31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Transcription Factor Binding Site Detection through Position Cross-Mutual Information variability analysis

Joan Maynou, Montserrat Vallverdú, Francesc Clarià, Joan-Josep Gallardo-Chacón,
Pere Caminal and Alexandre Perera

Abstract— **Regulatory sequence detection is a fundamental challenge in computational biology. One key process in protein synthesis starts with the binding of the transcription factor to its binding site. Different sites can show binding to the same factor. This variability found in binding sequences increases the difficulty of their detection using computational algorithms. In this manuscript, a method for the detection of binding sites is proposed, based on the correlation between binding sequence positions through information theoretical measures. Efficiency values of the method are reported in the form of Receiver Operating Characteristic curves on the detection of different transcription factors of the *Saccharomyces cerevisiae* organism. We compare our results with other known motif detection Motif Discovery scan (MDscan).**

I. INTRODUCTION

Each cell in an organism contains the information for the synthesis of any machinery for biological processes. For its survival, it is necessary a very strict control of gene expression networks in space (cell proliferation, and tissue differentiation) and in time (response to stimuli) [1]. During gene transcription the genetic information is initially transferred from deoxyribonucleic acid (DNA) onto messenger ribonucleic acid (mRNA). The primary mode of transcription control is by the association of specific proteins with their target binding sites in DNA [2]. In addition, they also bind other modulation factors and the RNA polymerase enzyme. These proteins, that are located within gene regulatory regions, are known as transcription factors. In eukaryote organisms, the transcription begins by means of RNA polymerase recruitment by different proteins that recognize specific signals in the region previous to the gene called promoter. One of them is a nucleotide sequence which has the signal to start the transcription. Once the mRNA has been synthesized, it is translated into an aminoacid sequence. This process is known as translation. These polypeptides, after post-translational changes form structural proteins and enzymes that control the metabolic processes in cells. A given transcription factor shows the ability to bind to different sites with different sequences along the genome. Due to this intrinsic variability it is difficult to establish a consensus sequence for the detection of binding sites [3]. Consequently, any detection method of binding sites within DNA sequences must consider the variability of these ones. This has originated several efforts of research, employing different methods to detect patterns in DNA sequences. One of the most relevant are the probabilistic methods, where the most representative models are based on Position Weight Matrices (PWM) also called position-specific weight matrices (PSWM). A PWM is a matrix of score values where there is one row for each symbol of the alphabet, and one column for each position in the pattern. There are several types of PWMs [4]: frequency matrices contain the absolute frequency of a nucleotide at each motif position, weight matrices contain the relative frequency of a nucleotide at a motif position as an estimation of the probability of this fact, and finally, log-odds matrices contain at each position the log of the quotient between the probability of finding particular nucleotide at such a position in sequences containing the real motif and the background frequency of the letter at the same position. One of these methods, *Motif Discovery scan* software, is based on the combination of word enumeration and position-specific weight matrix [5]. Information theoretical measures have been used in genetics to visualize and characterize the information of a sequence set [6], [7]. Detectors based on entropy measurements have also been published, measuring total information content in the binding site by means of Shannon and parametric entropies [8]. This previous work considered that binding site positions are independent to each other whereas other studies have suggested that mutually covarying base-amino acid positions may indicate possible protein-DNA contacts. This covariation can be measured by checking the correlation of the different positions on the binding site. In this manuscript we propose a motif detector applied on transcription factor binding site determination using a differential measure based on mutual information. The method starts from an aligned set of sequences with known binding and analyzes the total cross-information change when the candidate sequence is included in the set.

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2007-63637/TCM and the CIBER in Bioengineering, Biomaterials and Nanomedicin.

J. Maynou, M. Vallverdú, A. Perera and P. Caminal are with Dep. ESAII, Centre for Biomedical Engineering Research, Technical University of Catalonia (UPC), Barcelona, Gargallo, 5, 08028 Barcelona, Spain. http://www.creb.upc.es, http://www.upc.edu. e-mail: joan.maynou, montserrat.vallverdu, pere.caminal, alexandre.perera@upc.edu

J.J. Gallardo as member in CIBER Bioengineering, Biomaterials and Nanomedicin. http://www.isciii.es/htdocs/redes/ciber.jsp e-mail:joan.josep.gallardo@upc.edu

F.Clarià is Dep. Informática y Ingeniería Industrial, Universidad de Lleida, LLeida, Spain. e-mail:Claria@eup.udll.es

II. MATERIALS AND METHODS

A. Method

The proposed method starts with a matrix of aligned sequences with binding evidence. Any new candidate sequence added to the training matrix will cause a variation on the

TABLE I
SUMMARY OF THE RECOGNIZERS ANALYZED

Organism	Recognized	Base	Aligned Sequences
S. cerevisiae	*MCM1*	38	16
S. cerevisiae	*ABF1*	37	22

mutual information of the set of aligned sequences. The detection of an active site is considered depending on the actual change on the mutual information matrix of aligned sequences if the candidate sequence was added to the set of aligned sequences. For random sequences the correlation between the site positions will decrease, whereas for a true binding sequence the overall mutual information of the aligned sequence set is not expected to be modified. Therefore, this measurement allows to build a detector based on the dependence between binding site positions. The validation of the detector has been done by employing a "Leave one-out" cross-validation. Each individual sequence is used as a validation sequence, while the classifier is built on the rest of $n-1$ sequences as training set. The results have been obtained with randomly generated candidate sequences with *1000* sequence repeats, tested successively for each sequence within the training matrix.

B. Position Cross-Mutual Information

Mutual Information is a quantity that measures the mutual dependence of two variables. With two discrete random variables, X and Y, with N possible states $(X_1, X_2, \cdots, X_N)$ and $(Y_1, Y_2, \cdots, Y_N)$, the mutual information can be defined as,

$$\begin{aligned} I(X;Y) &= \sum_N \sum_N p(X,Y) log_2 \left(\frac{p(X,Y)}{p(x)p(y)} \right) \\ &= H(X) + H(Y) + H(X,Y) \end{aligned} \quad (1)$$

where $H(X)$ and $H(Y)$ are the marginal entropies, and $H(X,Y)$ is the joint entropy of X and Y. The mutual information measure is symmetric and non-negative. $I(X;Y) = 0$ holds if and only if two variables (X, Y) are statistically independent under no finite sample effects. In DNA signal, variables X and Y are nucleotides in two different positions. The probability is estimated by frequency estimation from the training matrix (set of sequences with known binding). The Position Cross-Mutual Information (PCMI) measure will allow the study of the dependence between non adjacent positions, providing with information about the correlation of the nucleotides.

C. Database Description

The algorithm requires a group of aligned nucleotide sequences with binding evidence. These sequences comes from the organism *Saccharomyces cerevisiae* which was the first eukaryotic organism with its completed genome known [9]. This organism contains around sixteen million of nucleotides distributed among sixteen chromosomes. We have considered the recognizers *MCM1* and *ABF1*, summarized in table I. The dataset has been obtained from the *TRANSFAC* database, `http://www.genregulation.com/pub/databases.html` , using for the extraction of DNA sequences an own R library for automatic sequence extraction from a transcription factor name. Finally, these sequences have been lined up by means of *MUSCLE* [10], to obtain the different nucleotides involved in each position.

D. Motif Detection

Using the matrix of aligned sequences we perform a measurement of the correlation between positions of the binding sites using mutual information. The values of mutual information for highly correlation positions are close to H_i (Shannon entropy in the site position i). On non-correlated positions the mutual information has values close to the zero. Using this propriety, the developed algorithm performs the comparison between the mutual information of the training matrix and the mutual information of the training matrix when the candidate sequence is added to the set. We test these modifications considering a set of functions. These are,

$$Difference = \left(\sum \gamma \right)^{-1} \quad (2)$$

$$Power = \left[\sum MI_{matrix} \gamma \right]^{-1/2} \quad (3)$$

$$Normalization = \left[\frac{\left(\sum MI_{matrix} \gamma \beta^{-1} \right)}{max(H)} \right]^{-1} \quad (4)$$

where , γ and β are

$$\gamma = |MI_{matrix} - MI_{matrix+seq}| \quad (5)$$

$$\beta = |MI_{matrix} + MI_{matrix+seq}| \quad (6)$$

where, MI_{matrix} is the mutual information matrix of the set of aligned sequences. The $MI_{matrix+seq}$ measurement determines the correlation between training matrix positions when the studied sequence is added. We consider the variation of the information matrix when adding a new sequence by means of the variation produced in the total cross-site mutual information. For a random sequence the dependence between positions decreases, increasing the value of γ. On the other hand, if the sequence added is assumed to be a binding sequence, γ value will be equal or lower and β will be higher than the random case as the binding sequence does not modify the aligned sequence set information. In this manner, we can define a detector that allows for the discrimination between a random sequence and a sequence that belongs to a binding site. The developed method, based on the criterion defined previously, is as follows:

1) For each position within the training matrix, we estimate the probability and the joint probability corresponding to each nucleotide. We impute the missing values to a multi-state nucleotide with probabilities corresponding to the frequency of each nucleotide in the corresponding organism.
2) Marginal and joint entropies are calculated from the PWM, correcting finite sample effects [11].

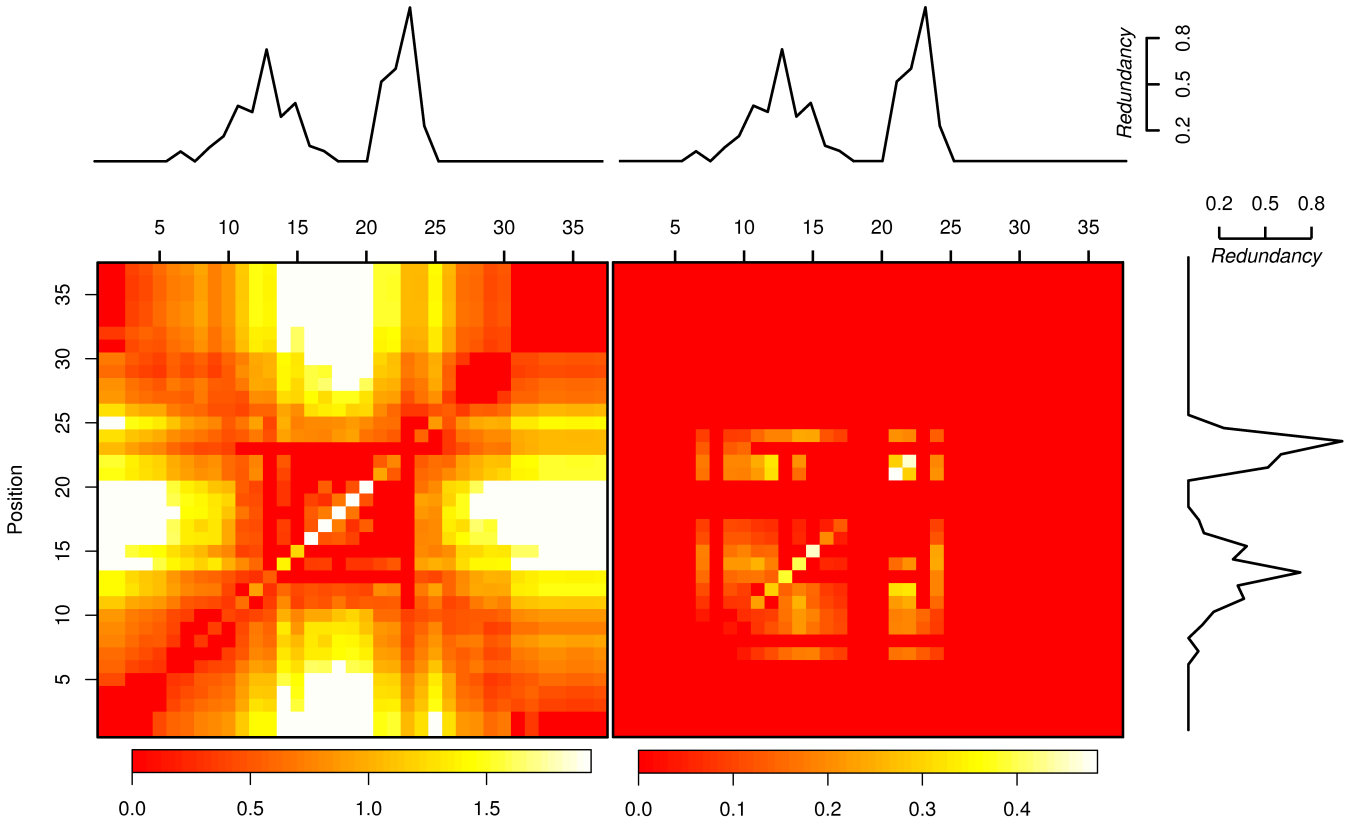

Fig. 1. (left) Mutual Information heatmap between binding site positions for ABF1. Redundancy is plotted on top; (right) Product between mutual Information matrix weighted by the exterior product of the redundancy profile.

3) Mutual information is computed from marginal and joint entropies.
4) 1, 2 and 3 points are repeated, considering the training matrix with the new sequence added.
5) For each mutual information obtained from the studied sequences, a scalar quantity has to be computed using different functions defined in (2), (3), (4), (5) and (6).

TABLE II
AREA UNDER CONVEX SURFACE

	MCM1	ABF1
Difference	0.9916	0.9801
Power	0.9923	0.9650
Normalization	0.9962	0.9848
MDscan	0.9793	0.9754

III. RESULTS

The mutual information between binding site positions and the variability of each position of the ABF1 transcription factor is shown in Figure 1 (left), where both, mutual information matrix and redundancy profile are shown in the plot. The redundancy measurement is a normalized entropy that compares the entropy of the variable to its theoretical maximum value, given as $R = 1 - H/H_{max}$ [8], [12]. The measurement of redundancy provides information about the symbolic variance observed in a position of the set of aligned sequences. The lower the symbolic variance, the higher the value of redundancy. In fact, the redundancy gives information on how much a particular position has been *conserved* on the set of sequences. On the other hand, Figure 1 (right) shows the product between the mutual information matrix between positions and the exterior product of redundancy profile. This measurement helps to determine the correlation between binding site positions that play an important role in the binding from a conservation point of view. In some sense, this graph shows not just the conserved sites among different binding sites but which *correlation* between the sites has been conserved on a number of binding examples. When this measurement is positive we consider that exists dependence between site positions. The detector proposed in this paper evaluates the perturbation into this matrix to check whether this *conserved correlation* is destroyed with the addition of the candidate sequence on the set of sequences. The performance of the detector in the case of ABF1 and MCM1 is shown as a Receiver Operating Characteristic (ROC) for different functions in Figures 2 and 3, respectively. The best learning system will be the one which produces a bigger area under the convex surface (AUC). The performance of the MI based detector is compared against a publicly available detector MDscan [5]. In table II, it can be observed that the detector has a different behavior depending on the functional used. Moreover, the area under convex surface (AUC) for the

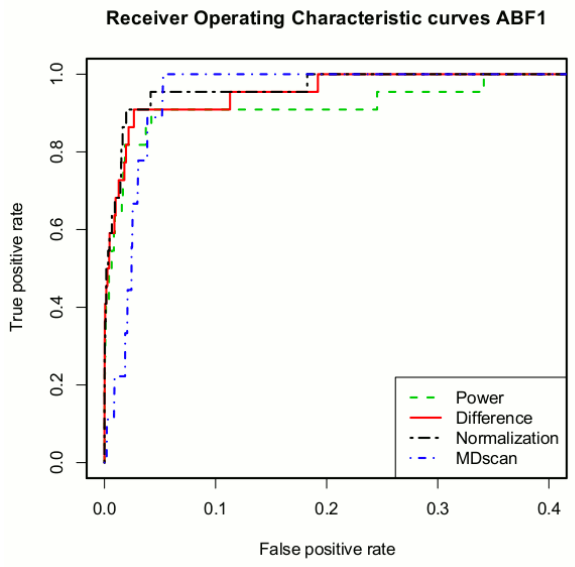

Fig. 2. ROC curve for the different detectors in ABF1

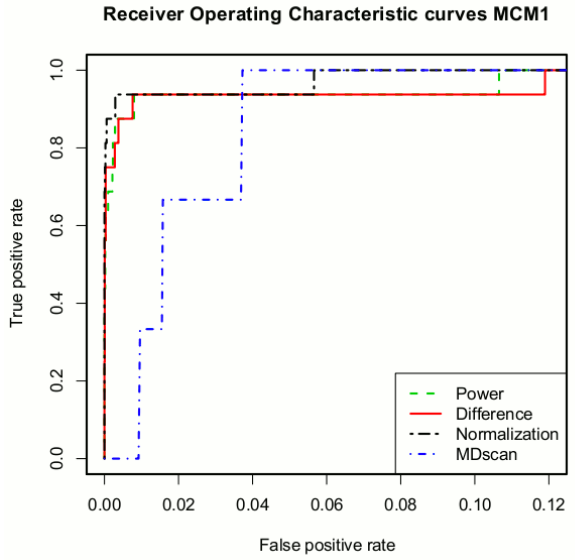

Fig. 3. ROC curve for the different detectors in MCM1

mutual information method is larger than MDscan except for the power functional in ABF1. Therefore, assuming position dependence through mutual information method helps to improve over MDscan in these examples. In Figures 2 and 3 it is observed how the number of true positives (TP) and false positives (FP) depends on the transcription factor binding site and the functional considered (e.g. given a number of true positives the number of false positives changes depending on the functional). The best functional can be selected for the final application given the cost criterion established for miss classifications of True Positives and the area under convex surface maximum.

IV. CONCLUSIONS AND FUTURE WORKS

In this contribution, we have presented a methodology to detect the transcription factor binding sites (TFBS). This method is based on the variation of the Position Cross-Mutual Information from a set of known binding sequences. The proposed algorithm has been applied on the detection of *ABF1* and *MCM1* recognizers from a random sequence. The obtained results improve binding site detection based on first order Shannon and Rényi total variation as reported. The mutual information measurement provides additional information related to the binding process, like the correlation between binding site positions. The proposed method behaves better than MDscan, which is a combined word enumeration and position-specific weight matrix in the case of binding site discrimination against random generated sequences. Future studies will extend the study of the dependence between the positions in the binding site employing parametric uncertainty measurements.

V. ACKNOWLEDGMENTS

CIBER de Bioingeniería, Biomateriales y Nanomedicina is an initiative of the ISCIII.

REFERENCES

[1] D. Latchman, *Eukaryotic Transcription Factors*, 5th ed. Academic Press, 2007.

[2] D. T. Holloway, M. Kon, and C. DeLisi, "Classifying transcription factor targets and discovering relevant biological features." *Biol Direct*, vol. 3, p. 22, 2008. [Online]. Available: http://dx.doi.org/10.1186/1745-6150-3-22

[3] R. Mutihac, A. Cicuttin, and R. Mutihac, "Entropic approach to information coding in DNA molecules," *Materials Science & Engineering C*, vol. 18, no. 1-2, pp. 51–60, 2001.

[4] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements." *Nat Rev Genet*, vol. 5, no. 4, pp. 276–287, Apr 2004. [Online]. Available: http://dx.doi.org/10.1038/nrg1315

[5] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments." *Nat Biotechnol*, vol. 20, no. 8, pp. 835–839, Aug 2002. [Online]. Available: http://dx.doi.org/10.1038/nbt717

[6] T. D. Schneider, "Information content of individual genetic sequences." *J Theor Biol*, vol. 189, no. 4, pp. 427–441, Dec 1997. [Online]. Available: http://dx.doi.org/10.1006/jtbi.1997.0540

[7] G. D. Stormo, "Dna binding sites: representation and discovery." *Bioinformatics*, vol. 16, no. 1, pp. 16–23, Jan 2000.

[8] J. Maynou, M. Vallverdu, F. Claria, A. Perera, and P. Caminal, "Detection of transcription factor binding sites using r&yi entropy," in *Proc. 8th IEEE International Conference on BioInformatics and BioEngineering BIBE 2008*, 8–10 Oct. 2008, pp. 1–5.

[9] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer, "TRANSFAC: an integrated system for gene expression regulation." *Nucleic Acids Res*, vol. 28, no. 1, pp. 316–319, Jan 2000.

[10] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, 2004. [Online]. Available: http://dx.doi.org/10.1093/nar/gkh340

[11] T. D. Schneider, G. Stormo, L. Gold, and A. Ehrenfeuch, "The information content of binding sites on nucleotide sequences," *J Mol Biol*, vol. 188, pp. 415–431, Nov 1986.

[12] A. Perera, M. Vallverdu, F. Claria, J. M. Soria, and P. Caminal, "Dna binding site characterization by means of rényi entropy measures on nucleotide transitions." *IEEE Trans Nanobioscience*, vol. 7, no. 2, pp. 133–141, Jun 2008. [Online]. Available: http://dx.doi.org/10.1109/TNB.2008.2000744