing

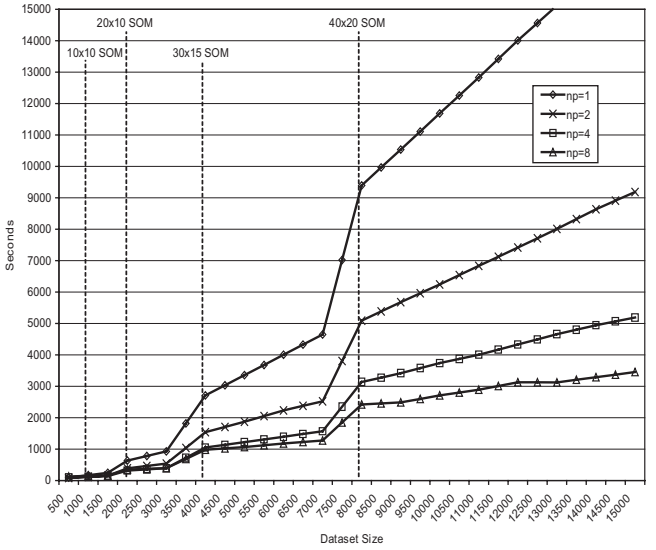


Fig. 3. Timing information for SOMBRERO running with various numbers of processors (np) on different dataset sizes. The points at which different SOM sizes are used are shown using dotted lines.

no PSSM family often remain on the lattice at the end of training. While these disadvantages are not important in terms of biasing a motif-finder as outlined here, they would affect the usefulness of the algorithm for classification purposes. However, many other automatic clustering algorithms should be suitable for classification applications, or for exploring the relationships between FBPs. We are currently evaluating the application of hierarchical clustering algorithms in this domain.

ACKNOWLEDGEMENTS

S.M. would like to thank the Department of Human Genetics, GSPH, University of Pittsburgh for the hospitality during his visit. S.M. was supported by an Embark postgraduate scholarship from the Irish Research Council for Science, Engineering and Technology. P.V.B. was supported by NSF grant MCB0316255.

REFERENCES

Abe,T., Kanaya,S., Kinouchi,M., Ichiba,Y., Kozuki,T. and Ikemura,T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res.*, **13**, 693–702.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB)*, Berline, Germany, ACM Press, pp. 28–37.

Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.

Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *D. melanogaster*. *Bioinformatics*, **21**, 1747–1749.

Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.

GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.

Gupta,M. and Liu,J.S. (2003) Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Stat. Assoc.*, **98**, 55–66.

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*. *J. Mol. Biol.*, **296**, 1205–1214.

Kohonen,T. (1995) *Self-Organizing Maps*. Springer-Verlag, Berlin.

Kohonen,T. and Somervuo,P. (2002) How to make large self-organizing maps for nonvectorial data. *Neural Netw.*, **15**, 945–952.

Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.

Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.

Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.

Mahony,S., Hendrix,D., Golden,A., Smith,T.J. and Rokhsar,D.S. (2005) Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, **21**, 1807–1814.

Mahony,S., McInerney,J.O., Smith,T.J. and Golden,A. (2004) Gene prediction using the self-organizing map: automatic generation of multiple gene models. *BMC Bioinformatics*, **5**, 23.

McCue,L.A., Thompson,W., Carmack,C.S. and Lawrence,C.E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.*, **12**, 1523–1532.

Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 269–278.

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

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32** (Database issue), D91–D94.

Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.

Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **30**, 5549–5560.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.

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. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol.*, **23**, 137–144.

Wang,H.C., Badger,J., Kearney,P. and Li,M. (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol. Biol. Evol.*, **18**, 792–800.

Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.

Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467–478.

Xing,E.P. and Karp,R.M. (2004) MotifPrototyper: a Bayesian profile model for motif families. *Proc. Natl Acad. Sci. USA*, **101**, 10523–10528.

Xing,E.P., Wu,W., Jordan,M.I. and Karp,R.M. (2004) Logos: a modular bayesian model for *de novo* motif detection. *J. Bioinform. Comput. Biol.*, **2**, 127–154.