BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–8

Context-specific Independence Mixture Modeling for Positional Weight Matrices

Benjamin Georgi[1] and Alexander Schliep[1]

[1]Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnestrasse 73, 14195 Berlin, Germany

ABSTRACT

Motivation: A positional weight matrix (PWM) is a statistical representation of the binding pattern of a transcription factor estimated from known binding site sequences. Previous studies showed that for factors which bind to divergent binding sites, mixtures of multiple PWMs increase performance. However, estimating a conventional mixture distribution for each position will in many cases cause overfitting.

Results: We propose a *context-specific independence* (CSI) mixture model and a learning algorithm based on a Bayesian approach. The CSI model adjusts complexity to fit the amount of variation observed on the sequence level in each position of a site. This not only yields a more parsimonious description of binding patterns, which improves parameter estimates, it also increases robustness as the model automatically adapts the number of components to fit the data.

Evaluation of the CSI model on simulated data showed favorable results compared to conventional mixtures. We demonstrate its adaptive properties in a classical model selection setup. The increased parsimony of the CSI model was shown for the transcription factor Leu3 where two binding-energy subgroups were distinguished equally well as with a conventional mixture but requiring 30% less parameters. Analysis of the human-mouse conservation of predicted binding sites of 64 JASPAR TFs showed that CSI was as good or better than a conventional mixture for 89% of the TFs and for 70% for a single PWM model.

Availability: http://algorithmics.molgen.mpg.de/mixture

Contact: {georgi | schliep}@molgen.mpg.de

1 INTRODUCTION

The reliable identification of putative transcription factor binding sites (TFBS) in genomic sequences is a problem of considerable importance for understanding gene regulation. The accepted approach is to formulate a mathematical representation of the binding pattern of a given factor based on collections of confirmed binding site sequences. This representation is subsequently used to score candidate sequences for occurrences of said pattern. The effectiveness of this approach depends on the models ability to accurately formalize the regularities found in the confirmed sites. Positional weight matrices (PWM) [30, 31, 38, 33, 34] are a statistical approach to modelling the factor-specific binding site composition. A PWM is derived from a multiple alignment of confirmed binding sites. For each position in the alignment a distribution over the four bases is estimated from the corresponding alignment column. Assuming independence between positions, this gives a probabilistic model of the binding site of a specific factor which subsequently can be used to score whether a DNA sequence contains a binding site for this factor [15, 18].

However, this approach relies on two strong assumptions, namely that *all* positions within the site are independent and, more importantly, that all binding sites of a factor are slight variations of the *same* sequence. The former has been shown to be a simplification of biological reality for such examples as the Zinc finger motive [40] or the Mnt repressor [21]. For the latter there is ample biological evidence to make it at least doubtful: It is well known that TFBS occur in clusters of functionally interacting transcription factors (TF) in promotor regions, so called transcriptional modules [6, 20, 36]. A single factor may have many different interaction partners for different genes and it has been shown that the topology of these modules has an impact on the binding site sequences found for about nine thousand sites in S. cerevisiae [5]. Also, it is known that a single change in a binding site can have profound effects on both the interaction behavior of a factor [24] or the level of induced gene expression [39]. Moreover, in [17] the authors find increased levels of conservation for non-consensus binding site positions for 16 factors in 10 bacterial genomes, concluding that these sites are subject to evolutionary pressure. This gives further evidence for

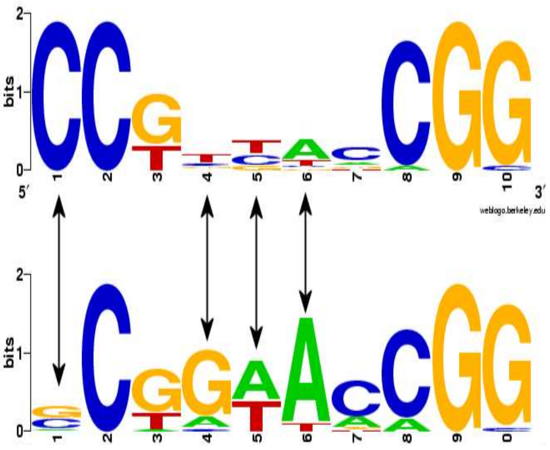

Fig. 1. WebLogos (*http://weblogo.berkeley.edu*) for the two subgroups of Leu3 binding sites. It can be seen that sequence variability is limited to positions 1, 4, 5 and 6 (indicated by arrows).

a level of biological complexity of binding site sequences beyond the “single site” hypothesis and motivates the development of more sophisticated methods.

This issue has received some attention in recent years. In [4] the authors successfully used subclasses of Bayesian networks for *de novo* motive discovery, among them mixtures of PWMs. More recently, in [14] binding sites have also been described as mixtures of PWMs. There it was shown, that a two component mixture model yielded improved conservation scores and higher expression coherence when compared to using a single PWM for a collection of 64 PWMs taken from the JASPAR data base [26].

However, the conventional mixture approach has severe drawbacks. First, it is an essentially unsolved problem to choose an appropriate number of mixture components, in particular if data is sparse and the classical model selection techniques [2, 29] do not apply. In general too few components lead to suboptimal performance due to insufficient generalization, while, more severely, too many components will cause overfitting. To circumvent this issue the number of components was fixed to two in [14]. Secondly, it seems plausible that for most factors which have several types of binding sites (and can thus be modeled more precisely by a mixture), the different subgroups will not consists of distinct, dissimilar sequences. Rather, the variability between sites will be concentrated on specific positions. Estimating a full PWM for each mixture component will then introduce unnecessary parameters into the model. This increases model complexity unnecessarily and leads to less robust parameter estimates.

We present an extension of the conventional mixture framework that addresses these problems by learning an explicit dependency structure between the components of a PWM mixture. The basic idea of the method is to reduce the number of parameters required in the model by representing binding site positions with little variability in the different components by the same distribution. A biological example for such a situation is the TF Leu3. In [14] the authors showed that a two component mixture naturally separated the known binding sites [19] into one high and one low binding-energy subgroup. Now, consider Fig. 1. The figure shows the sequence logos [27] for these subgroups. It can be seen that sequence variability is only present in position 1, 4, 5 and 6 (indicated by arrows) while the other sites are highly conserved. Another example is the factor Reb1. Reb1 binds with different affinities to motives TTACCC<u>G</u> and TTACCC<u>T</u> [37], that is the two subgroups differ in a single position only.

This notion of adapting model complexity to the data is known as *context-specific independence* (CSI) and has received considerable attention in the probabilistic reasoning community [7, 8, 12]. In the context of mixture modeling, CSI has been successfully used for the analysis of gene expression data [3].

The advantage of the CSI model in settings such as the Leu3 and Reb1 data is that in a conventional mixture random sequence deviations will cause the parameters in the different components for the same position to vary slightly, even if there is no meaningful variability on the sequence level. This overfitting introduces a distortion in the scores produced by the model that may result in a decrease in performance. Therefore, learning a CSI structure does not only yield a more parsimonious model, as less parameters are required, but also increases robustness for noisy data. Moreover, if components share the same group in the CSI structure for all positions, they can be merged thus reducing the number of components in the model. Therefore learning of a CSI structure allows for an automatic reduction of the number of components to a value more appropriate for a data set as an integral part of model training.

In the following sections we are going to introduce notation for the CSI mixture model and present the structure learning algorithm. We will then evaluate the performance of our method based on both simulated and real biological data.

2 METHODS

2.1 CSI Mixture Models

Before we begin defining the CSI mixture model we briefly introduce notation for conventional mixture models (refer to [23] for a detailed coverage of the subject). Let $X_1, ..., X_p$ denote random variables (RV) over the four bases (A,C,G,T) corresponding to a binding site with p positions. Given a data set D consisting of N samples $x_i, i = 1, ..., N$ where each x_i consists of an realization $x_{i1}, ..., x_{ip}$ of $X_1, ..., X_p$ a K component mixture distribution is given by

$$P(x_i) = \sum_{k=1}^{K} P(C = k) \prod_{j=1}^{p} P_j(x_{ij}|C = k), \qquad (1)$$

where C is a RV representing the component number, the $P(C = k)$ are the component priors ($\sum_{k=1}^{K} P(C = k) = 1$) and the $P(x_{ij}|C = k)$ are discrete distributions over the four bases, conditional on the component RV C. That is, each $P(x_{ij}|C = k)$ is parameterized by a 4-component probability vector $\theta_{j|k}$. Define the collection of all $\theta_{j|k}$ and the weight vector $\theta_\pi = (P(C = 1), ..., P(C = K))$ as $\theta_M = (\theta_\pi, \theta_{j|k})$. Then θ_M completely parameterizes the mixture M. The likelihood $P(D|M)$ for data set D is simply the product over the mixture densities of each independent sample

$$P(D|M) = \prod_{i=1}^{N} P(x_i). \qquad (2)$$

At this point we would like to point out that mixtures models and the extensions we are about to describe are not limited to discrete features. Rather the $P_j(x_{ij}|C = k)$ can be of any parametric family, be it discrete or continuous and that in particular the domains of the X_j can be heterogenous.

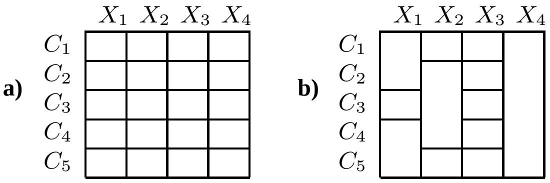

Fig. 2. a) Model structure for a conventional mixture with 5 components and four RV. Each cell of the matrix represents a distribution in the mixture and every RV has an unique distribution in each component. **b)** CSI model structure. Multiple components may share the same distribution for a RV as indicated by the matrix cells spanning multiple rows. In example C_2,C_3 and C_4 share the same distribution for X_2.

In order to formally define the CSI mixture model it is helpful to first review the independence assumptions implicit in the conventional mixture model. In addition to the routine assumption of independence between the different data samples x_i, we also assume conditional independence between the X_j given a component k. This leads to a model structure as shown in Fig. 2a. The figure shows the structure matrix for a five component mixture with $p = 4$, each cell representing an uniquely parameterized distribution over the corresponding X_j. In a CSI model we qualify the general assumption of conditional independence between the X_j by representing multiple components with the same set of parameters. Essentially this amounts to learning a parameter tying structure for each X_j over the range of C. This is closely related to learning parameter ties in the topology of a *Hidden Markov Model* (HMM) [10, 32] as a mixture can be seen as an HMM with strongly constrained topology. In Fig. 2b we show one possible CSI structure for this model. Again, each cell of the matrix represents an uniquely parameterized distribution. This means that for example C_1 and C_2 are represented by the same distribution for X_1 and all components share the same distribution for X_4.

Formally we define the CSI mixture model as follows: For the set of component indexes $\mathcal{C} = \{1, .., K\}$ and variables $X_1, ..., X_p$ let $G = \{g_j\}_{(j=1,...,p)}$ be the CSI structure of the model M. Then $g_j = (g_{j1}, ... g_{jZ_j})$ where Z_j is the number of subgroups for X_j and each $g_{jr}, r = 1, ..., Z_j$ is a subset of component indexes from $\mathcal{C}$. That is, each g_j is a partition of $\mathcal{C}$ into distinct subsets where each g_{jr} represents a subgroup of components which share the same distribution for X_j. The CSI mixture distribution is then obtained by replacing $P_j(x_{ij}|C = k)$ with $P_j(x_{ij}|g_j(k))$ in (1) where $g_j(k) = r$ such that $k \in g_{jr}$. Accordingly $\theta_M = (\theta_\pi, \theta_{X_j|g_{jr}})$ is the model parameterization. The complete CSI model M is then given by $M = (G, \theta_M)$

The usefulness of the CSI approach for real world applications obviously depends on the ability to accurately and reliably determine an appropriate structure from data. This problem is addressed in the following section.

2.2 Structure Learning

The task of learning a CSI model from data consists of assigning values to the group structure variables g_j and estimating parameters for the induced distributions. For the latter the *Expectation Maximization* (EM) [9, 22] algorithm is the standard technique to arrive at parameter estimates. For the former, we adopted a Bayesian approach in the *Structural EM* algorithm framework [11]. This means that we score different candidate model structures based on the model posterior $P(M|D)$ which according to Bayes rule is given by

$$P(M|D) \propto P(M)P(D|M),$$

where $P(M)$ is a prior over the model structure and $P(D|M)$ is the Bayesian likelihood based on the data D and the *maximum a posteriori* (MAP) parameter estimates $\overrightarrow{\theta}_M$. That is

$$P(D|M) = P(D|\overrightarrow{\theta}_M)P(\overrightarrow{\theta}_M),$$

where $P(\overrightarrow{\theta}_M)$ is a prior over the model parameters in form of a product of conjugate Dirichlet priors over the individual elements of θ_M. The prior over the mixture weights θ_π was uniform, the priors over the $\theta_{X_j|g_{jr}}$ were chosen to be almost uniform with a small bias towards uniform θ (i.e., all hyper-parameters of the Dirichlets were set to 1.02). This was done to guard against overfitting by setting zero probabilities in the parameter estimation.

For the model prior $P(M)$ we adopted a fairly simple factored form

$$P(M) \propto P(K)P(G), \qquad (3)$$

where the $P(K)$ is the prior over the number of components and $P(G)$ is the model structure prior. We set $P(K) = \gamma^K$ and $P(G) = \prod_{j=1}^{p} \alpha^{Z_j}$ with both γ and $\alpha < 1$. Thus by means of the prior we introduce a bias towards smaller models and simpler structures into the model posterior.

2.3 Learning Algorithm

For a CSI mixture with K components over p RVs there are B_K^p possible model structures, where B_K is the Kth Bell number [1]. B_K gives the number of possible partitions of a set with K elements. This makes an exhaustive search over the structure space infeasible even for moderate sizes of K and p. For example for $K = 3$ and $p = 8$ there are 390,625 different structures. Instead we adopt an iterative greedy backward-selection procedure to learn a CSI model $M = (G, \theta_M)$. We initialize the procedure with $M^0 = (G^0, \theta_M^0)$, such that G^0 is the structure of maximal complexity (which is equivalent to a conventional mixture) and the initial parameters θ_M^0 are obtained by a single EM update based on a random assignment of data to components, followed by conventional parametric EM to obtain the MAP parameters.

In each following steps l we then use the current model $M^l = (G^l, \theta_M^l)$ to score the candidate structures G based on possible merges $(g_{jr}^l, g_{jz}^l) \rightarrow g_{jr}^l \cup g_{jz}^l \;\; (r, z = 1, ..., Z_j, r \neq z)$ by computing the posteriors and accepting the candidate model with maximal posterior as M^{l+1}. Due to the independence assumption between the X_j we can score the candidate structures of each variable separately. In the framework of *Structural EM* [11] this scoring can be done efficiently by computing the expected sufficient statistics of a candidate based on the current model M^l. Once we have determined G^{l+1} we can obtain the parameterization θ_M^{l+1} by running parametric EM. The procedure terminates when all candidate models have a posterior worse than M^l.

In summary, the structure learning procedure for an initial model M^0 consists of iterations over the following steps:

- Score possible candidates M^{l+1} based on M^l, accept candidate with maximal posterior.
- Optimize θ_M^{l+1} by running parametric EM.

2.4 Choosing the Structure Prior

One important aspect of the Bayesian approach to structure learning is the choice of the hyper parameters in the model prior. There are techniques for estimating these parameters directly from data [25] or by simulation techniques such as Gibbs sampling [13]. In our application and for this first analysis we choose the structure prior parameter α directly based on a simple heuristic.

In general the prior $P(M)$ encodes the preference for a simpler model. This is contrasted with the data likelihood $P(D|M)$ which increases with model complexity. One way of thinking about the relation between prior and likelihood is that the prior acts as a regularization of the likelihood to prevent overfitting. From the perspective of the CSI structure learning task, the choice of the hyper parameter α of the structure prior $P(G)$ expresses our preference of a simpler, less complex structure. One way to look at this is that α puts a threshold on the decrease in likelihood we are willing to accept in exchange for a less complex structure. Since the likelihood of a data set is dependent on the sample size N the same must be true for α. To make this explicit, consider the decision rule between a model M^l and a candidate model M during an iteration of the learning algorithm. Recall that M^l and M are identical except for a single merge in a g_j. This merge is accepted if

$$\frac{P(M^l|D)}{P(M|D)} = \frac{P(D|M^l)P(M^l)}{P(D|M)P(M)} \leq 1.$$

Substituting Eq. 2 and (3) and cancelling terms we obtain

$$\prod_{i=1}^{N} \frac{P(x_i|M^l)}{P(x_i|M)} \quad \alpha \quad \leq 1.$$

Each of the N fractions gives the decrease in likelihood of a x_i for moving from M^0 to the less complex model M. That is, we can think of each fraction as $(1+\delta_i)$ where δ_i is the relative decrease in likelihood for x_i. Under the simplifying assumption that all of the δ_i are equal, i.e. $\delta_i = \delta$, we can now choose a δ as the *maximal relative decrease* in likelihood we are willing to accept in exchange for a less complex model. Then α is given by

$$\alpha = \alpha(\delta, N) = \frac{1}{(1+\delta)^N} \; .$$

It is important to stress that at this point all we have done is to replace the choice of α with the choice of δ. However this is advantageous for two reasons: First, the formula given above explicitly shows the impact of the data set size N. Secondly, δ has a straightforward interpretation as the reduction in likelihood between simple discrete distributions. As such it is easier to make an informed choice for δ based on the specific application. In our case it seemed reasonable to use a strong prior, such that the structure only introduced additional complexity into the model if clearly warranted by the data. In the following we chose the prior according to $\alpha(0.18, N)$ (unless noted otherwise). As an example for 20 sequences we obtain $\alpha(0.18, 20) = 0.036$

2.5 Sequence Scoring

One practical advantage of the model extensions described above is that it refines the models ability to represent TF binding patterns without abandoning the framework of probabilistic models. This means that the CSI model can be seamlessly and easily combined with established techniques for finding hits with significant scores in genomic sequences [15, 18]. Here, as in [14], the score of a mixture was defined as the maximum score over all components. This means that the score of a sequence was given by the strongest signal found among the components. Similar scoring schemes have been used for instance in the field of speech recognition.

3 RESULTS

3.1 Simulation Studies

In order to examine the difference in performance between normal mixture and CSI models we generated artificial data sets from mixtures with differing numbers of components and structures.

In the first experiment the generating model was a two component CSI mixture with $p = 10$ and random weights θ_π. The CSI structure was set up as follows: Out of the ten positions, six were represented by single distributions in both components and four had a unique distribution in each component. The parameters of the distributions $\theta_{X_j|g_j}$ were chosen randomly.

	Best model	Best avg. BIC
G_1	M_1	10851
G_2	M_2	11444
G_{CSI}	M_{CSI}	12266
G_4	M_{CSI}	12350

Table 1. Optimal model for the four data sets according to the average BIC over 30 repetitions.

First we evaluated the ability of our method to adapt to the structure in the data and thus to avoid overfitting. We trained one conventional and one CSI mixture model, both using three components on a training data set with 40 samples. The first result was that the structure learning algorithm recovered the generating models two component CSI structure with high accuracy (not shown). In order to quantify the advantage of the CSI model for sequence scoring we generated test data sets with 500 samples. We used a uniform background model to obtain the scores for each sample and the scores were then converted to p-values based on a score distribution on 1Mb of random sequence. We repeated the simulation for 30 different randomly generated data sets and observed that the CSI mixture yielded better (lower) p-values than the conventional mixture. The one-sided Wilcoxon test for paired samples assigned a significance of 0.02 to this result. Repeating the experiment with only 25 training samples confirmed these results with a Wilcoxon test significance of 0.04.

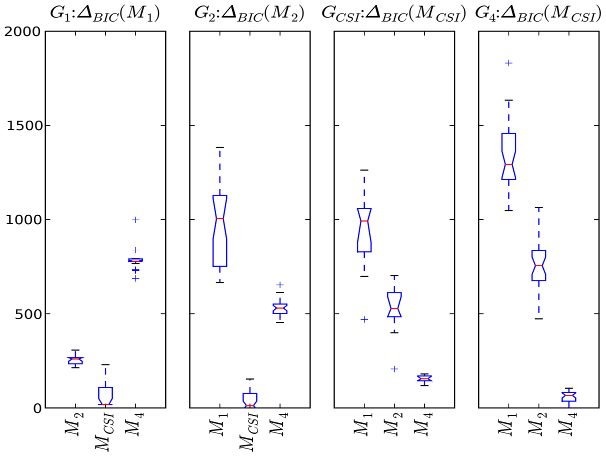

Fig. 3. Distributions of the difference in BIC to best scoring model for the four simulated data sets on 30 repetitions.

The next question we addressed was how the CSI model performed for different data sets in a classical model selection setup. We generated data sets of size 500 with $p = 10$ from four different models: a single PWM model G_1, a conventional two component mixture G_2, a CSI mixture with four components G_{CSI} and a conventional four component mixture G_4. The parameters of the discrete distributions in θ_M were chosen such that one base β was assigned a random probability sampled uniformly from [0.5,0.8] and the remaining mass split evenly over the other bases. In each case β was chosen such that it adhered to the CSI structure of the respective model, that is components that did not share a group for a X_j also had a dissimilar β. The structure in G_{CSI} consisted of 6 positions with four groups and two positions with three and two groups each. Subsequently we trained 30 models M of each of the four types (i.e. M_1, M_2, M_{CSI} and M_4) on each of the four data sets. Model fit was assessed by the *Bayesian Information Criterion* (BIC) [29]. The best scoring model for each data set and its average BIC value based on the 30 repetitions is shown in Table 1. As one would expect, the model type that best matches the respective generating model yields the optimal BIC. A more interesting point to consider was the distributions of the differences of the remaining models to the optimal BIC shown in Fig. 3. It can be seen that for data sets where M_{CSI} is not optimal it achieves BIC scores very similar to the best. These results illustrate the inherent ability of CSI models to adapt to different data settings. This makes CSI a preferable choice of model for practical applications where the true number of components is unknown.

3.2 Analysis of TF LEU3

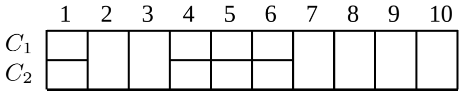

Fig. 4. Two component CSI mixture structure for known Leu3 binding sites. Each cell represents a discrete distribution, where cells spanning both rows identify positions with high conservation in both subgroups.

It was shown that 46 known binding sites of the TF Leu3 [19] can be separated into a high and low binding-energy subgroup using a two component mixture with highly significant p-value [14]. We repeated this analysis by training a two component CSI mixture. Since we were using the model in a clustering context a weak prior of $\alpha(0.05, 46) = 0.11$ was used. Fig. 4 shows the resulting CSI structure. Note the correspondence between the fully parameterized positions (1, 4, 5, 6) and the group specific sequence variability as visualized in Fig. 1. The CSI mixture yielded a subgroup division of the Leu3 sites that was practically identical to the one previously reported. However there are two important differences between the two models: First, the conventional mixture requires the estimation of 61 free parameters while due to the tying expressed in the CSI structure our model only needs 43 parameters. This means that CSI gave equivalent results using about 30% less parameters. Secondly, the CSI structure makes information about the subgroup and position specific sequence variability an explicit part of the model. Having this information readily available will facilitate further investigations, especially for large-scale studies where hundreds or more factors are involved.

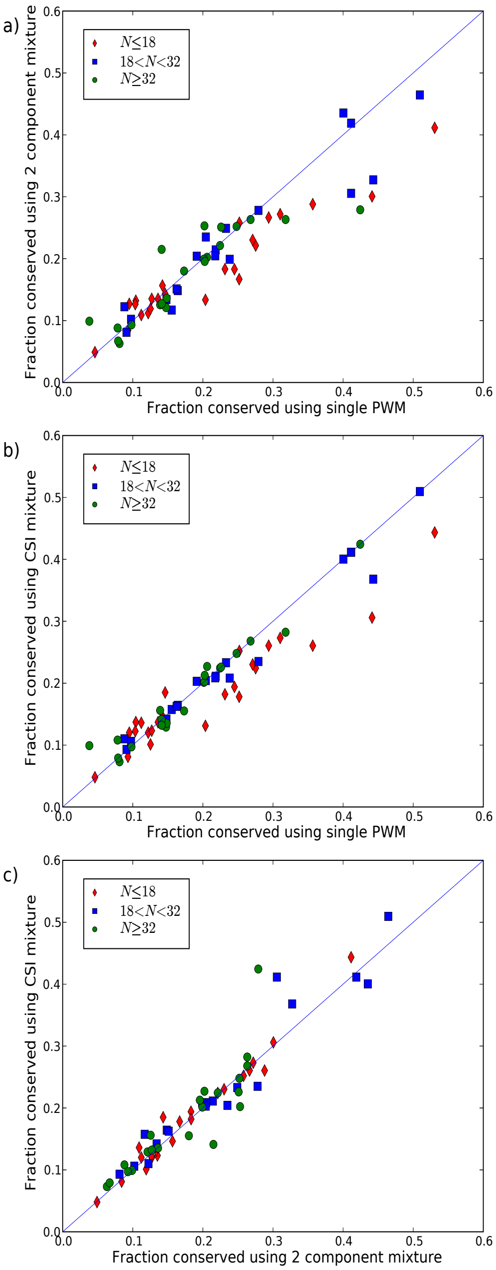

Fig. 5. a) Conserved fractions of hits for M_1 and M_2. The mixture M_2 is as good or better for 67% (43) of the TFs. **b)** Conserved fractions for M_{CSI} and M_1. For 70% (45) of the TFs the conservation of M_{CSI} was as good or better than for M_1. Outliers with strong preference for M_1 model had very few known sequences. If we only consider TFs with at least 20 sequences, the CSI yields as good or better conservation in 85% (34/40) of the cases. **c)** Comparison of conservation statistics of M_2 and M_{CSI}. For 89% (57) of the TFs M_{CSI} yields higher or equal conservation.

3.3 Conservation Statistics

The validation of predicted binding sites with respect to their biological functionality is a difficult problem as functionality cannot be assessed directly. One surrogate for functionality found in the literature is the degree of conservation in genomic sequences between related species [35]. For the sake of comparability with the results reported in [14] we follow the same evaluation approach taken there and evaluate the different models by the fraction of conserved predicted binding sites.

In the following we are going to evaluate the performance of a single PWM M_1, a two component mixture M_2 and a two component CSI mixture M_{CSI} based on human-mouse conservation. We used the same 64 JASPAR TFs as in [14]. We downloaded the 1kb upstream regions of the **hg17** assembly (May 2004) from the UCSC genome data base [16]. The mouse conservation data (**mm7**) was extracted from the axtNet data set [28] (also UCSC). For each of the 64 TFs and each of the three models under consideration, we then computed the 1000 best scoring hits in the 1kb upstream regions. The overall base composition of the sequences was used as the background model. For the mixtures the hits were chosen proportionally to the mixing weights. This means that for a $\theta_\pi = (0.6, 0.4)$ we would chose the 600 best hits from the first component and the 400 best from the second. The fraction of hits that was conserved in mouse was then computed based on a 80% sequence identity cutoff.

Evaluation: In order to decrease the impact of random variation on the analysis we considered TFs with very similar fractions of conserved hits for two model types as not giving conclusive preference to any of the two. That is, if the difference in the conserved fraction was less than ten percent of the maximal conserved fraction observed for any of the three model types, the scores were considered to be ”equal” for the purposes of this analysis. This has the effect of making the results more conservative in the sense that the impact of factors with very small differences in the conservation statistics was suppressed.

Fig. 5 shows the comparison of conserved fraction for the three model types. To illustrate the impact of the available number of training samples N for a factor on performance, we depict TFs differently based on the number of associated sequences. TFs with less than 18 sequences are shown as red diamonds, TFs with 19 - 31 sequences are shown as blue rectangles and TFs with more than 31 sequences are shown as green dots. The numbers were chosen as to split the 64 TFs into three roughly equally sized groups.

M_1 **vs** M_2: In 5a) you can see the conserved fraction of M_1 and M_2 for the 64 TFs in the data set. The mixture model M_2 was as good or better than M_1 in 67% (43) of the cases. For 33% (21) of the TFs the mixture was strictly better. This means that the performance of the two component mixture was somewhat weaker in our analysis than reported in [14] . Recall, that our data set differed from the one in [14] as it was based on a later genome freeze and, more importantly, it did not contain any downstream sequences. To the best of our knowledge the rest of our analysis was identical to the one conducted in [14].

M_{CSI} **vs** M_1: The comparison between the fraction of conserved hits of the CSI mixture M_{CSI} and the single PWM model M_1 can be seen in Fig. 5b). In 70% (45) of the TFs under consideration M_{CSI} showed a conserved fraction as good or better than M_1, with 28% (18) being strictly better. One important observation is that in most instances where M_1 had a strong advantage in conserved hits, the factor had only a small number of known binding sites. This can be seen by the large number of diamonds below the diagonal. For instance the rightmost point in Fig. 5b) at (0.53, 0.43) corresponds to MA0062 which has 7 known sites. In such a situation a little random variation in the sequences can have a strong impact on the trained model and lead to spurious structures. This is supported by the correlation between the number of available sequences for a factor and the increase in conservation for the CSI model. If we only considered TFs with 15 or more sequences, M_{CSI} is as good or better in 74% (40/54) of the cases, for 20 or more sequences in 85% (34/40) and for 40 or more in 94% (15/16). The fraction of TFs where M_{CSI} is strictly better remained in the range of 30% independent of the number of sequences.

	$M_2 \geq M_1$ (43)	$M_1 > M_2$ (21)
$M_{CSI} \geq M_2$	84% (36)	100% (21)
$M_{CSI} > M_2$	47% (20)	81% (17)
$M_{CSI} \geq M_1$	89% (38)	33% (7)
$M_{CSI} > M_1$	37% (16)	10% (2)

Table 2. Comparison of the conserved fraction of the 1000 best scoring hits for M_{CSI}, M_1 and M_2 in the two subsets of the TF data given the conditions $(M_2 \geq M_1)$ and $(M_1 > M_2)$ respectively.

M_{CSI} **vs** M_2: In Fig. 5c) we show the fraction of conserved hits for M_{CSI} and the conventional two component mixture M_2. For 89% (57) of the TFs the CSI model yields higher or equal conservation, 58% (37) being strictly greater.

Performance of M_{CSI}: Applying the two conditions $(M_2 \geq M_1)$ and $(M_1 > M_2)$ on the conserved fractions of hits split the 64 TFs in two subsets of size 43 and 21. We can think of the first subset as those TFs where a mixture model is appropriate and the second subset as being better represented by a single PWM. In the following we examined the performance of our CSI models within these two subsets. The results are summarized in Table 2. For the subset induced by $(M_2 \geq M_1)$ M_{CSI} was as good or better then M_1 or M_2 for a strong majority of 84% (36) and 89% (38) of the TFs respectively. M_{CSI} was strictly better for 47% and 37% respectively. This means that for TFs where a two component mixture improves performance as compared to a single PWM, the CSI model will in most cases outperform both of the other models. M_2 due to the reduction in overfitting and the more robust parameter estimates, M_1 because of the improved description of the binding pattern.

For the subset where a single PWM yielded a larger conserved fraction than the two component mixture (given by the condition $(M_1 > M_2)$) M_{CSI} was as good or better than M_2 for all the TFs in the subset (100% (19)) and strictly better for 81% (17). This illustrates the property of the CSI model to adapt to the number of subgroups supported by the data (one in this case) by means of the structure learning. M_{CSI} is equivalent or better than M_1 in 33% (7) of the TFs in the subset. This rather low number again shows the impact of spurious structures for TFs with few known binding sites. If we only consider the 11 TFs in the subset with 20 or more annotated binding sites, the value for $(M_{CSI} \geq M_1)$ goes up to 64% (7/11). Finally, M_{CSI} is strictly better than M_1 for a negligible 10% (2). This is not surprising as we would not expect CSI to outperform M_1 in situation where a single PWM is the appropriate model. Rather a successful application of the structure learning in such a case makes M_{CSI} equivalent to M_1. This corresponds to the points which lie directly on the diagonal (i.e. the conserved fractions are equal) in Fig. 5b).

4 DISCUSSION

The results of our simulation studies show that the CSI formalism yields more parsimonious and robust representations for TFs that exhibit a position-wise subgroup structure in their binding pattern. The greater parsimony of the CSI model as compared to conventional mixtures was demonstrated for a subgrouping of known Leu3 binding sites. In this example CSI required 30% less parameters than a conventional mixture for equal performance. The analysis of the conserved fraction of predicted binding sites in human upstream regions in mouse showed that a two component CSI model is clearly superior to a conventional two component mixture. This means that learning the CSI structures led to a more biologically meaningful characterization of the binding patterns of the TFs under consideration. For the TFs where the CSI model increased performance, we can assess that the known binding sites apparently exhibited a biologically relevant subgroup structure. The exact nature of the biological mechanisms underlying these subgroups remains elusive at this point. One possible explanation though would be the existence of different conformations of the TFs which show distinct binding patterns.

A strong advantage of the CSI (or conventional mixture) model over the single PWM model could not be observed on this data set. This was due to the occurrence of spurious structures for TFs with very few known binding sites and the large number of TFs where the single PWM model seems to be appropriate. This makes sense as one would expect the structure learning to be more vulnerable to outliers in situations where data is extremely sparse. The conclusion we draw from this result is twofold: First, CSI is a practical tool for the search for putative TFBS that fits in seamlessly within the probabilistic framework for scoring hits that has been established for the single PWM model (e.g. [18]). For a practical analysis using CSI though it seems important to require a minimum number of available binding sites (say 18) in order to attempt to fit a CSI model and to use the single PWM model otherwise. This could be easily included into the model prior. Secondly, we would expect the general usefulness of the CSI approach to increase in the future as the pool of known confirmed binding sites increases.

For future research we consider the development of more complex structure priors and improvements to the structure learning algorithm for sparse data. Also, it might be interesting to quantify the impact of different sequence scoring schemes on the performance the model. Moreover, since the probabilistic framework we work in is fully general, there are numerous biological applications where our method might yield improved results. In particular we consider applying our methods on donor splicing site detection, as larger data sets are available in this setting.

5 ACKNOWLEDGEMENTS

We would like to thank Martin Vingron and Gunnar Rätsch for helpful discussions and the reviewers for the excellent feedback they provided.

REFERENCES

[1] M. Aigner. A characterization of the Bell numbers. *Discrete Math.*, 205:207–210, 1999.

[2] H. Akaike. Information theory and the extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281, 1973.

[3] Y. Barash and N. Friedman. Context-specific bayesian clustering for gene expression data. *J Comput Biol*, 9(2):169–91, 2002.

[4] Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. Modeling dependencies in protein - dna binding sites. In *Proceedings of RECOMB '03*, pages 28–37, New York, NY, USA, 2003. ACM Press.

[5] Yonatan Bilu and Naama Barkai. The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol*, 6(12):R103, 2005.

[6] Hamid Bolouri and Eric H Davidson. Modeling DNA sequence-based cis-regulatory gene networks. *Dev Biol*, 246(1):2–13, Jun 2002.

[7] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

[8] David Maxwell Chickering and David Heckerman. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Mach. Learn.*, 29(2-3):181–212, 1997.

[9] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–38, 1977.

[10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[11] Nir Friedman. Learning belief networks in the presence of missing values and hidden variables. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 125–133, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[12] Nir Friedman and Moises Goldszmidt. Learning bayesian networks with local structure. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 421–459, Norwell, MA, USA, 1998. Kluwer Academic Publishers.

[13] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis, 2nd edition*. CRC Press, 2003.

[14] S. Hannenhalli and L.-S. Wang. Enhanced position weight matrices using mixture models. *Bioinformatics*, 21 Suppl 1:i204–i212, Jun 2005.

[15] G Z Hertz and G D Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577, Jul 1999.

[16] A.S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*, 34(Database issue):590–598, Jan 2006.

[17] Ekaterina A Kotelnikova, Vsevolod J Makeev, and Mikhail S Gelfand. Evolution of transcription factor DNA binding sites. *Gene*, 347(2):255–263, Mar 2005.

[18] S. Levy and S. Hannenhalli. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome*, 13(9):510–514, Sep 2002.

[19] Xiao Liu and Neil D Clarke. Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J Mol Biol*, 323(1):1–8, Oct 2002.

[20] M Z Ludwig, N H Patel, and M Kreitman. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, 125(5):949–958, Mar 1998.

[21] T K Man and G D Stormo. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*, 29(12):2471–2478, Jun 2001.

[22] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[23] G.J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[24] Mark Ptashne. *A Genetic Switch: Gene Control and Phage l*. Cold Spring Harbor Laboratory Press, 2004.

[25] H. Robbins. An empirical bayes approach to statistics. In *Proc. Third Berkeley Symposium on Math. Statist. and Prob.*, pages 157–164, 1956.

[26] Albin Sandelin, Wynand Alkema, Par Engstrom, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):91–94, Jan 2004.

[27] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990.

[28]Scott Schwartz, W James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C Hardison, David Haussler, and Webb Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, Jan 2003.

[29]G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[30]R Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res*, 12(1 Pt 2):505–519, Jan 1984.

[31]R Staden. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci*, 5(2):89–96, Apr 1989.

[32]A. Stolcke and S. M. Omohundro. Best-first model merging for hidden Markov model induction. Technical Report TR-94-003, 1947 Center Street, Berkeley, CA, 1994.

[33]G D Stormo. Consensus patterns in DNA. *Methods Enzymol*, 183:211–221, 1990.

[34]G D Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000. Historical Article.

[35]J.W. Thomas, J.W. Touchman, R.W. Blakesley, GG Bouffard, SM Beckstrom-Sternberg, EH Margulies, M Blanchette, AC Siepel, PJ Thomas, JC McDowell, B Maskeri, NF Hansen, MS Schwartz, RJ Weber, WJ Kent, D Karolchik, TC Bruen, R Bevan, DJ Cutler, S Schwartz, L Elnitski, JR Idol, AB Prasad, SQ Lee-Lin, VV Maduro, TJ Summers, ME Portnoy, NL Dietrich, N Akhter, K Ayele, B Benjamin, K Cariaga, CP Brinkley, SY Brooks, S Granite, X Guan, J Gupta, P Haghighi, SL Ho, MC Huang, E Karlins, PL Laric, R Legaspi, MJ Lim, QL Maduro, CA Masiello, SD Mastrian, JC McCloskey, R Pearson, S Stantripop, EE Tiongson, JT Tran, C Tsurgeon, JL Vogt, MA Walker, KD Wetherby, LS Wiggins, AC Young, LH Zhang, K Osoegawa, B Zhu, B Zhao, CL Shu, PJ De Jong, CE Lawrence, AF Smit, A Chakravarti, D Haussler, P Green, W Miller, and ED. Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793, Aug 2003.

[36]William Thompson, Michael J Palumbo, Wyeth W Wasserman, Jun S Liu, and Charles E Lawrence. Decoding human regulatory circuits. *Genome Res*, 14(10A):1967–1974, Oct 2004.

[37]K L Wang and J R Warner. Positive and negative autoregulation of REB1 transcription in Saccharomyces cerevisiae. *Mol Cell Biol*, 18(7):4368–4376, Jul 1998.

[38]T Werner. Models for prediction and recognition of eukaryotic promoters. *Mamm Genome*, 10(2):168–175, Feb 1999.

[39]J R Williams, C Thayyullathil, and N E Freitag. Sequence variations within PrfA DNA binding sites and effects on Listeria monocytogenes virulence gene expression. *J Bacteriol*, 182(3):837–841, Feb 2000.

[40]S A Wolfe, H A Greisman, E I Ramm, and C O Pabo. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J Mol Biol*, 285(5):1917–1934, Feb 1999.