Recognition of splice sites and transcription factor binding sites using generalized maximum entropy models

Ralf Eggeling[1*], Jens Keilwagen[2], Ivo Grosse[1]

[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany

[2]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

[*]corresponding author: eggeling@informatik.uni-halle.de

The maximum entropy principle has been a promising approach for learning models with complex dependency constraints in many areas of science. Here we propose a generalization of the maximum entropy principle based on the Tsallis entropy and apply it to the recognition of splice donor sites and transcription factor binding sites.

1 Introduction

The recognition of statistical patterns within nucleotide sequences is a recurring task in computational biology. In eukaryotes, one important subtask in the process of gene finding is the recognition of splice donor and acceptor sites. Scanning a whole genome for a known pattern, such as a sequence motif from literature, is necessary if all target genes of a certain transcription factor (TF) are of interest. Both problems can be perceived as standard classification problems. Utilizing a likelihood ratio classifier based on a pair of statistical models is a common approach for solving such problems. The simplest and most popular model is a position weight matrix (PWM) model [17, 16], which assumes statistical independence among all positions. However, many studies have shown that dependencies within binding sites exist and that modelling these dependencies improves classification performance significantly [10, 1, 19, 4]. With the rise of next-generation sequencing based technologies like ChIP-seq [11], the amount of available data is dramatically increasing in the near future. Complex statistical models, which have been previously handicapped by overfitting due to small data samples, are becoming of increasing interest.

One of the most successful algorithms for splice site classification utilizes *maximum entropy models* (MEMs) [19]. MEMs choose the probability distribution that maximizes the Shannon entropy [14] under given *constraints* and are applied in many fields of sciences, ranging from linguistics [2] to biology [19]. The advantage of MEMs is the great flexibility with respect to the structure of statistical dependencies that can be taken into account. One popular

example, which yields the best result for the classification of splice donor sites [19], are pairwise dinucleotide frequencies among all positions in the sequence.

Several generalizations of the Shannon entropy have been proposed in the past, with notable examples being the *Rényi entropy* [12] and the *Tsallis entropy* [18]. In this work, we propose the classification of nucleotide sequences based on a generalized maximum entropy model utilizing the Tsallis entropy. Whereas the Shannon entropy assumes a uniform distribution over all events that are equivalent with respect to the constraints, the Tsallis entropy relaxes this assumption. Depending on an external parameter α, it modifies the shape of the probability distribution. We propose a criterion for selecting the optimal α-parameter and evaluate whether the corresponding Tsallis-based MEP yields better classification results than the Shannon MEP.

2 Methods

In this section, we introduce the generalized maximum entropy principle, which contains the traditional MEP as special case. Let S denote the set of different events in the state space, that is, all possible sequences of length L over the DNA alphabet {A,C,G,T}. We denote the probability of each event $i \in S$ by $p_i \in (0,1)$ with $\sum_{i \in S} p_i = 1$.

2.1 Shannon and Tsallis entropy

The Shannon entropy [14] of a probability distribution $\vec{p}$ is defined by

$$H\left(\vec{p}\right) = -\sum_{i \in S} p_i \ln p_i. \tag{1}$$

It is a concave function, since the Hessian of $H\left(\vec{p}\right)$ is a negative definite matrix. The Tsallis entropy is a parameterized generalization of the Shannon entropy and defined by

$$H_\alpha\left(\vec{p}\right) = \frac{\sum_{i \in S} p_i^\alpha - 1}{1 - \alpha}. \tag{2}$$

It has been originally proposed by Havrda and Charvát [5] and popularized by Tsallis [18]. Because of the additional parameter $\alpha \in \mathbb{R}$, is it also known as α-entropy [5]. In the special case of $\alpha = 1$, the Tsallis entropy is undefined. Using l'Hôpital's rule, we find that

$$\lim_{\alpha \to 1} H_\alpha(\vec{p}) = H(\vec{p}), \tag{3}$$

stating that the family of Tsallis entropies contains the Shannon entropy as limiting case for $\alpha \to 1$.

The shape of $H_\alpha\left(\vec{p}\right)$ depends on the parameter α. Considering the Hessian of the Tsallis entropy, we find that $H_\alpha\left(\vec{p}\right)$ is concave for $\alpha > 0$. For $\alpha < 0$ the Tsallis entropy is a convex function, having no well-defined maximum. So we restrict α to $\mathbb{R}^+$ when utilizing the maximum Tsallis entropy principle in the following.

2.2 Constraints and models

Let $T \in S$ denote a subset of events from the state space and let $w \in (0,1)$ denote an arbitrary weight. We define a single constraint by $C = (T, w)$, denote the number of all contraints of a model by J, and denote all constraints of a model by $\vec{C} = (C_1, \ldots, C_J)$. We further define the indicator function

$$\chi(i,j) = \begin{cases} 1: & i \in T_j \\ 0: & \text{else} \end{cases} \tag{4}$$

for a convenient access to the events that belong to a particular constraint. We define the j-th *constraint function* as

$$h_j(\vec{p}) = \sum_{i \in S} \chi(i,j) p_i - w_j. \tag{5}$$

A probability distribution satisfies a set of constraints (denoted by $\vec{p} \in \vec{C}$) if and only if all constraint functions are zero, that is,

$$\vec{p} \in \vec{C} \Leftrightarrow \vec{h}(\vec{p}) = \vec{0}. \tag{6}$$

The most simple constraint, which guarantees the normalization of $\vec{p}$, is $(S, 1)$. A *maximum Tsallis entropy model* (MTEM) is the set of probability distributions for a given indicator function χ and a given parameter α. In this work, we always use the same set of constraints corresponding to marginal dinucleotide frequencies among all pairs of positions, which has been shown to be optimal for the classification of splice donor sites using MEMs [19]. Hence, a specific MTEM is parameterized only by α and thus denoted by MTEM(α). We perceive the MTEM(1) as MEM.

2.3 Learning

Learning a maximum Tsallis entropy model is equivalent to the following problem

$$\vec{p}_\star = \underset{\vec{p} \in \vec{C}}{\operatorname{argmax}}\ H_\alpha(\vec{p}). \tag{7}$$

For solving this *constrained optimization problem*, we introduce a Lagrange multiplier λ_j for each constraint C_j, yielding the Lagrange function

$$L_\alpha\left(\vec{p}, \vec{\lambda}\right) = H_\alpha\left(\vec{p}\right) - \sum_{j=1}^{J} \lambda_j h_j(\vec{p}). \tag{8}$$

Setting the partial derivatives $\frac{\partial}{\partial p_i} L_\alpha\left(\vec{p}, \vec{\lambda}\right)$ zero for each $i \in S$ yields a system of equations that contains each p_i as a function of $\vec{\lambda}$. For $\alpha = 1$, the system of equations can often be solved analytically, but for $\alpha \neq 1$, we obtain the following nonlinear coupled system of equations

$$\forall_{i \in S} : p_i(\vec{\lambda}) = \sqrt[\alpha-1]{\frac{1-\alpha}{\alpha}\left(\sum_{j=1}^{J} \lambda_j \chi(i,j)\right)}, \tag{9}$$

for which we do not find a closed-form solution.

Equation 7 is called *primal problem* and the corresponding target function 2 is called *primal function*. Optimizing a function on a constrained state space is typically hard, so we transform the primal problem into an equivalent unconstrained optimization problem over the state space of $\vec{\lambda}$ by applying equation 9 to equation 8. The resulting function $L_\alpha\left(\vec{p}(\vec{\lambda}), \vec{\lambda}\right)$ is called *dual function* and denoted by $\Psi_\alpha(\vec{\lambda})$. In our case, we obtain

$$\Psi_\alpha\left(\vec{\lambda}\right) = \left(\frac{1-\alpha}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \sum_{i \in S} \left(\sum_{j=1}^{J} \lambda_j \chi(i,j)\right)^{\frac{\alpha}{\alpha-1}} + \frac{1}{\alpha-1} + \sum_{j=1}^{J} \lambda_j w_j, \tag{10}$$

where $\Psi_\alpha(\vec{\lambda})$ is defined on $\mathbb{R}^J$, as it only depends on $\vec{\lambda}$. Minimizing the dual function by

$$\vec{\lambda}_\star = \underset{\vec{\lambda}}{\operatorname{argmin}}\ \Psi_\alpha(\vec{\lambda}) \tag{11}$$

is called the *dual problem*. The primal problem and the dual problem are equivalent, which is commonly known as Lagrangian duality principle. Because of $\vec{p}_\star = \vec{p}(\vec{\lambda}_\star)$, it is sufficient to solve the unconstrained optimization problem of equation 11 to solve the constrained optimization problem of equation 7. We obtain the gradient $\nabla\Psi_\alpha$ by computing the partial derivative for each $\gamma \in (1, \ldots, J)$:

$$\frac{\partial \Psi_\alpha\left(\vec{\lambda}\right)}{\partial \lambda_\gamma} = -\left(\frac{1-\alpha}{\alpha}\right)^{\frac{1}{\alpha-1}} \sum_{i \in S} \chi(i,\gamma) \left(\sum_{j=1}^{J} \lambda_j \chi(i,j)\right)^{\frac{1}{\alpha-1}} + w_\gamma \tag{12}$$

Since $\Psi_\alpha(\vec{\lambda})$ is convex and its gradient exists, an arbitrary numerical optimization algorithm can be used to obtain a global minimum. Here, we use the conjugate gradiate algorithm of Polak and Ribière [9].

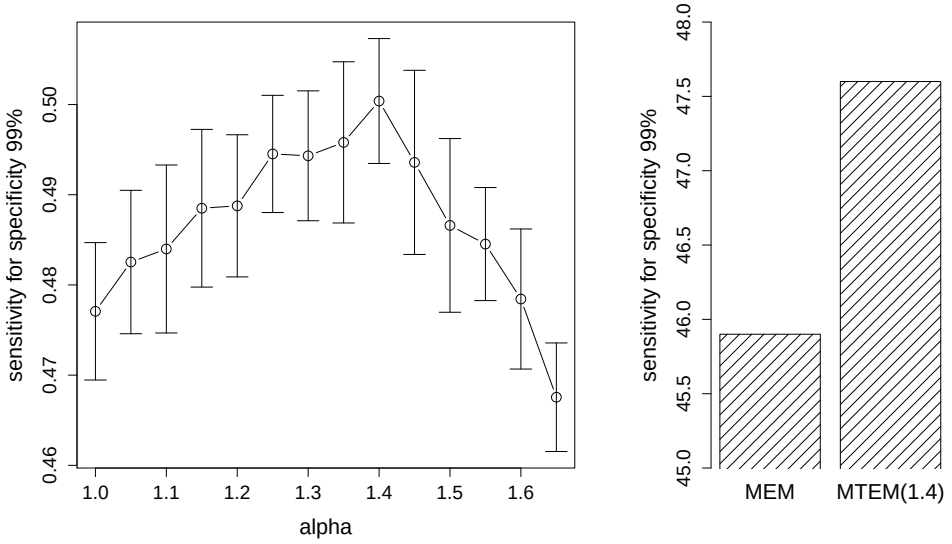
2.4 Model selection

Training a MTEM requires the additional task of determining the optimal α, since there is no possibility to decide a priori which value of α is suitable for a particular data set. To this end, we utilize a K-fold cross validation on the training data set T. After dividing T into equally large subsets T_k with $k \in (1, \ldots, K)$, we train for different values of α and for each partition $k \in (1, \ldots, K)$ a MTEM(α) on $T \setminus T_k$ and evaluate its classification performance on T_k. Next, we average the resulting performance measures for each α and select the parameter $\hat{\alpha}$ that yields the highest average measure. If the training data set contains less than 500 sequences, we suggest using a repeated holdout strategy instead of a cross validation. We subsequently train MTEM($\hat{\alpha}$) on T and return this model for a classification of independent test sequences.

3 Results and discussion

3.1 Classification of human splice sites

In order to evaluate the performance of MTEMs, we utilize the data set from the MEM publication of Yeo and Burge [19]. It consists of 12,623 canonic human splice donor sites and 269,155 decoys. Yeo and Burge divide both sets into training and test data at a ratio of 2:1, which we also use for the following studies. Sequences are 9 bp long including the canonic GT-dinucleotide at positions 4 and 5. Since they do not contribute any additional information, we remove both positions from all data sets and retain sequences of length 7.



(a) Cross validation

(b) Test

Figure 1: **Sensitivity for MEM and MTEM on Yeo/Burge data set.** Figure 1a displays the results of a 50-fold cross validation experiment on the training data set. $\alpha = 1.4$ yields an optimal sensitivity and is subsequently used for training a MTEM on the complete training data set. Figure 1b displays the sensitivity of this optimal model and a standard MEM on independent test data. The MTEM outperforms the MEM by 1.7%.

We classify splice sites versus decoys using a likelihood ratio classifier. In contrast to Yeo and Burge [19], we choose a MTEM for modelling the splice donor sites and a PWM model for modelling the decoys.

First, we perform a model selection step as specified in section 2.4. In a 50-fold cross validation on the training data set, we evaluate the sensitivity for a fixed specificity of 99% for different values of α. The result is shown in figure 1a. A standard MEM yields a sensitivity of 47.8%. With increasing α, the sensitivity increases up to 50% at $\alpha = 1.4$. When further increasing α, the sensitivity decreases quickly to a sensitivity below that of a standard MEM. In this cross validation, we find a maximal increase in sensitivity of 2.2% and conclude that MTEM(1.4) is the *optimal* MTEM for this data set.

Next, we evaluate how well this model performs on the independent test data set. We train MEM and a MTEM(1.4) on the complete training data and utilize both models to classify sequences in the test data set of Yeo and Burge [19]. The MEM yields a sensitivity of 45.9%, whereas the MTEM(1.4) obtains a sensitivity of 47.6% (figure 1b). Even though the difference is smaller compared to that of the cross validation experiment, we still find an increase of sensitivity by 1.7% by using the Tsallis entropy. We conclude, that the Tsallis entropy is – at least for this data set – more suitable to distinguish splice donor sites from decoys than the standard Shannon entropy.

3.2 Classification of splice sites from different organisms

In a second study, we apply MTEMs on splice sites from five different organisms: *Homo sapiens*, *Danio rerio*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans* [15]. Splice sites and decoys of each organism are partitioned into five sets [15].

In analogy to the previous study, we define the union of data sets one, two, three and four as training data and use data set five as independent test data. These data sets are substantially larger than the Yeo/Burge data set, so we perform a 5-fold cross validation during the model selection step. We obtain an optimal parameter $\hat{\alpha}$ for each organism, and use the resulting MTEM to classify the test data. We compare the results with those of a standard MEM in table 1.

Table 1: **Classification results on data sets of different organisms.** The sensitivity for a specificity of 99% is shown for the optimal MTEM and a standard MEM.

organism	MEM	MTEM($\hat{\alpha}$)	Δ	$\hat{\alpha}$
H. sapiens	48.4	**48.5**	0.1	0.91
D. rerio	52.8	**53.2**	0.7	0.80
A. thaliana	47.7	**47.8**	0.1	0.90
D. melanogaster	70.4	**70.8**	0.4	0.91
C. elegans	61.7	**62.6**	0.9	0.80

For *D. rerio*, *D. melanogaster*, and *C. elegans*, we observe an improved sensitivity. For *H. sapiens* and *A. thaliana*, MTEM($\hat{\alpha}$) and MEM perform almost equally well. Interestingly, there are no cases in which the MTEM($\hat{\alpha}$) is outperformed by a standard MEM, thus the application of the maximum Tsallis entropy principle is never disadvantageous.

3.3 Classification of TFBS

After having evaluated the performance of MTEMs for splice site classification, we investigate if they might be also useful for the classification of transcription factor binding sites (TFBS). Sequence motifs from databases such as JASPAR [13] or TRANSFAC® [7] are often based on a computational de-novo discovery. Since most de-novo motif discovery tools use a PWM model to infer the statistics of the motif, the predictions are biased towards statistical independence of nucleotides. In order to avoid this effect, we utilize data from protein binding matrix (PBM) experiments [3] that determine the in-vitro binding affinity of each possible oligonucleotide of length eight to given proteins. Even though there are also some computational postprocessing steps of the experimental output, the set of high scoring oligomers should be less biased towards statistical simplicity than computational PWM-based predictions.

Here, we focus on PBM data of transcription factors from the yeast *Saccharomyces cerevisiae* [20], which are available through the UniPROBE database [8]. PBM experiments assign a score in the interval $(-0.5, 0.5)$ to each oligomer. We define a threshold by considering all oligomers with a score greater than 0.35 to be bound by the TF and declare them as positive data set. We further assume all oligomers with a score less than 0 not to be bound by the TF and use them as negative data set. Next, we choose the ten TFs with largest positive data sets to perform classification experiments.

Using PBM data requires an additional preprocessing step, as the high scoring oligomers are not necessarily aligned. Some are shifted by a position with respect to a true motif, and approximately half of them differ in strand orientation. In order to to cope with that problem, we apply MotifAdjuster [6] with demanding a common oligonucleotide of length six. We randomly fill empty positions of the resulting alignment of length ten with nucleotides according to the relative nucleotide frequencies of the column and extract the oligomer from positions 2 to 9.

Table 2: **Classification results on PBM data of yeast transcription factors.** The sensitivity for a fixed specificity of 99.9% of a MTEM is compared with the performance of a standard MEM. The MTEM yields an improvement up to 1.3%.

TF	MEM	MTEM($\hat{\alpha}$)	Δ	$\hat{\alpha}$
Put3-11	98.2	**99.5**	1.3	1.12
Rdr1-9	92.5	**93.2**	0.7	1.18
Rds1	94.9	**95.5**	0.6	1.17
Tbs-1	**99.1**	98.8	-0.3	1.04
Yox1	96.4	**97.0**	0.6	1.15

In contrast to splice sites, the classification of TFBS is orders of magnitude easier, since all motifs contain at least some highly conserved nucleotides, while the negative data does not share this property. Hence, we measure the sensitivity fixed specificity of 99.9%.

In analogy to the previous experiments, we first determine $\hat{\alpha}$ via the model selection procedure of section 2.4. We find $\hat{\alpha} > 1$ for all TF and compare the corresponding models with a standard MEM.

In five of ten cases (Sum1-11, Sum1-9, Ume6-11, Yll054-9, Asg1), the sensitivity of the optimal MTEM and the MEM is identical and varies between 98.2% and 99.7%. In these five cases, the classification is nearly perfect, yielding almost no room for improvement. The results of the five remaining TF are shown in table 2. In four of five cases (Put3-11, Rdr1-9, Rds1, Yox1), we find an increase of sensitivity of more than 0.5%. However, the model selection step of the MTEM is misleading in the case of Tbs-1, where a straightforward application of a MEM would have lead to a more accurate classification.

3.4 Conclusions

We developed a generalization of maximum entropy models by utilizing the Tsallis entropy. We studied the efficacy of MTEMs for classifying splice sites and transcription factor binding sites. Apart from one exception, we found that MTEMs increase the classification accuracy over MEMs or that both models classify equally well. These results make it tempting to speculate that the maximum Tsallis entropy principle might possibly be useful for other classification problems in computational biology or beyond. We implemented the model including all learning algorithms in the open source Java library Jstacs[1] and make them publicly available with the next release.

[1] http://www.jstacs.de

References

[1] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 28–37, 2003.

[2] A. Berger, S. Pietra, and V. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[3] M.F. Berger, A.A. Philippakis, A.M. Qureshi, F.S. He, P.W. Estep, and M.L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, 24:1429–1435, 2006.

[4] S. Gunewardena, P. Jeavons, and Z. Zhang. Enhancing the prediction of transcription factor binding sites by incorporating structural properties and nucleotide covariations. *Journal of Computational Biology*, 13:929–945, 2006.

[5] J. Havrda and F. Charvát. Quantification method of classification processes. *Kybernatika*, 3(1):30–35, 1967.

[6] J. Keilwagen, J. Baumbach, T.A. Kohl, and I. Grosse. MotifAdjuster: a tool for computational reassessment of transcription factor binding site annotations. *Genome Biology*, 10:R46, 2009.

[7] V. Matys, E. Fricke, R. Geffers, E. Gling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 33:374–378, 2003.

[8] D.E. Newburger and M.L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37:D77–D82, 2009.

[9] E. Polak and G. Ribière. Note sur la convergence de méthodes de directions conjuguées. *Revue Française d'Informatique et de Recherche Opérationnelle*, 16:35–43, 1969.

[10] M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, O.A. Podkolodnaya, D.G. Vorobyev, N.A. Kolchanov, and G.C. Overton. Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. *Bioinformatics*, 15:631–643, 1999.

[11] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, December 2000.

[12] Albert Rényi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960.

[13] A. Sandelin, W. Alkema, P. Engström, W.W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:D91–D94, 2004.

[14] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[15] S. Sonneburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8:S7, 2007.

[16] Rodger Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519, 1984.

[17] G.D. Stormo, T.D. Schneider, and L.M. Gold. Characterization of translational initiation sites in e.coli. *Nucleic Acids Research*, 10(2):2971–2996, 1982.

[18] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

[19] G. Yeo and C.B. Burge. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology*, 11(2/3):377–394, 2004.

[20] C. Zhu, K. Byers, R. McCord, Z. Shi, M. Berger, D. Newburger, K. Saulrieta, Z. Smith, M. Shah, M. Radkakrishnan, A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T.V.S. Murthy, J. LaBaer, and M. Bulyk. High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Research*, 19:556–566, January 2009.