SOMBRERO: INTEGRATING SELF-ORGANIZING NEURAL NETWORKS IN THE SEARCH FOR DNA BINDING MOTIFS

Shaun Mahony[1], Panayiotis V. Benos[2], Terry J. Smith[1], Aaron Golden[1,3]
[1]National Centre for Biomedical Engineering Science, NUI Galway, Galway, Ireland.
[2]Department of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[3]Department of Information Technology, NUI Galway, Galway, Ireland
email: {shaun.mahony, terry.smith, aaron.golden}@nuigalway.ie and benos@pitt.edu

Abstract – *Identification of the short DNA sequence motifs that serve as binding domains for transcription factors continues to be a challenging problem in computational biology. Currently popular methods of motif discovery are based on unsupervised techniques from the statistical learning theory literature. We present here a working prototype of a neural networks based system that aims to tackle the DNA regulatory motif identification problem. The system consists of three modules, the core module being a SOM-based motif-finder named SOMBRERO. The motif-finder is integrated in the prototype with a SOM-based pre-processing method that initialises SOMBRERO with relevant biological knowledge, as well as a self-organizing tree method that helps the user to interpret SOMBRERO's results. The system is demonstrated here using various datasets.*

Key words – Transcription factor binding sites, motif-finding, Self-Organizing Map, Self-Organizing Tree.

1 Introduction

Given a collection of DNA regions that are believed to contain common regulatory elements, computational methods that aim to find transcription factor binding sites (TFBSs) typically proceed by identifying short DNA sequence "motifs" that are statistically overrepresented in the input. The motif identification problem is notoriously difficult, however, as motifs are short signals (6-20bp long) that are hidden amongst a great amount of statistical noise (promoter regions are typically thousands of base pairs long). No information usually exists as to the number of individual TFBSs contained in the input sample, and sequence variation exists such that one TFBS can be quite dissimilar from another of the same type. More than one transcription factor (TF) may have binding sites in the input sample, so it is not typically known even how many distinct motifs we expect to find.

Despite the difficulties, numerous motif prediction techniques have become available over the past few years. Many methods are based on statistical learning theory methods such as expectation-maximisation (e.g. MEME [1]) and Gibbs sampling (e.g. AlignACE [2], Co-Bind [3] and BioProspector [4]). Such methods work through maximum likelihood parameter estimation of the motif model. Neural networks have rarely been applied to the motif-identification problem, one notable exception being ANN-Spec [5], where a Perceptron was combined with a Gibbs-sampler to increase the specificity of the estimated motif models. Alternative motif identification methods have also been proposed, including word enumeration, winnowing, and dictionary construction based methods [6-8].

An alternative approach to the motif identification problem can be defined by phrasing it as a clustering problem. For example, instead of defining the problem in terms of two models (the motif and the background) whose parameters have to be estimated by expectation maximisation or other such methods, consider the input sequence collection as a set of short overlapping substrings which may be clustered into a number of bins according to sequence similarity. After clustering, each bin would contain an alignment of similar substrings and therefore a motif. Given a large number of bins, a correspondingly large number of motifs would be found by the clustering approach. The vast majority of these motifs would not be TFBS motifs, and would instead be due to the background mutation patterns of the genome. Given an appropriate background model, TFBS motifs can be distinguished from motifs that represent background noise.

One unsupervised clustering algorithm suitable to our alternative phrasing of the motif-identification problem is the Self-Organizing Map (SOM) [9]. We have previously shown that the SOM can be applied to the motif identification problem, and the SOMBRERO (Self-Organizing Map for Biological Regulatory Element Recognition and Ordering) framework resulted [10]. In the previous publication, we demonstrated that SOMBRERO's approach to simultaneously characterising a complete set of motifs for a given dataset has advantages over traditional approaches. Specifically, the self-organized approach to motif identification helps to separate weak motif signals from large datasets, and improved motif detection performance in real biological datasets is observed [10].

In the current work, we show how two distinct self-organizing neural networks can be usefully integrated with SOMBRERO. One of the additional modules, based on the SOM, seeks to initialise the motif-finder with relevant biological information on known TFBSs, thus turning the original unsupervised approach into a semi-unsupervised one. Another subsystem, based on a self-organizing tree algorithm, helps to resolve differences between the motifs that have been found by the core motif-finder.

2 Methods

2.1 Position Specific Scoring Matrices

It is usually possible for TFs to bind to a set of related sequences that share some highly conserved positions as well as some more stochastically determined positions. The set of related sequences can be approximately represented by a consensus sequence, or more accurately by a position specific scoring matrix (PSSM). In describing an alignment of sequences, a PSSM gives the frequency of each base at each position in the alignment (see Figure 1). Similarity between a DNA substring and a PSSM is provided by a log-likelihood ratio score, *S(x)*, defined as:

$$S(x) = \sum_{b=A}^{T} \sum_{i=1}^{\ell} x_{ib} \log\left(\frac{f_{ib}}{p_b} \right) \tag{1}$$

where p_b is the background probability for base *b* and x_{ib}, a position in the indicator matrix for the string *x*, is 1 if base *b* is at position *i* of the string and 0 otherwise. A high score *S(x)* indicates that the string *x* is more similar to the motif characterised by the PSSM *f* than to the background model.

2.2 SOMBRERO

SOMBRERO is based on the SOM, whose general structure is a two-dimensional (2-D) lattice of interconnected nodes. In SOMBRERO, PSSMs are embedded as models at each node on the SOM grid. The motif discovery problem aims to find over-represented features of length ℓ in an input dataset of DNA sequences. SOMBRERO, therefore, aims to align similar ℓ-mer sequences at the SOM nodes. With this aim in mind, the training algorithm proceeds as follows:

1. An *X* x *Y* grid of nodes is created, and the coordinates of the nodes are denoted by z = (z_1, z_2). Each node contains a PSSM model f^z and a count matrix c^z that contains the number of base *b* at each position *i* in the current alignment.

2. A length ℓ is chosen, typically between 8 and 20, and the input sequences are segmented into every overlapping ℓ-mer (x_j , $j=1,\ldots,N$). The PSSM models are initialised using an ordered gradient random initialisation [10].
3. Each x_j is assigned to the node with the maximum likelihood, i.e. the highest score $S_{\mathbf{z}}(x_j)$.
4. Update step:
    - 4.1. The count matrix $c^{\mathbf{z}}$ is updated for each node, according to the current set of ℓ-mers aligned at the node.
    - 4.2. New models are generated by augmenting the profile matrix:

$$f_{ib}^{z} = \frac{\sum_{z'} \Phi\left(\left|z-z'\right|\right) c_{ib}^{z'} + \beta p_b}{\sum_{b'} \sum_{z'} \Phi\left(\left|z-z'\right|\right) c_{ib}^{z'} + \beta} \tag{2}$$

where p_b is the background probability model, β is a small scaling factor that helps to avoid zero probabilities, and $\Phi(|\mathbf{z} - \mathbf{z}'|)$ is a neighbourhood function that defines the proportion that a node will contribute to another node that is a distance $|\mathbf{z} - \mathbf{z}'|$ away on the SOM. For our purposes, the Gaussian neighbourhood function

$$\Phi\left(\left|z-z'\right|\right) = e^{-[(z_1-z_1')^2+(z_2-z_2')^2]/\gamma} \tag{3}$$

is used. Here the term γ is a measure of the sharpness of the neighbourhood function and is defined as $\gamma \equiv 1/\log(\delta)$ so that adjacent nodes will contribute $1/\delta$ of their counts to each other. In practice, δ ranges from 4 to 15 over the course of training. Thus, the contributions from $f_{ib}^{\mathbf{z}}$ to the counts of neighbouring nodes initially strongly enforce the similarity of nearby nodes, and end up contributing little at the end of training.

5. Training repeats from step 3 until convergence (defined here as 100 cycles). Once convergence is reached, each string x_j is assigned to its most similar node. In the case where two or more strings at a given node are overlapping strings in the input sequences, only the string with the larger $S_{\mathbf{z}}(x_j)$, is kept. At this point, each node will have a PSSM motif in its final state as well as a list of ℓ-mers that contributed to the motif's construction.
6. In practice, various motifs of different lengths can exist in a single dataset. Therefore, separate SOMs can be trained from step 2 for various length ℓs. In our application, separate SOMs are typically trained across all even lengths between 8 and 20.

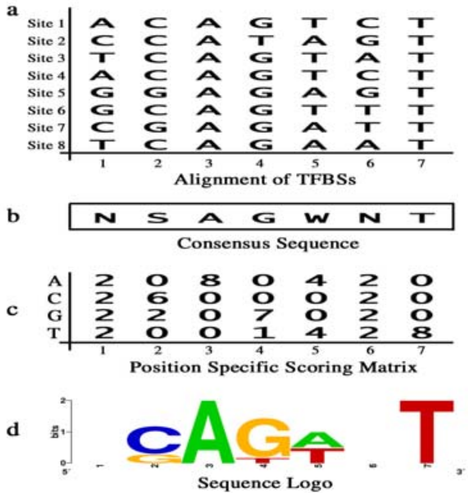


Figure 1. a) an alignment of binding sites, **b)** a consensus sequence representation of the alignment, where W = {A or T}, S = {C or G} and N represents any base, **c)** a PSSM representation, and **d)** a sequence logo representation, where the information content of each position is multiplied by the relative frequency of the letters to give letter height.

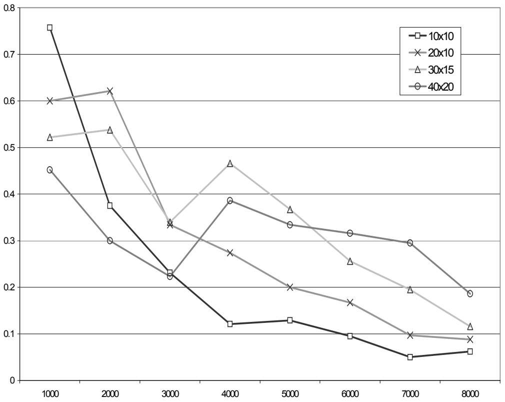


Figure 2. The effect on SOMBRERO performance of varying the SOM size as a factor of input dataset size.

7. Post processing steps:
   7.1. Significant features are distinguished from those that would be expected due to chance. A 3rd order Markov chain model of the relevant background is used to generate random datasets, and these sets are used to find the expected number of occurrences of each motif, thus yielding z-scores for each node's motif.
   7.2. Repetitive chains of DNA that exist throughout the genome may sometimes seem to be significant motifs, but are in fact uninteresting from the viewpoint of binding motif identification. Such repetitive motifs are filtered using a motif complexity score (given in [10]). Complexity here refers to a measure of the diversity of bases appearing in a PSSM.

In order to investigate the effect of varying the SOM lattice size, various sized SOMs were tested on artificial sequence datasets of varying lengths. Each sequence dataset was created using a 3rd order Markov model of *E. coli* intergenic sequences, and 10 instances of the gal4 binding motif were placed randomly in each dataset. The tests were run ten times to generate the average performance ratios displayed in Figure 2. Performance is defined here as $|K \cap P| / |K \cup P|$, where K is the set of known motif sites and P is the set of predicted motif sites. The trend of getting better performance in larger datasets by increasing the SOM lattice size is clear in Figure 2. From this test and others, it was found that the optimum SOM performance can be obtained by keeping the ratio of lattice nodes to input dataset size in the order of one node to 10 base pairs.

2.3 Initialising SOMBRERO using a PSSM SOM

In the original description of the SOMBRERO algorithm, randomised and ordered SOM lattice initialisation strategies were explored [10]. Although no significant difference in motif-finding accuracy was observed between the two strategies, the ordered initialisation, where the PSSMs in each corner of the lattice were biased towards a particular base (and gradients of preference existed in other nodes), was chosen for the smoothness it introduced into the SOM training procedure.

An alternative approach to SOMBRERO initialisation has since been developed [11]. Many known binding motifs exist in databases such as TRANSFAC (www.gene-regulation.com), and the binding preferences displayed in these motifs are not randomly distributed. For example, if two transcription factors are related evolutionarily, it is highly likely that their DNA binding motifs also display some similarity. The challenge is to incorporate such knowledge into methods that aim to find novel motifs in a set of sequences.

It has previously been demonstrated that incorporating the binding preferences of an entire family of related motifs as a "biasing prior" for a Gibbs-sampler based motif finder improves the detection of motifs related to the given family [12]. Knowledge of familial membership of an unknown motif is rarely available, however, and incorporating an incorrect prior has a detrimental effect on motif-finding performance. Current motif-finding methods can only incorporate a single prior in a given motif-finding run, and therefore the choice of biasing prior is critical. However, the SOMBRERO lattice contains many PSSM models, and thus the opportunity exists for multiple priors to be used to initialise the lattice. One effective way to order a set of known PSSMs into a structure suitable for initialising and biasing SOMBRERO is to train another SOM on the set of PSSMs and use the final node states from that SOM as the initial SOMBRERO node states.

When training a SOM on a set of PSSMs, PSSMs are again used as SOM node models. As distinct from the SOMBRERO algorithm, PSSM matrices, and not ℓ-mers, are to be clustered at the models. Clustering matrices on the SOM requires a matrix-matrix similarity measure. Pietrokovski's [13] methods for aligning two PSSMs are used in this study. Under this schema, Pearson's correlation coefficient is used for column-to-column comparisons, and a modified Smith-Waterman algorithm [14] is used to find optimal (gapless) local alignments of PSSM pairs. In order to compare alignments of different widths, the method for the calculation of empirical *p*-values described by Sandelin & Wasserman [12] is followed exactly. The method involves extensive analysis with a set of 10,000 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 [15]. The training algorithm for the SOM of PSSMs proceeds as follows:

1. The PSSM SOM lattice size is chosen as equal to the size required by the SOMBRERO grid, and each node model m_j is initialised as a PSSM with random values.
2. For each training set 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 above.

    2.2. The node w whose model m_w has the best p-value score to x_i is selected.

3. Update step:

    3.1. At each node j, all clustered members are aligned to give the weighted alignment matrix A_j. The weight (Z_v) of each member is calculated by the average p-value (p_v) obtained in comparisons of profile v to all other members of the same node: $Z_v = 1 - p_v$. 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:

$$m_j(t+1) = \sum_{i=1}^{N} align(x_{i,k} \cdot Z_{i,k} \cdot e^{-|j-k|^2/\gamma}) \tag{4}$$

where *align()* is a function that aligns the columns of each x_i (clustered at node k) to the relevant alignment positioning template at node j, and $|j\text{-}k|$ is the distance on the SOM grid between nodes j and k.. The Gaussian sharpness factor, γ, is defined as before, but here δ ranges from 4 to 30 during training. The length of the new model depends on the quality of the alignment. Flanking columns with low information content (<0.4 bits) are excluded from the new model, to a minimum model length of 8 columns. Finally, each m_j is normalised.

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

The algorithm results in a grid of PSSMs that can be used as the initial states for a SOMBRERO grid of equal size. The effect of varying the initialisation strategy on SOMBRERO's performance is illustrated in Figure 3. For this figure, two sets of artificial sequence data were created; one containing variable instances on the E4BP4 motif, and another containing variable instances of the CSRE motif. The motif instances were randomly embedded in sequences of various lengths. Three initialisation strategies were tested; the original ordered initialisation, a random initialisation and an initialisation using a PSSM SOM that has been previously trained on 257 mammalian PSSMs (including E4BP4, but not CSRE). From Figure 3, it may be seen that the latter initialisation improves motif-finding performance significantly in the E4BP4 set, especially in longer sequences. Performance is not significantly affected by using a PSSM SOM in the CSRE set, as CSRE was not in the dataset used to train the PSSM SOM.

2.4 Displaying relationships between discovered motifs using the Self-Organizing Tree Algorithm

As a consequence of SOMBRERO repeating the motif search over various values of ℓ, slightly different instances of the same motif may be discovered and reported as distinct motifs. In order to point out similarities between the discovered motifs to the user, a third subsystem is necessary. The Self-Organizing Tree Algorithm (SOTA) was first described by Dopazo & Carazo as an alternative means of automatically constructing a phylogenetic tree for a set of protein sequences [16]. The topology of the SOTA neural network takes the form of a binary tree. The tree begins with 2 external elements, denoted as cells, connected by an ancestor, named a node. Training proceeds similarly to the SOM algorithm, but at the end of a training cycle the tree grows by splitting one cell. The tree stops growing when a predefined threshold has been reached, or when every cell has a single datapoint clustered within (as in our usage). SOTA is based on Fritzke's growing self-organizing network concept [17]. Many other forms of growing self-organizing maps have been described, but SOTA's binary tree topology was deemed most suitable for this application.

In the current system, we apply SOTA to the hierarchical clustering of a set of the most over-represented PSSMs outputted by SOMBRERO. The SOTA nodes and cells each contain a PSSM that evolves over the training period to represent a PSSM or set of PSSMs from the input dataset. Other than the change of neural network topology and associated cell-growing mechanism, the PSSM SOTA methodology (including the PSSM alignment method) is similar to the PSSM SOM described above, with the following specifics:

1. Two cells, and a connecting node, are initialised as random value PSSMs m_j.
2. For each training set PSSM, x_i, (x_i, $i=1,\ldots,N$):

    2.1. x_i is aligned to every cell on the tree (m_j) using the alignment method described above.

    2.2. The node w whose model m_w has the best p-value score to x_i is selected.

3. Update step (only cells and immediate ancestors):

   3.1. An alignment matrix A_j is constructed at each cell in the same manner as that described for the PSSM SOM.
   3.2. New models are generated according to:

$$m_j(t+1) = m_j(t) + \sum\nolimits_i^n align(x_{i,k} \cdot \eta_j) \tag{5}$$

   where the notation follows that of the PSSM SOM, and the learning rate $\eta_j = \alpha_i (1 - t / M_t)$, where α_{sister}=1/2 and α_{mother}=1/8.

4. The training process repeats from step 2 until M_t cycles are reached (M_t = 50).
5. Growing phase: If the algorithm has not yet converged, the cell with the lowest resource value (Z_i, defined above) is split, giving rise to two (initially) identical descendants.
6. Training repeats from step 2 until convergence. Convergence is defined here as the point where every cell contains one and only one PSSM, although training can be stopped at any point in the growth of the tree.

The resulting tree structure can be displayed to the user, and this allows the visualisation of the distinct motifs discovered by SOMBRERO.

3. Results

The functionality of the complete SOMBRERO motif-finding system is briefly summarised in the demonstration outlined in Box 1. In this demonstration, SOMBRERO aims to identify motifs in an artificial dataset, generated using a 3rd order Markov model of yeast intergenic DNA. Within the artificial dataset are implanted 10 TFBSs for each of three TFs: GAL4, NF-κβ and CREB. SOMBRERO's grid is initialised using a PSSM SOM that has been previously trained on a collection of 257 mammalian PSSMs. The collection includes the NF-κβ and CREB PSSMs, but not the GAL4 motif. As can be seen in Box 1, SOMBRERO identifies multiple motifs that match the known GAL4, NF-κβ and CREB PSSMs.

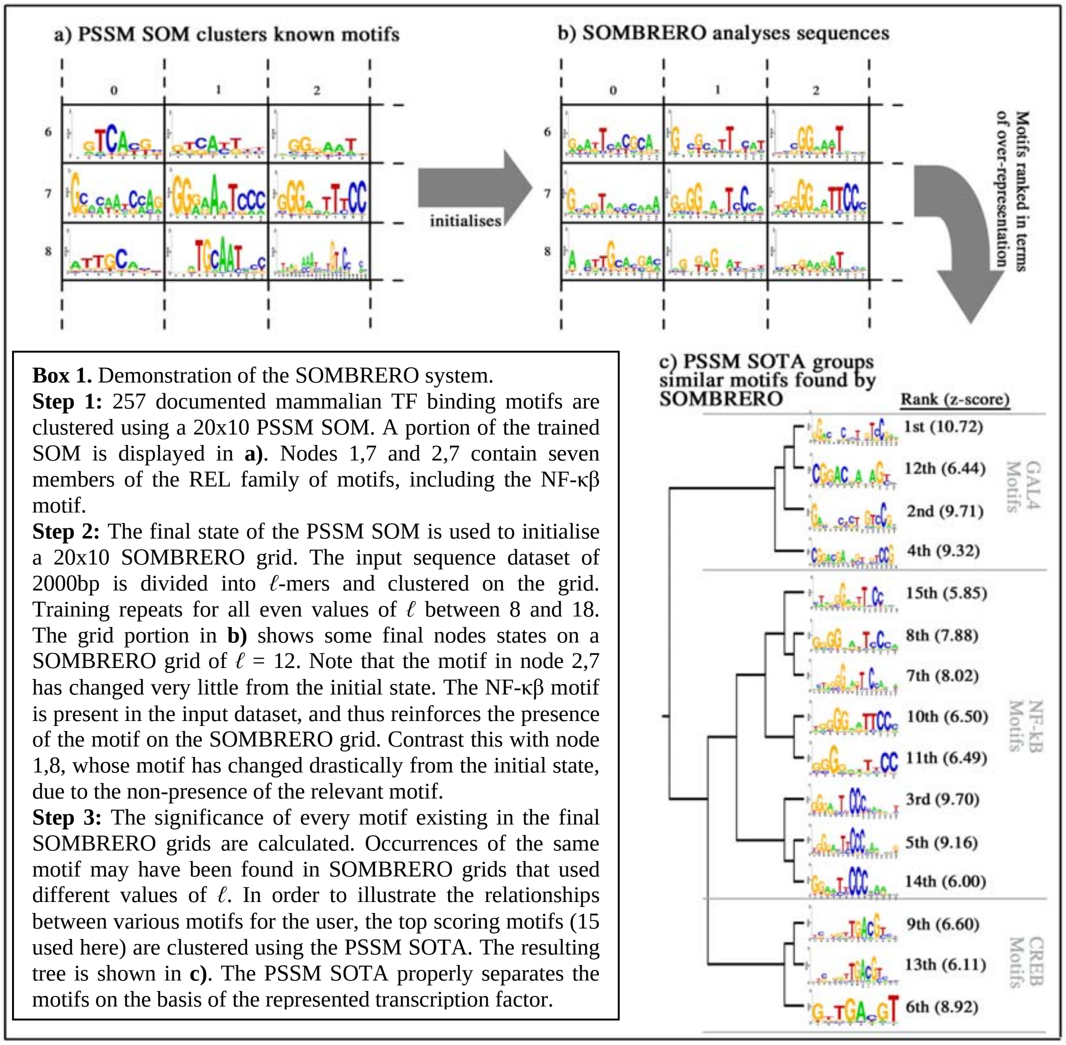
As an example of SOMBRERO's performance in real genomic datasets, 10 datasets from the yeast genome are used to evaluate the performance of SOMBRERO with and without a prior initialisation in comparison with the popular motif-finders MEME [1] and AlignACE [2]. Each dataset contains a number of instances (given by the *sites* column) of a particular yeast transcription factor binding motif (as denoted by the name of the dataset). The prior initialisation refers to a SOMBRERO run that has been initialised using a PSSM SOM trained on the entire set of known yeast motifs contained in the SCPD database. The results in each dataset are described in Table 1 in terms of false negative rates (*FN*), false positive rates (*FP*) and performance (*Perf*). It can be seen from the table that the use of the prior initialisation allows SOMBRERO to gain the best performance rate in 8 of the 10 datasets.

4. Conclusion

We have described an effective prototype of a neural networks based software pipeline that has the ability to automatically identify multiple DNA binding motifs in biological sequence datasets. The SOMBRERO system is the only existing motif-finding software program that can effectively incorporate a complete set of known TFBS PSSMs as prior knowledge, thereby offering improved performance. The SOMBRERO system is currently freely available as a standalone application for a variety of different systems (see http://bioinf.nuigalway.ie/sombrero for download). A web-based interface for the integrated system is currently being developed that will allow users to submit their data for analysis, and results will be returned on completion. One issue with the SOM-based approach to motif-finding is the computational cost of the SOMBRERO algorithm. Parallelisation has alleviated this problem somewhat (see Figure 4 for timing information on a SGI Origin 3800), but scaling up the SOMBRERO system to allow for the analysis of very large sequence datasets will require further optimisation and deployment on distributed computing resources.

Acknowledgements

S.M. was supported by an Embark postgraduate scholarship from the Irish Research Council for Science, Engineering and Technology.



Box 1. Demonstration of the SOMBRERO system.

Step 1: 257 documented mammalian TF binding motifs are clustered using a 20x10 PSSM SOM. A portion of the trained SOM is displayed in **a)**. Nodes 1,7 and 2,7 contain seven members of the REL family of motifs, including the NF-κβ motif.

Step 2: The final state of the PSSM SOM is used to initialise a 20x10 SOMBRERO grid. The input sequence dataset of 2000bp is divided into ℓ-mers and clustered on the grid. Training repeats for all even values of ℓ between 8 and 18. The grid portion in **b)** shows some final nodes states on a SOMBRERO grid of ℓ = 12. Note that the motif in node 2,7 has changed very little from the initial state. The NF-κβ motif is present in the input dataset, and thus reinforces the presence of the motif on the SOMBRERO grid. Contrast this with node 1,8, whose motif has changed drastically from the initial state, due to the non-presence of the relevant motif.

Step 3: The significance of every motif existing in the final SOMBRERO grids are calculated. Occurrences of the same motif may have been found in SOMBRERO grids that used different values of ℓ. In order to illustrate the relationships between various motifs for the user, the top scoring motifs (15 used here) are clustered using the PSSM SOTA. The resulting tree is shown in **c)**. The PSSM SOTA properly separates the motifs on the basis of the represented transcription factor.

Table 1. Comparison of motif detectors on 10 yeast promoter sequence datasets. The best performance rate in each dataset is highlighted in **bold**.

		SOMBRERO (original initialisation)			**SOMBRERO (with prior)**			**MEME**			**AlignACE**		
	sites	**FN**	**FP**	**Perf**	**FN**	**FP**	**Perf**	**FN**	**FP**	**Perf**	**FN**	**FP**	**Perf**
abf1	20	0.45	0.56	*0.324*	0.40	0.29	***0.480***	0.55	0.18	*0.409*	0.50	0.38	*0.385*
csre	4	0.25	0.73	***0.250***	0.00	0.75	***0.250***	0.50	0.67	***0.250***	0.25	0.82	*0.167*
gal4	14	0.07	0.24	*0.722*	0.07	0.07	***0.867***	0.29	0.17	*0.625*	0.21	0.08	*0.733*
gcn1	25	0.60	0.29	*0.345*	0.44	0.33	***0.438***	0.92	0.80	*0.061*	0.60	0.44	*0.303*
gcr1	9	0.22	0.69	*0.285*	0.00	0.41	***0.588***	0.44	0.44	*0.385*	0.33	0.63	*0.316*
hstf	9	0.11	0.57	*0.407*	0.11	0.53	***0.444***	0.33	0.75	*0.222*	0.11	0.56	*0.421*
mat	13	0.31	0.25	*0.563*	0.15	0.27	*0.647*	0.15	0.27	*0.647*	0.31	0.00	***0.692***
mcb	12	0.08	0.65	*0.344*	0.08	0.31	*0.647*	0.25	0.25	*0.600*	0.08	0.08	***0.846***
mig1	10	0.20	0.68	*0.296*	0.10	0.47	***0.500***	1.00	1.00	*0.000*	0.90	0.91	*0.050*
pho2	6	0.50	0.91	*0.083*	0.33	0.80	***0.182***	1.00	1.00	*0.000*	1.00	1.00	*0.000*
Avg		0.32	0.61	*0.333*	0.22	0.42	***0.500***	0.56	0.45	*0.323*	0.43	0.47	*0.379*

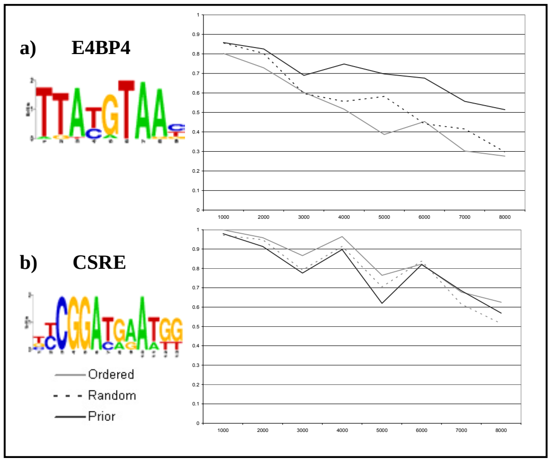


Figure 3. The effect on performance when initialising SOMBRERO with a PSSM SOM, **a)** when the motif being found is in the data clustered by the PSSM SOM and **b)** when it is not.

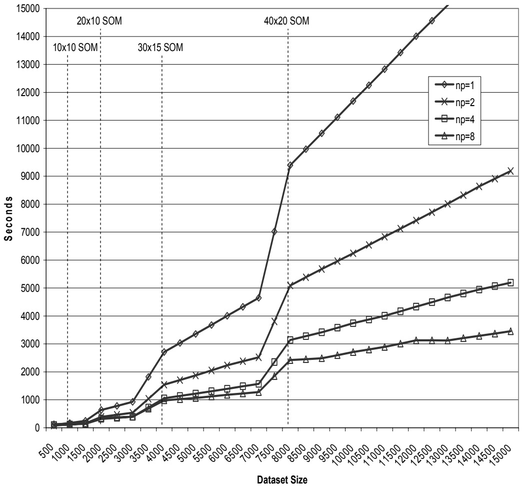


Figure 4. Timing information for SOMBRERO running with various numbers of processors (np) on different dataset sizes. The points at which different SOM sizes are used are shown using dotted lines.

References

[1] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proc Int Conf Intell Syst Mol Biol*, vol. 2, pp. 28-36, 1994.

[2] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae," *J Mol Biol*, vol. 296, pp. 1205-14, 2000.

[3] D. GuhaThakurta and G. D. Stormo, "Identifying target sites for cooperatively binding factors," *Bioinformatics*, vol. 17, pp. 608-21, 2001.

[4] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," *Pac Symp Biocomput*, pp. 127-38, 2001.

[5] C. T. Workman and G. D. Stormo, "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity," *Pac Symp Biocomput*, pp. 467-78, 2000.

[6] I. Rigoutsos and A. Floratos, "Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm," *Bioinformatics*, vol. 14, pp. 55-67, 1998.

[7] P. A. Pevzner and S. H. Sze, "Combinatorial approaches to finding subtle signals in DNA sequences," *Proc Int Conf Intell Syst Mol Biol*, vol. 8, pp. 269-78, 2000.

[8] M. Gupta and J. S. Liu, "Discovery of Conserved Sequence Patterns Using a Stochastic Dictionary Model," *Journal of the American Statistical Association*, vol. 98, pp. 55-66, 2003.

[9] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer-Verlag, 1995.

[10] S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D. S. Rokhsar, "Transcription factor binding site identification using the self-organizing map," *Bioinformatics*, vol. 21, pp. 1807-14, 2005.

[11] S. Mahony, A. Golden, T. J. Smith, and P. V. Benos, "Improved detection of DNA motifs using a self-organized clustering of familial binding profiles," *Bioinformatics*, vol. 21 Suppl 1, pp. i283-i291, 2005.

[12] A. Sandelin and W. W. Wasserman, "Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics," *J Mol Biol*, vol. 338, pp. 207-15, 2004.

[13] S. Pietrokovski, "Searching databases of conserved sequence regions by aligning protein multiple-alignments," *Nucleic Acids Res*, vol. 24, pp. 3836-45, 1996.

[14] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J Mol Biol*, vol. 147, pp. 195-7, 1981.

[15] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Res*, vol. 32 Database issue, pp. D91-4, 2004.

[16] J. Dopazo and J. M. Carazo, "Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree," *J Mol Evol*, vol. 44, pp. 226-33, 1997.

[17] B. Fritzke, "Growing cell-structures - a self-organizing network for unsupervised and supervised learning," *Neural Networks*, vol. 7, pp. 1141-1160, 1994.