BMC Bioinformatics

RESEARCH Open Access

SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model

Nung Kion Lee, Dianhui Wang*

From The Ninth Asia Pacific Bioinformatics Conference (APBC 2011)
Inchon, Korea. 11-14 January 2011

Abstract

Background: Discrimination of transcription factor binding sites (TFBS) from background sequences plays a key role in computational motif discovery. Current clustering based algorithms employ homogeneous model for problem solving, which assumes that motifs and background signals can be equivalently characterized. This assumption has some limitations because both sequence signals have distinct properties.

Results: This paper aims to develop a Self-Organizing Map (SOM) based clustering algorithm for extracting binding sites in DNA sequences. Our framework is based on a novel intra-node soft competitive procedure to achieve maximum discrimination of motifs from background signals in datasets. The intra-node competition is based on an adaptive weighting technique on two different signal models to better represent these two classes of signals. Using several real and artificial datasets, we compared our proposed method with several motif discovery tools. Compared to SOMBRERO, a state-of-the-art SOM based motif discovery tool, it is found that our algorithm can achieve significant improvements in the average precision rates (i.e., about 27%) on the real datasets without compromising its sensitivity. Our method also performed favourably comparing against other motif discovery tools.

Conclusions: Motif discovery with model based clustering framework should consider the use of heterogeneous model to represent the two classes of signals in DNA sequences. Such heterogeneous model can achieve better signal discrimination compared to the homogeneous model.

Background

Identification of transcription factor binding sites (TFBS) is fundamental of understanding gene regulations. Binding sites or motif instances are typically 10 ~ 15*bp* in length and degenerated in some positions. They are often buried in a large amount of non-functional background sequences, which causes low signal-to-noise ratio. Hence, using computational approaches to discriminate motif signals from background signals has not always brought satisfactory results. Development of advanced tools is necessary for more accurate motif predictions.

* Correspondence: dh.wang@latrobe.edu.au
Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia

An essence of computational approaches for motif discovery is to search for motifs that are over-represented in the input sequences compared to the background sequences. Motif over-representation can be explained by the existence of segments that have been evolutionarily preserved due to their functional significance to gene regulation. Hence, appearances of motif instances are rather similar to each other despite having variability in some of their positions [1]. Two issues that are closely related to motif discovery problem are: (i) how to construct a model to represent the motifs and, (ii) how to define a suitable search strategy to find putative motifs from the solution space. Position-specific-scoring-matrix (PSSM) [2] and its variations are the most widely used motif model. This model defines the maximum-likelihood estimation on the probability of

nucleotide occurrences in every position of a motif. The motif search strategies can be local or global. Local search algorithms begin with an initial guess of a motif model and iteratively refine this model in the search space to maximize a certain criterion. Two examples of such algorithm are MEME [3] (expectation-maximization) and ALIGNACE (gipps-sampling) [4]). The local search approaches find out one motif at a time. Global search algorithms such as clustering based algorithms (e.g., SOMBRERO [5] and MISCLUSTER [6]) and genetic algorithms based algorithms (e.g., GAME [7] and iGAPK [8]) perform simultaneous searches for multiple candidate motifs by exploring the whole solution space.

In this paper, we aim to develop a SOM [9] based Extraction Algorithm (SOMEA) to discover over-represented motifs in DNA datasets. We seek to use SOM to project *k*-mers (i.e. a subsequence with length *k* of DNA sequences) onto a 2-dimensional (2D) lattice of nodes. Through this projection, input patterns (i.e., *k*-mers) with closely related features are projected onto the same or adjacent nodes on the map. Hence, the complex similarity relationships of the high-dimensional input sequence space become apparent on the map. Analysis of selected nodes, therefore, can reveal potential patterns (i.e., motifs) in the dataset.

Previous studies have applied a standard (e.g. [5,10,11]) and hierarchical (e.g. [12]) SOM to discover motifs in protein or DNA sequences. Those studies have made a common assumption that *the motif and the background signals can be analogously modeled by using a homogeneous node model.* This assumption is weak because the two classes of signals have some distinct statistical properties [13]. Hence, homogeneous model of these two signal classes may cause unfaithful map representation and produce clusters with many false positives. The traditional homogeneous modeling of two signal classes implies that, both signals are clusterable under a single type of model. However, mutational events are more rapid in background regions compared to binding regions, causing most of the nucleotide bases in background regions to be random. Thus, they have relatively lower clusterability compared against binding site regions [14]. Therefore, nodes' vectorial or string [9] based prototypes given by SOM in traditional tools, can represent motifs reasonably well, but do not well suit for background sequences since two different classes of signals are tried to be expressed through a homogeneous modeling. Hence, an alternative modeling approach, preferably a heterogeneous modeling approach, that takes these two signal properties in consideration is necessary.

In the development of SOMEA, we have proposed a hybrid node model to address some of the limitations of current SOM approaches. This hybrid node model is constituted by PSSM [2] and markov chain (MC) [15] model. These two model components perform soft-competition through an adaptive weighting scheme within a node to represent the mixture of signals in it. We hypothesized that, the fitness of each model's components (i.e., PSSM and MC) with respect to the sequences in a node, is a fuzzy indication of its signal class composition. Heuristic learning rules are proposed in this paper to adjust the model parameters during learning stage. We have evaluated our proposed SOMEA algorithm against several motif discovery tools using real and artificial DNA datasets. Results have shown that, our approach performs significantly better than a state-of-the-art clustering algorithm for motif discovery, named SOMBRERO [5].

Results and discussion

We now present an experimental evaluation of our SOMEA approach. We have used eight real datasets to compare the performances of our approach against SOMBRERO, MEME, ALIGNACE and WEEDER [16] in terms of sensitivity and specificity. Then, to evaluate SOMEA's ability in multiple motif discovery, we used five artificial datasets.

For performance quantification, we employed three measures i.e., precision(P), recall(R) and F-measure(F) [17]. They can be computed as: $P = TP/(TP + FP)$, $R = TP/Y$, $F = 2/(1/P + 1/R)$, where TP, FP, and Y are the numbers of true positives, false positives, and true binding sites in the dataset, respectively. We have considered a predicted site as a true positive if it is overlapped with a true binding site location by at least x nucleotides, where x is selected according to the length of the true motif consensus.

Performance on real datasets

The eight test datasets used in this experiments are composed of seven datasets used in [7] and a dataset collected from the Promoter Database of *S. cerevisiae* (SCPD) [18]. Each sequence contains at least one true binding site. These datasets consist of motifs from *Escherichia coli*(CRP), *homo sapiens*(ERE, MEF2, SRF, CREB, E2F, MYOD) and *S. cerevisiae*(GCN4).

SOMEA was run with map sizes that were arbitrarily selected between 10 × 10 to 20 × 20 depending on the size of the dataset. In each case, SOMEA was trained for 100 epochs with a motif length value in $[l - 3, l + 3]$, where l is the known motif consensus length. The top 10 highest ranked motifs according to their MAP score [19] were saved for evaluation purpose. A 3rd order markov chain model [15] was used to compute MAP score. The learning rate parameter was fixed at 0.005 in all the experiments. Whereas, the neighborhood function parameter value, σ was set at 3.0.

For WEEDER, we used the online tool [20] with the following options: sites might appear more than once, both strands, and normal or complete scan. The interesting motifs and their instances that scored at least 90 were used in the evaluation. SOMBRERO was run with the default map sizes and random initialization method. The standalone tool was downloaded from [21]. We evaluated all the "best-motifs" returned by the tools. MEME was run with the "any number" model option and minimum and maximum length value as discussed above. AlignACE was run online [22] with default arguments in most cases.

Table 1 shows recall (R), precision(P) and F-measure (F) rates for a ten run average for each program on the eight real DNA datasets. Comparison shows that, in terms of recall rates, SOMEA performs better than or equally to other tools in four(4) of the eight(8) datasets. Compared to SOMBRERO, SOMEA performs better in terms of recall rates in six(6) of the datasets. Also, SOMEA has higher precision rates in six(6) of the datasets and has better F-measure values in seven of the test datasets(except ERE). Notably, for the MEF2 dataset, SOMEA obtained a much higher precision rate (0.99 vs 0.22) in comparison with SOMBRERO. The performances on all datasets show that, SOMEA achieves significant improvements in the average precision rate (26.9%) and recall rate (13.8%) in comparison with SOMBRERO. This clearly shows that, SOMEA with heterogeneous node model can represent the *k*-mers distribution in DNA sequences better than the algorithms with homogeneous model.

It can be noticed that, SOMEA performance is comparable or better than ALIGNACE, MEME and WEEDER. For example, in terms of F-measure rates, SOMEA produces the best results for five of the eight datasets due to its higher precision rates (note that, both SOMEA and ALIGNACE achieve the same F-measure value for the CRP dataset). SOMEA's average F-measure value for all datasets (i.e. 0.72) is found better than MEME (0.65), ALIGNACE(0.69) and SOMBRERO(0.55) and equally good as WEEDER(0.72).

It should be noted that, the comparison results between programs cannot be completely fair as every program has its own strengths and weaknesses. For example, some programs might perform rather well for strong motifs; whereas some are designed to discover motifs with certain characteristics (e.g. gapped motifs). The nature of the datasets can be an influential factor to the success of each program. Therefore, the results reported here should serve only as reference.

Performance on artificial datasets with multiple planted motifs

Practically, we can often find multiple motifs in upstream region of a set of co-regulated genes. These motifs often work as cis-regulatory module to regulate gene expressions. Motif discovery programs should be able to return all of these potential motifs. Local search algorithms, such as MEME, perform a search for single motif at one time; whereas SOMEA and SOMBRERO search for all motifs simultaneously. It is interesting to compare these two strategies.

We have prepared five artificial DNA datasets generated from Annotated regulatory Binding Sites (ABS, v1.0) database [23]. Every DNA dataset has twenty(20) sequences (each with 500bp in length) with three planted real motifs. We run MEME, WEEDER, SOMEA and SOMBRERO five times on each dataset. We asked SOMEA and MEME to return the top 20 motifs for the evaluation purposes. Again, we evaluate all best motifs returned by WEEDER and SOMBRERO.

Table 2 shows the results of comparison between the four algorithms. Overall, SOMEA has the best recall rates in seven(7) of fifteen(15) of the motifs. However, such higher recall rates come at the price of having lower precision rates compared to MEME and WEEDER. Compared to SOMBRERO, SOMEA performs significantly better in most of the datasets in all performance measures. For example, in terms of recall rates, SOMEA is higher in ten(10) of the motifs; whereas, in terms of F-measure values, SOMEA has

Table 1 Evaluation results with comparisons

	SOMEA			SOMBRERO			MEME			ALIGNACE			WEEDER		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
CRP	**0.91**	0.89	**0.9**	0.83	0.43	0.56	0.59	0.88	0.69	0.83	0.98	**0.9**	0.75	0.83	0.79
GCN4	0.69	0.45	0.54	**0.8**	0.41	0.53	0.52	0.52	0.52	0.61	0.62	0.6	0.64	**0.87**	**0.73**
ERE	0.74	0.58	0.65	0.8	0.59	0.67	0.72	**0.82**	0.77	0.75	0.77	0.76	**0.76**	0.54	0.63
MEF2	0.81	**0.99**	**0.89**	0.35	0.22	0.27	**0.92**	0.8	0.85	0.86	0.87	0.86	0.88	0.88	0.88
SRF	0.84	0.74	**0.79**	0.67	**0.83**	0.74	**0.87**	0.72	**0.79**	0.83	0.71	0.77	0.83	0.71	0.76
CREB	**0.89**	0.67	**0.77**	0.83	0.43	0.56	0.59	**0.88**	0.69	0.52	0.66	0.57	0.79	0.71	0.75
E2F	**0.82**	0.64	0.71	0.76	0.67	0.71	0.68	0.64	0.65	0.75	**0.68**	**0.71**	0.89	0.67	**0.76**
MYOD	**0.66**	0.39	**0.49**	0.5	0.32	0.39	0.23	0.38	0.27	0.34	0.31	0.32	0.43	**0.5**	0.46
Average	**0.80**	0.67	**0.72**	0.69	0.49	0.55	0.64	**0.71**	0.65	0.69	0.70	0.69	0.75	**0.71**	0.72

Table 2 Evaluation results with comparisons for multiple motifs datasets

		SOMEA			SOMBRERO			MEME			WEEDER			
		R	P	F	R	P	R	R	P	F	R	P	F	
Dataset1	CREB	0.43	0.26	**0.33**	**0.44**	0.26	**0.33**	0.20	**1.00**	**0.33**	0.00	0.00	0.00	
	MYOD	**0.48**	**0.23**	**0.31**	0.20	0.08	0.11	0.00	0.00	0.00	0.00	0.00	0.00	
	TBP	**0.36**	0.21	**0.26**	0.20	0.12	0.15	0.07	**0.50**	0.12	0.00	0.00	0.00	
	Avg	**0.42**	0.23	**0.30**	0.28	0.15	0.20	0.09	**0.50**	0.15	0.00	0.00	0.00	
	Dataset2	NFAT	0.39	0.27	0.31	0.36	0.21	0.26	**0.44**	**0.78**	**0.56**	0.00	0.00	0.00
		HNF4	0.57	0.40	0.47	**0.63**	0.39	0.48	0.60	0.82	**0.69**	0.40	**1.00**	0.57
SP1		0.50	0.53	**0.50**	**0.53**	0.35	0.42	0.38	**0.54**	0.44	0.00	0.00	0.00	
Avg		0.49	0.40	0.43	**0.51**	0.32	0.39	0.47	**0.71**	**0.56**	0.13	0.33	0.19	
Dataset3		CAAT	**0.43**	0.21	0.25	0.32	0.17	0.22	0.29	**0.80**	**0.42**	0.00	0.00	0.00
		SRF	**0.70**	0.40	**0.50**	0.59	0.28	0.38	0.29	**0.57**	0.38	0.00	0.00	0.00
	MEF2	0.79	0.45	0.57	0.65	0.31	0.27	**0.80**	0.57	**0.67**	0.27	**1.00**	0.42	
	Avg	**0.64**	0.35	0.44	0.52	0.25	0.29	0.46	**0.65**	**0.49**	0.09	0.33	0.14	
	Dataset4	USF	0.68	0.39	0.48	**0.73**	0.48	**0.57**	0.41	**0.88**	0.56	0.00	0.00	0.00
		HNF3B	**0.47**	0.25	**0.31**	0.26	0.13	0.17	0.15	**1.00**	0.27	0.00	0.00	0.00
NFKB		0.71	0.47	0.56	0.66	0.46	0.54	**0.80**	0.57	**0.67**	0.33	**1.00**	0.50	
Avg		**0.62**	0.37	0.45	0.55	0.36	0.43	0.45	**0.82**	**0.50**	0.11	0.33	0.17	
Dataset5		GATA3	**0.61**	0.37	0.46	0.49	0.33	0.36	0.40	0.75	0.52	0.40	**1.00**	**0.57**
		CMYC	0.74	0.47	0.57	**0.89**	0.70	0.84	0.75	**1.00**	**0.86**	0.19	0.75	0.30
	EGR1	**0.66**	0.36	0.47	0.47	0.26	0.33	0.64	**0.81**	**0.72**	0.00	0.00	0.00	
	Avg	**0.67**	0.40	0.50	0.62	0.43	0.51	0.60	**0.85**	**0.70**	0.20	0.58	0.29	

better results in twelve(12) of the motifs. Hence, it can be observed that, our SOMEA has better signal discrimination ability than SOMBRERO. MEME performs better than SOMEA in terms of average F-measure values in four of five of the datasets. Nonetheless, SOMEA has higher average recall rates in all of the datasets. WEEDER performs poorly in most of the test datasets most likely due to the inability of its scoring function to rank the true motifs highly when planted in the artificial sequences (see WEEDER manual [20]). In summary, both global and local search techniques perform equally well and each strategy has its own strengths and weaknesses. Coupling them could be a feasible approach to enhance motif discovery result.

It should be noted, there are some biases in the comparisons for two reasons. Firstly, both SOMEA and SOMBRERO are rather sensitive to the motif length parameter. As the motif consensuses in a dataset have different lengths, a single run with a fixed length value might not be suited for all motifs. On the contrary, MEME is able to find a length value that suits better each motif. Consequently, in some of SOMEA/SOMBRERO runs, some motifs might appear to be performed better in the experiments. Secondly, the lower precision rates for SOMEA and SOMBRERO could be explained by the fact that the optimal map sizes are not known. Improper map sizes can, to some extend, affect the results for multiple motif datasets.

Robustness analysis

We have conducted some analysis on the robustness of SOMEA with respect to different map sizes. We have computed recall, precision and F-measure analysis on SOMEA using the eight real datasets for map sizes 10 × 10, 15 × 15, and 20 × 20. Each dataset is run for five times and their average recall, precision, and F-measure is computed.

Table 3 shows the F-measure of eight datasets with different map sizes. It can be seen that, different map sizes affect the performance on the datasets. From the comparisons, it can be noted that the performance of SOMEA and SOMBRERO shows a similar trend. Their F-measure rates reach maximum for most datasets when the map size 15 × 15 is used. The map size 10 × 10 is too small to represent the *k*-mers distribution in the original space for all the datasets. For a smaller map size, naturally the average number of *k*-mers in each cluster increases, hence, the precision rates become lower. In contrast, for a larger map size (i.e., 20 × 20) the precision rates naturally become higher. However, the recall rates can be lower as the true binding site *k*-mers may suffer from sparse distribution among several nodes in the map.

The computational time of SOMEA is mainly imposed by three operations: a) finding winner node for each kmer; b) updating winner and its neighboring nodes models; and c) updating node model at the end of an

Table 3 Comparisons of performance with different map sizes

	SOMEA	SOMBRERO	SOMEA	SOMBRERO	SOMEA	SOMBRERO
	10 × 10		15 × 15		20 × 20	
CREB	0.70	0.41	0.76	0.67	0.72	0.67
CRP	0.81	0.71	0.66	0.71	0.58	0.52
E2F	0.58	0.73	0.69	0.63	0.72	0.67
ERE	0.53	0.42	0.66	0.60	0.61	0.74
GCN	0.41	0.44	0.51	0.52	0.58	0.60
MEF	0.68	0.92	0.91	0.80	0.82	0.44
MYOD	0.32	0.23	0.49	0.42	0.47	0.49
SRF	0.70	0.67	0.77	0.72	0.71	0.71

Showing is the average F-measure from five runs of eight real datasets using three map sizes 10 × 10,15 × 15 and 20 × 20.

epoch. The time complexity of SOMEA with map size $R \times C$ is $O((L \times R \times C) + (P \times L) + (R \times C))$, where L is the total length of DNA sequences and P is the size of neighborhood during k-mers assignment. Here, the $O(L \times R \times C)$ term is due to the computation of finding winning node for every k-mer; $(P \times L)$ operations are needed for the computation of updating the temporary model variables during the k-mers assignment stage; and $(R \times C)$ operations for updating the node models at the end of an epoch. Self-organizing map based algorithm is known to suffer from heavy computational time due to the global search to simultaneously discover all clusters. We have recorded the execution time of SOMEA for the eight real datasets and found that it has the highest average computational time of 1364s as compared to WEEDER (825s), SOMBRERO (326s), MEME (126.7s) and ALIGNACE (101s). The slower computational time of SOMEA compared to SOMBRERO is due to the fact that we have to update the ΔM_{pssm} and ΔM_{mc} parameters (see Methods) for the winner and its neighborhood during k-mers assignment (i.e. see Eqs (9) and (10)). In SOMBRERO, the update of node models only occurs at the end of an epoch. Also, some heuristic optimizations are included in it to reduce computational time. It can be observed that, current version of SOMEA requires slightly larger computational time, however, its better sensitivity and specificity performances can offer a good trade-off.

Conclusions

Motif discovery in DNA datasets is a challenging problem domain because of our lack of understanding of the nature of the data, and the mechanisms to which proteins recognize and interact with its binding sites are still perplexing to biologist. Hence, predicting binding sites by using computational algorithms is still far from satisfaction.

In this paper, we have proposed a SOM based Extraction Algorithm (SOMEA) for simultaneous identification of multiple-motifs in DNA dataset. We have made two main contributions in this work. Firstly, it is shown that, the use of node model that considers the distinct properties of the motif and background signals is helpful in mining DNA motifs. We have proposed a hybrid model that is composed of PSSM and MC model to better represent these two classes of signals. Secondly, it has been highlighted that, clustering based DNA motif mining requires some customizations in the clustering system design, as standard clustering frameworks may not be sufficient. In addition to these, we have proposed heuristic learning rules to update the node model's parameters during learning.

Many computational motif discovery algorithms have been proposed in the past decade. Like most of these algorithms, SOMEA shares some common challenges that require further investigation. The first is the scalability of the system for large scale dataset such as ChIP sequences. The scalability is the ability of a tool to maintain its prediction performances and efficiency while the size of the datasets increases. To the best of our knowledge, most motif discovery algorithms are not designed to handle large scale datasets. In a recent study [24], using ChIP datasets as benchmark, it is shown that local search techniques such as MEME does not scale well with the increase in dataset sizes. This finding is consistent to an early study by [25]. Currently, SOMEA is not proposed to handle large scale dataset either. However, it can potentially be used to reduce the sequence search space by pre-cluster sequences into lower-dimensional topological space. Then we can perform the motif searches in this lower-dimensional space instead of the original sequence space. It would be interesting to further investigate the feasibility of this search space reduction strategy to enable system scalability.

The second critical issue is the system's robustness, which relates to the ability of the pattern recognition system to maintain its performance with the changes of parameters and noise in the inputs [26]. Currently, the critical parameters for SOMEA are the map size and the motif length. From our experiences with SOMEA, we

found that setting improper map sizes have caused poorer performance. If the map sizes are two small, the precision(recall) rates might be poor(better); whereas if the map sizes are two large, opposite results are expected. Choosing a proper motif length value is important to reveal the true motif patterns. Setting improper length values caused motif discovery algorithms to return only partial motif consensus patterns. We can overcome this shortcoming by running the system with different length values. Through some analysis on the produced results from different runs, we will be able to reveal the true motif consensuses.

In conclusion, clustering biological sequences for motif discovery should consider the use of heterogeneous model to efficiently represent both motif and background signals. We have shown that, our proposed SOMEA using a heterogeneous model, can perform better in terms of sensitivity and specificity than the tools that use homogeneous model.

System and methods

Overview

The main idea of our SOMEA algorithm is to use a hybrid node model, where a model is heterogeneously composed of PSSM and MC. We assume each node on the map is a fuzzy composition between a motif signal and background noise. Since we do not have prior knowledge on the type of each node, we use a *soft competitive weighting scheme* for the two components (i.e., PSSM and MC) of each node model. We refer it as intra-node competition. Our framework design is inspired by the fact that, the two sequence classes (i.e. motif and background noise) in the DNA dataset have distinctive properties. Subsequently, it is necessary to represent them using appropriate signal's models.

SOMEA starts with converting the input DNA sequences (both strands) into a set of k-mers using k length window shifting through the sequence. Then, the size of the map is defined (user input) and nodes' model parameters are randomly initialized. Then, the following two learning steps are repeated for each input k-mer in the dataset:

1. **Inter-nodes competition:** to find the best matching unit (BMU) of current input k-mer K_j.

2. **Models updating:** update model parameters of the BMU including its topological neighborhood.

The two steps above define a recursive regression process [9], where the optimal models parameters are estimated by iteratively applying the k-mers to the system. After some training epochs, similar k-mers from supposing motif or background class are projected onto the same or adjacent nodes on the 2D grid map. The k-mers projected in the vicinity on the map, generally forming clusters. This implies the similarity of their respective features. Once the nodes' models have been stabilized, we can identify candidate motifs using a motif model evaluation metric.

Basic concepts and problem formulation

We first give some notations used in this paper, and then describe the SOMEA algorithm. Denoted by $D = \{S_1, S_2,..., S_N\}$, a DNA dataset with N sequences. Let a k-mer $K_i = (b_1 b_2 ... b_k)$ be a continuous subsequence of length k in a sequence, and $i = 1,..., Z$, with Z is the total number of k-mers in that sequence. For a sequence with length L, there are $L - k + 1$ number of k-mers can be produced using k length window shifting process.

We can represent a k-mer K as a $4 \times k$ matrix [27]. Let the matrix representation be $e(K) = [a_{ij}]_{4\times k}$, where $(a_{1j}, a_{2j}, a_{3j}, a_{4j}) = (A, C, G, T)$ and $j = 1,..., k$. The matrix has a column j representing certain nucleotide i at that position j in the k-mer.

A 2D SOM map is a lattice of $R \times C$ nodes, where R, C is the number of rows and columns respectively. Each node V_{ij}, $i = 1,..., R$ and $j = 1,..., C$, has a subset of k-mers assigned to it. For convenience, we use the notation V_l to represent a node, where $1 \le l \le (R \times C)$. The coordinate of a node V_l in the lattice is expressed as $z_l = (i, j)$. Then, each node V_l has a parameterized model Θ_l associated with it.

Let us formulate the clustering based motif discovery task. Clustering on the k-mers dataset aims to partition the dataset into a set of non-overlapping clusters $\{C_1, C_2,..., C_U\}$, where each cluster C_i holds a subset of k-mers. In our study, each node in the SOM 2D-lattice represents a cluster (i.e. $U = R \times C$). After forming the clusters, each cluster C_i's potential is evaluated as true motif using motif model evaluation metric and rank the clusters based on their obtained scores. In SOMEA, we used Maximum A Posteriori score (MAP score) as the model evaluation metric. Then, top H highest ranked clusters are selected as putative motifs, and k-mers from those clusters indicate the motif locations in the sequences.

PSSM based motif model M_{pssm}

We use the Position-Specific-Scoring-Matrix (PSSM) [2] to model the motif signals. The PSSM based motif model, let it denoted by M_{pssm}, is a matrix, i.e., $M_{pssm} = [f(b_i, i)]_{4\times k}$, where $b_i \in \{A, C, G, T\}$ and $i = 1,..., k$. Here, each entry $f(b_i, i)$ represents the probability of nucleotide b_i in position i and $\sum_{i=1}^{4} f(b_i, i) = 1$. In our SOMEA, the M_{pssm} for a node V_l can be calculated from the k-mers in a node using the maximum likelihood principle, with a pseudo-count value added as under sample correction to the probabilistic model. We follow the Bayesian

estimation method for this purpose [28]. The PSSM entries are computed as follows:

$$f(b_i, i) = (c(b_i, i) + g(b_i)) / (N + 1), \quad (1)$$

where N is the number of k-mers, $c(b_i, i)$ is the frequency of nucleotide b_i at position i of a set of k-mers in a node, $g(b_i) = [n(b_i) + 0.25]/(N \times k + 1)$ and $n(b_i) = \sum_{i=1}^{k} c(b_i, i)$.

Markov chain based background model M_{mc}

In our approach, the background signal is modeled by using the markov chain (MC) model [15]. The MC is a commonly used background signal model to distinguish over-represented motifs from background signals (e.g. in [16,19]). The stochastic and temporal nature of this model can effectively model the complex relationship of the background bases. The MC model assumes that, the probability of occurrence of a nucleotide b_i at position i in a DNA sequence is dependent only on the occurrences of m previous nucleotides. This relationship can be expressed by the conditional probability $p(b_i|b_{i-m}...b_{i-1})$, where $b_{i-m}...b_{i-1}$ are bases that precede base b_i, and m is the markov order. In our approach, the first order MC (i.e. $m = 1$) is used because higher order model usually requires more input data to avoid overfitting. The maximum likelihood estimation of the conditional probability $p(b_i|b_{i-m}...b_{i-1})$ is given by [15] as:

$$p(b_i \mid b_{i-m} \dots b_{i-1}) = \frac{c'(b_{i-m} \dots b_{i-1} b_i)}{\sum_{\forall b_i} c'(b_{i-m} \dots b_{i-1} b_i)}, \ \forall b_i \in \{A, C, G, T\}, \quad (2)$$

where $c'(x)$ is the number of times sub-sequence x found in a set of k-mers in a node.

Let us denote $\pi(a, a')$ to represent the conditional probability $p(a'|a)$ of the first order MC, where $a, a' \in \{A, C, G, T\}$. Then the MC transition matrix gives the background model M_{mc} to be used in SOMEA, i.e., $M_{mc} = [\pi(a, a')]_{4\times4}$, where $\sum_{a'=\{A,C,G,T\}} \pi(a, a') = 1$.

Similarity score

A similarity metric is needed for k-mers assignment to the nodes during the learning. The score of a k-mer $K_j = (b_1 b_2 \dots b_k)$ in respect with the PSSM based model M^l_{pssm} assigned to node V_l, can be computed as,

$$Score(M^l_{pssm}, K_j) = -\log(\prod_{i=1}^{k} f(b_i, i)). \quad (3)$$

Here, k is the length of k-mer, and $f(b_i, i)$ represents the probability of nucleotide b_i in position i. Then, the score of a k-mer K_j to the MC [15] based model M_{mc} of node V_l is computed as:

$$Score(M^l_{mc}, K_j) = -\log\left(p(b_1) \prod_{i=2}^{k} \pi(b_{i-1}, b_i) \right), \quad (4)$$

Here, $p(b_1)$ is the independent and identically distribution (i.i.d) probability of nucleotide b_1 in current node, which is estimated from the k-mers of node V_l.

Hybrid model

In practice, we are unable to certainly deduce if a SOMEA's node is a motif or background at any stage of the learning process. Also, before the system converged, the members of a node are likely to be composed of mixed signals. Therefore, neither PSSM or MC based models (*i.e.*M_{pssm} and M_{mc}) alone would satisfactorily model such composition. However, we can weigh the fitness of MC and PSSM models with respect to the k-mers in a node. In other words, when a set of k-mers fit with a certain model, (i.e., either motif model given by M_{pssm} or background model given by M_{mc}), it is more likely that those k-mers represent that class. Note that both signal models, can represent signal features from opposite class to some extent.

In this work, we aimed to combine the expression abilities of both of the models (i.e., i.e. M_{pssm} and M_{mc}) in an unified mechanism to improve the distinguishing ability of the system, since each node given by SOMEA (or any clustering based approach) contains a fuzzy mixture of motif signals and background signals.

In implementation, we adopted a simple linear weighting scheme to combine these two models for a node V_l as follows:

$$\Theta_l(K_j) = \left[\frac{\alpha}{Score(M^l_{pssm}, K_j)} + \frac{\beta}{Score(M^l_{mc}, K_j)} \right]^{\lambda} \ where\ \alpha + \beta = 1, \quad (5)$$

Equation (5) gives a linear combination of the two models to produce a heterogeneous model for a node V_l. Here, λ is a scaling factor, for simplicity default value of λ is set as 0.5. If a k-mer K_j gets a higher score by this heterogeneous model based scoring $\Theta_l(K_j)$, that indicates K_j has a better fit to the combined model of node V_l.

Motif ranking

Once the SOMEA is stabilized after training, we have to perform an evaluation on the nodes in order to identify the most prominent candidate motifs. The candidate motifs can be identified using either motif evaluation metric or statistical significance value. These metrics usually require the use of background sequences model for computation.

In this work, we adopt the Maximum A Posteriori score (MAP score) [19] for motif ranking. The MAP score measures the conservation property of a motif with respect to the species background sequences [19]. Since, rare motifs in the background can achieve a higher MAP score, this measure can be used to distinguish a true motif from false ones based on their scores ranking. The background sequences can be modeled by using the markov chain model generated from the intergenic sequences of a species under study. This model can be used to assign a probability of a K, namely $p(K \mid M_{mc}^{B})$, under the background model given by M_{mc}^{B}. The MAP score of a node V_l can be calculated as follows:

$$F(V_l) = -\frac{\ln(N_l)}{k}\left[E(V_l) + \frac{1}{N_l}\sum_{K \in V_l} \ln p(K \mid M_{mc}^{B}) \right], \quad (6)$$

where N_l is the number of k-mers in node V_l and $p(K \mid M_{mc}^{B})$ refer to background probability of a k-mer K in respect with background model M_{mc}^{B}. $p(K \mid M_{mc}^{B})$ can be written as,

$$p(K \mid M_{mc}^{B}) = p(b_1, b_2, \ldots, b_m)\prod_{i=m+1}^{k} p(b_i \mid b_{i-m}, b_{i-m+1}, \ldots, b_{i-1}). \quad (7)$$

Here, m is the Markov chain order, k is the length of k-mers, $p(b_1, b_2,..., b_m)$ is the probability of subsequence b_1, b_2, ..., b_m, and $p(b_i|b_{i-m}, b_{i-m+1},..., b_{i-1})$ is the conditional probability of the subsequence b_i under b_{i-m}, $b_{i-m+1,...,}b_{i-1}$ occurrence constraints. For instance, using the 3rd order model, the probability of the sequence $ATGCG$ can be calculated as:
$p(ATGCG \mid M_{mc}^{B}) = p(ATG) \times p(C \mid ATG) \times p(G \mid TGC)$.
This background probability is usually pre-computed on the sequences of interest. In Eq (6), $E(V_l)$ is the Shannon's entropy, that can be written as,

$$E(V_l) = -\sum_{i=1}^{k}\sum_{\forall b_i} f(b_i, i)\log_2 f(b_i, i), \ \forall b_i \in \{A, C, G, T\} \quad (8)$$

Here, $f(b_i, i)$ is the probability of nucleotide base $b_i \in \{A, C, G, T\}$ to occur in i-th position of the PSSM.

Algorithm

In this Section, we describe our SOMEA learning algorithm, which includes the similarity metric used for k-mer assignments, model parameters adaptation, and the finding of BMU for a k-mer. According to [29], any arbitrary set of items, for which a similarity or distance measure between its elements is definable, can be mapped onto the SOM grid in an orderly fashion. Hence, the standard SOM learning algorithm is applicable for our purposes with some modifications.

Adaptation process

We opted for the more speed efficient batch training scheme to update the nodes' model parameters. This method delays the update of the model parameters at the end of an epoch. Heuristic rules are proposed to update each node's PSSM and MC model parameters. We associate each node with three computing components including: two matrices ΔM_{pssm}, ΔM_{mc} and a counter r. Let V_l^* be BMU of an input k-mer $K = (b_1,b_2,...,b_k)$. Denoted $\Delta M_{pssm} = [\Delta f(b_i, i)]_{4\times k}$ for $b_i \in \{A, C, G, T\}$ and $i = 1,... ,k$. Similarly, let $\Delta M_{mc} = [\Delta\pi(a, a')]_{4\times 4}$ for $a, a' \in \{A, C, G, T\}$. We initialize all entries in both matrices ΔM_{pssm} and ΔM_{mc} as 0. Also let $r = 0$. Once a winning node for a k-mer K is found, the matrices of a node V_l^* are updated as follows.

$$\Delta f(b_i, i) = \Delta f(b_i, i) + h(z_l^*, z_j, \sigma)a_{b_i i}, \quad (9)$$

$$\Delta\pi(a, a') = \Delta\pi(a, a') + h(z_l^*, z_j, \sigma)count(a, a') / (k-1), \quad (10)$$

where $a_{b_i i}$ is an entry of the binary matrix $e(K)$, $count$ (a, a') is the frequency of di-nucleotide (aa') in k-mer K and h is a neighborhood function. The neighborhood function h is defined as

$$h(z_l, z_j, \sigma) = \exp\left(-\frac{\left\| z_l - z_j \right\|}{2\sigma^2} \right), \quad (11)$$

where σ is the variance whose value is fixed throughout the learning stage. We also update $r = r + 1$. Upon completion of an epoch, all nodes' model parameters will be updated as follows:

$$f(b_i, i)_{new} = f(b_i, i) + \eta\frac{\Delta f(b_i, i)}{r}, \quad (12)$$

$$\pi(a, a')_{new} = \pi(a, a') + \eta\frac{\Delta\pi(a, a')}{r}, \quad (13)$$

where η is the learning rate and $f(b_i, i)$ and π (a, a') is defined in Eq (1) and Eq (2) respectively. Note that, in the computation of Eq (12) and Eq (13), we first compute $f(b_i, i)$ and π (a, a') using the current set of kmers assigned to a node.

It is also necessary to update the weighting parameters α and β. Assuming a set of N_l k-mers $\{K_1 ..., K_{Nl}\}$ is

assigned to a node V_l at the end of an epoch, the weighting parameters update equations are

$$\alpha_{new} = \frac{\sum_{i=1}^{N_l} Score(M_{pssm}^{l}, K_i)}{\sum_{i=1}^{N_l} (Score(M_{pssm}^{l}, K_i) + Score(M_{mc}^{l}, K_i))} \quad (14)$$

and

$\beta_{new} = 1 - \alpha_{new}$. (15)

Training

Assuming a set of k-mers X is available. The high-level training algorithm for SOMEA is as follows.

1. **Inputs.** k-mer length k, number of top motifs to return in the results H, markov chain background model, and DNA sequences.

2. **Architecture setup.** The SOMEA lattice size ($U = R \times C$)*is* arbitrarily chosen. The default size is 10 × 10. Each node's model, Θ_i, is initialized with random values.

3. **Training.**

Let the BMU index for a k-mer K is $q(K)$.

for epoch=1 to max_epoch **do**

for each $K \in X$ **do**

•Compute $\Theta_i(K), \forall i = 1, \ldots, U$.

•Find the BMU of K as $q(K) = \arg\max_i \{\Theta_i(K)\}$

•Assign k-mer K to node $q(K)$.

•Update ΔM_{pssm}, ΔM_{mc}, r of node $q(K)$ and its neighboring nodes.

end for

Update model parameters of all nodes using Eqs (12) and (13).

end for

4. **Finalizing.**

(a) Compute the MAP score $F(V_i), \forall i = 1, \ldots, U$.

(b) Rank V_i according to their MAP score values.

(c) Save the top H ranked V_i as result.

Acknowledgements

We would be grateful to the group members, Sarwar Tapan, Li Xi, Paul Conilione and Hai Thanh Do, for their comments on the technical aspects and some useful discussions at group meetings. The authors express their sincere appreciation to Sarwar Tapan, who helped in improving the linguistic quality of this paper. NK express his thanks to the Universiti Malaysia Sarawak, which sponsors his PhD study at La Trobe University.

This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 1, 2011: Selected articles from the Ninth Asia Pacific Bioinformatics Conference (APBC 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S1.

Author details

Intelligent Search and Discovery Laboratory, Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia.

Authors' contributions

DH conceived the idea of this study and gave the direction on the framework and experimental studies. Under DH supervision, NK implemented the algorithm and carried out the experimental studies. NK drafted the manuscript and amended by DH. Both authors agreed on the final manuscript.

Competing interests

The authors declare they have no competing interests.

Published: 15 February 2011

References

1. Moses A, Chiang D, Kellis M, Lander E, Eisen M: **Position specific variation in the rate of evolution in transcription factor binding sites.** *BMC Evolutionary Biology* 2003, **3**:19.
2. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16**:16-23.
3. Bailey TL, Elkan C: **Unsupervised learning of multiple motifs in biopolymers using expectation maximization.** *Machine Learning* 1995, **21**:51-80.
4. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16(10)**:939-945.
5. Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS: **Transcription factor binding site identification using the self-organizing map.** *Bioinformatics* 2005, **21(9)**:1807-1814.
6. Wang D, Lee NK: **Computational discovery of motifs using hierarchical clustering techniques.** *ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* Washington, DC USA: IEEE Computer Society; 2008, 1073-1078.
7. Wei Z, Jensen ST: **GAME: detecting cis-regulatory elements using a genetic algorithm.** *Bioinformatics* 2006, **22(13)**:1577-1584.
8. Wang D, Li X: **iGAPK:Improved GAPK algorithm for regulatory DNA motif discovery.** In *Neural Information Processing. Models and Applications, 17th International Conference (ICONIP2008), Volume 6444 of Lecture Notes in Computer Science.* K. W. Wong, B. S. U. Mendis, A. Bouzerdoum. Springer; 2010: 217-225 .
9. Kohonen T: *Self-organizing maps.* 3rd edition. Springer series in information sciences, 30, Springer; 2001.
10. Ferrán EA, Ferrara P: **Clustering proteins into families using artificial neural networks.** *Comput. Appl. Biosci* 1992, **8**:39-44.
11. Giuliano F, Arrigo P, Scalia F, Cardo PP, Damiani G: **Potentially functional regions of nucleic acids recognized by a Kohonen's self-organizing map.** *Comput. Appl. Biosci* 1993, **9(6)**:687-693.
12. Liu D, Xiong X, DasGupta B, Zhang H: **Motif discoveries in unaligned molecular sequences using self-organizing neural networks.** *IEEE Transactions on Neural Networks* 2006, **17(4)**:919-928.
13. Gunewardena S, Zhang Z: **A hybrid model for robust detection of transcription factor binding sites.** *Bioinformatics* 2008, **24(4)**:484-491.
14. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED: **Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics.** *PNAS* 2002, **99(11)**:7323-7328.
15. Robin S, Rodolphe F, Schbath S: **DNA, Words and Models.** New York: Cambridge University Press; 2005.
16. Pavesi G, Mauri G, Pesole G: **An algorithm for finding signals of unknown length in DNA sequences.** *Bioinformatics* 2001, **17(suppl 1)**:S207-214.
17. Fawcett T: **An introduction to ROC analysis.** *Pattern Recognition Letters* 2006, **27(8)**:861-874.
18. **SCPD database.** [http://cgsigma.cshl.org/jian/].
19. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20(8)**:835-839.
20. **Weeder.** [http://159.149.109.9/modtools/].
21. **SOMBRERO.** [http://bioinf.nuigalway.ie/sombrero].
22. **ALIGNACE.** [http://atlas.med.harvard.edu/cgi-bin/alignace.pl].
23. Blanco E, Farre D, Alba MM, Messeguer X, Guigo R: **ABS: a database of annotated regulatory binding sites from orthologous promoters.** *Nucleic Acids Res* 2006, **34(Database issue)**:D63-D67.
24. Li L: **GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery.** *Journal of Computational Biology* 2009, **16(2)**:317-329.

25. Hu J, Li B, Kihara D: **Limitations and potentials of current motif discovery algorithms.** *Nucleic Acids Res* 2005, **33(15)**:4899-4913.
26. Wang D, Li X: **GAPK: genetic algorithms with prior knowledge for motif discovery in DNA sequences.** *CEC'09: Proceedings of the Eleventh conference on Congress on Evolutionary Computation* Piscataway, NJ, USA: IEEE Press; 2009, 277-284.
27. Wang D, Lee NK: **MISCORE: Mismatch-based matrix similarity scores for DNA motif detection.** In *Advances in Neuro-Information Processing, 15th International Conference (ICONIP2008), Volume 5506 of Lecture Notes in Computer Science.* Springer;Köppen M, Kasabov NK, Coghill GG 2009:478-485.
28. Osada R, Zaslavsky E, Singh M: **Comparative analysis of methods for representing and searching for transcription factor binding sites.** *Bioinformatics* 2004, **20(18)**:3516-3525.
29. Kohonen T, Somervuo P: **How to make large self-organizing maps for nonvectorial data.** *Neural Networks* 2002, **15(8-9)**:945-952.

doi:10.1186/1471-2105-12-S1-S16
Cite this article as: Lee and Wang: **SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model.** *BMC Bioinformatics* 2011 **12**(Suppl 1):S16.