BayCis: A Bayesian Hierarchical HMM for Cis-Regulatory Module Decoding in Metazoan Genomes

Tien-ho Lin[1,⋆], Pradipta Ray[1,⋆] , Geir K. Sandve[2], Selen Uguroglu[3], and Eric P. Xing[1,⋆⋆]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Dept of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway
[3] Dept of Computer Science and Engineering, Sabanci University, Istanbul, Turkey
epxing@cs.cmu.edu

Abstract. The transcriptional regulatory sequences in metazoan genomes often consist of multiple *cis-regulatory modules* (CRMs). Each CRM contains locally enriched occurrences of binding sites (motifs) for a certain array of regulatory proteins, capable of integrating, amplifying or attenuating multiple regulatory signals via combinatorial interaction with these proteins. The architecture of CRM organizations is reminiscent of the grammatical rules underlying a natural language, and presents a particular challenge to computational motif and CRM identification in metazoan genomes. In this paper, we present BayCis, a Bayesian hierarchical HMM that attempts to capture the stochastic syntactic rules of CRM organization. Under the BayCis model, all candidate sites are evaluated based on a posterior probability measure that takes into consideration their similarity to known BSs, their contrasts against local genomic context, their first-order dependencies on upstream sequence elements, as well as priors reflecting general knowledge of CRM structure. We compare our approach to five existing methods for the discovery of CRMs, and demonstrate competitive or superior prediction results evaluated against experimentally based annotations on a comprehensive selection of *Drosophila* regulatory regions. The software, database and Supplementary Materials will be available at http://www.sailing.cs.cmu.edu/baycis.

1 Introduction

Rules determining the spatio-temporal variations of gene expression in multi-cellular organisms are believed to be encoded as "*cis*-regulatory sequences", known to account for a large portion of a metazoan genome [15]. While recent years have seen substantial progress in *in silico* prediction of protein coding sequences from metazoan genomes, our understanding of the vocabulary and rules governing *cis*-regulatory sequences is limited, and remains a major open problem.

⋆ The first two authors in the list contributed equally to the paper and should be acknowledged as co-first authors.
⋆⋆ Correspondence should be addressed.

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 66–81, 2008.
© Springer-Verlag Berlin Heidelberg 2008

Unlike prokaryotes or uni-cellular organisms like yeast, metazoan transcription factor binding sites (TFBS, also known as motifs) are usually neither located immediately upstream of the proximal promoter element, nor are they distributed uniformly and independently in the extended surrounding region. Instead, the distributions of these motifs exhibit apparent general principles referred to as *modular organizations* – being organized into a series of discrete regions of roughly 200-1000 bp in length, each of which controls a distinct aspect of a gene's expression pattern [3]. Each CRM consists of a locally enriched collection of motifs of certain combination and ordering, capable of integrating, amplifying, or attenuating multiple regulatory signals via combinatorial physical interaction with multiple transcriptional regulatory proteins (i.e., TFs) [2]. Furthermore, it is believed that there also exist dependencies among CRMs so that coordinations between regulatory signals can be orchestrated.

Motif models of TFBSs for a single transcription factor have existed for many years, currently the most common model being the position weight matrix (PWM) introduced more than twenty years ago [25]. In recent years, focus has shifted from predicting TFBSs for a single TF towards predicting CRMs comprising several TFBSs, often for several distinct TFs. Several models have been proposed, making use of certain architectural features of the CRMs. Some of these models apply comparative genomic methods for CRM prediction [12,16,22,23]. These approaches are, however, restricted to very closely related organisms, because non-coding sequences are hard to align and more subject to events like duplication and shuffling which make orthology prediction difficult. A large number of CRM and motif prediction algorithms, including the one we propose in this paper thus rely on single species data.

One line of methods for the discovery of CRMs count the number of matches (of some minimal strength) to given motif patterns within a certain window of DNA sequence [19,21,20,4]. From a modeling point of view, this family of algorithms assumes that motifs are uniformly and independently distributed within each window; an *ad hoc* window size needs to be specified, and careful statistical analysis of matching strength is required to determine a good cutoff or scoring scheme [21,10]. Rajewsky *et al.* addressed the issue of compensating the matching scores for co-occurring weak motif sites using an updatable word frequency measure, leading to higher scores for motifs co-occurring more frequently within a given window size [1] [19].

A second line of methods takes an entirely different approach by modeling the occurrences of motifs and CRMs as the output of a first-order hidden Markov process. This approach alleviates the necessity of both the window size and the score cutoff, and takes into account not only the strengths of motif matches, but also the spatial distances between matches (arguably more informative than co-occurrence within a

[1] Their algorithm also contains an important extension for unsupervised CRM prediction, where representations of novel motifs are estimated directly from input DNA sequences. However, under a modular formulation of the CRM prediction problem (cf. the LOGOS model [30]), prediction of motif instances from given representations, and estimation of motif representations from predicted instances, can be treated as two orthogonal sub-problems to be solved separately and coupled as components of a higher-level joint model with estimates exchanged in iterative fashion. In this paper, we only focus on CRM prediction given motif representations and defer implementing the fully autonomous *de novo* motif-finding program to a later paper.

window). The hidden Markov model (HMM) translates to a set of soft specifications of the expected CRM length and the inter-CRM distance (i.e., in terms of geometric distributions). However, since training data for fitting the HMM parameters hardly exist, these parameters typically have to be specified based on empirical guesses. HMMs and similar models that captures TFBS distributions, as well as intra-CRM and inter-CRM backgrounds, have been used in several CRM discovery methods, e.g. in Cister [7], Cluster-Buster [6], CisModule [31] and EMCModule [9]. As these methods employ a general inter-motif background, they do not infer any ordering between motifs. This model is extended to include distinct motif-to-motif transition probabilities in the methods Stubb [24] and Module Sampler [27].

In this paper, we present a new method, *BayCis*, which implements a Bayesian hierarchical HMM for CRM search. *BayCis* represents a step further along the direction of HMM-based CRM models. It uses a more sophisticated HMM model that is intended to capture, to a reasonable degree, the detailed syntactic structure of CRM and cis-regulatory regions containing CRMs. By combining general intra-CRM background, motif specific background surrounding motif instances, as well as specific motif-to-motif transitions, it allows couplings between motifs to be captured. We also introduced more advanced approaches to model the background, using separate inter-CRM, intra-CRM and motif-specific higher-order Markov backgrounds. Furthermore, inter-motif distances may be modeled with more flexible distributions (rather than only simple geometric distributions). Finally, as detailed in the following sections, we treat parameters of the HMM grammar as stochastic variables for which Bayesian priors are applied, instead of regarding the state-transition parameters of the HMM grammar as fixed parameters that solely rely on empirical default values or user specification like in previous methods. This technique in principle alleviates user specification of model parameters (although advanced users could choose to decide the "strength" of the priors, or define their own priors). On the computational front, we developed an efficient variational inference algorithm for posterior inference of sequence annotation and Bayesian parameter estimation. This algorithm enjoys a desirable convergence guarantee and is much more efficient than the classical Gibbs sampling methods without compromising much accuracy.

BayCis has several advantages over existing methods for CRM discovery. The explicit model of CRMs makes architectural assumptions clear, and supports rich interpretation of results by analyzing likelihoods at states and transitions. The sophisticated modeling, including motif-to-motif specific transitions and several distinct background states should allow more specific CRM predictions at the same level of sensitivity. Finally, by relying on soft priors instead of hard specification of model parameters, the Bayesian approach adds generality and user convenience to the method.

2 Methods

To model the complex architecture of metazoan transcriptional regulatory sequences (TRS), we propose to use a *hierarchical hidden Markov model* (hHMM) that can encode a set of stochastic syntactic rules presumably underlying the CRM organizations and motif dependencies. A first-order Markov process over a hierarchy of states allows us

to describe the structure of regulatory regions at different levels of granularity, offering more modeling power than existing methods.

2.1 A Hierarchical HMM of TRS

As first proposed in [5], the hHMM is an extension of the classical HMM for modeling domains with hierarchical structures. In an hHMM, all hidden states are not equal, but follow a hierarchical organization that constrains stochastic transitions among states—transitions are only permissible for (certain pairs of) states at the same level or adjacent levels in the hierarchy; different states can emit either single observations or strings of observations, depending on their position in the state hierarchy; and the strings emitted from the non-leaf states in the hierarchy are themselves governed by a sub-hHMM (or more generally, by an arbitrary generative model, which would further extend the overall model beyond an hHMM).

An hHMM can explicitly capture nested generative structures (e.g., TRS → CRM → Motif → Single Nucleotide Site) underlying complex sequential data, and dependencies among elements at different levels of granularity (e.g., motif versus motif, site versus site, etc.), which makes it a powerful and natural approach to model genomic regions harboring transcriptional regulatory sequences. Fig. 1 shows an example of an hHMM encoding typical hierarchical structures of the metazoan TRSs we are concerned with in this study. At the top (i.e., coarsest) level, this hHMM represents a TRS as a concatenation of long stretches of sequences corresponding to global backgrounds and CRMs. We can think of this top level as an HMM whose states emit whole CRMs and inter-CRM (global) background sequences. Formally, we let $\mathbb{Q}^1 \equiv \{b_g, c_1, c_2, \ldots, c_I\}$ denote the set of these possible states. At the next level, each CRM is represented as a sequence of motifs and intra-CRM (local) background states. Accordingly we have $\mathbb{Q}^2 \equiv \{b_c, m_1, m_2, \ldots, m_K\}$. At a finer level below, each motif is represented as a sequence of buffer states and nucleotide sites. (We will explain shortly why we include non-motif buffer states at this level.) Accordingly, we define $\mathbb{Q}^3 \equiv \mathbb{B} \cup (\cup_i \mathbb{M}_i)$, where $\mathbb{B}$ corresponds to the non-motif buffer states padding right before and after the motif sequences and $\mathbb{M}_i$ corresponds to all possible sites within motif i. More

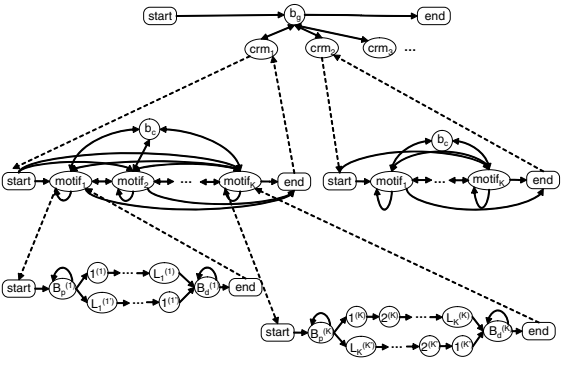

Fig. 1. The BayCis hHMM state transition diagram with 3-level hierarchy. Circular nodes represent functional states in DNA sequences, and round boxes represent start and end states in each sub-model. CRM and motif states are sub-models invoked by higher level models. Arrows between nodes represent permissible state transtions, including horizontal transitions denoted as black arrows, and verticle transitions denoted as dashed arrows.

specifically, we define: $\mathbb{M}_i \equiv \mathbb{M}_i^f \cup \mathbb{M}_i^r$, where $\mathbb{M}_i^f = \{1^{(i)} \dots L_i^{(i)}\}$ is the set of all possible sites within motif i on the forward DNA strand, and $\mathbb{M}_i^r$ is the set of all possible sites within motif i if it is on the reverse complementary DNA strand.; $\mathbb{B} \equiv \mathbb{B}^p \cup \mathbb{B}^d$, where $\mathbb{B}^p = \{b_p^{(1)}, \dots, b_p^{(K)}\}$ denotes the set of *proximal-buffer* states associated with each type of motif [2], and $\mathbb{B}^d = \{b_d^{(1)}, \dots, b_d^{(K)}\}$ denotes the set of *distal-buffer* states associated with each type of motif.

The possible transitions between these states are made explicit by the arrows in the hierarchical state diagram in Fig. 1. (To make the hHMM model well-defined, we also introduce *dummy* states `START` and `END` at appropriate levels to enable instantiation of state-traversal, and proper termination of subsequences at each level.) The biological motivation for such a state hierarchy is that we expect to see occasional motif clusters in a large ocean of global background sequences (represented by state b_g); each motif instance in a cluster is like an island in a sea of intra-cluster background sequences (b_c); and adjacent motifs may be statistically coupled (we will elaborate on this point in the next section). Our model assumes that the distance between clusters is geometrically distributed with mean $1/(1-\beta_{g,g})$, and the span of the intra-cluster background is also geometrically distributed with mean $1/(1-\beta_{c,c})$. These modeling choices are intended to not only reflect our uncertainty about the CRM structure, but also to offer substantial flexibility to accommodate potential 1st-order syntactic characteristics within the CRMs. In this hHMM, only the bottom-level motif-site and motif-buffer states, as well as the global and local background states, are capable of emitting individual nucleotides constituting the TRS, according to a stochastic emission model (which we will elaborate later). A stochastic traversal of the hHMM states according to the hHMM state-transition diagram would generate a TRS of arbitrary length but with a structure consistent with our empirical knowledge of the functional organization of the metazoan TRS. Note that this hHMM model does not impose rigid constrains on the number of motif instances or CRMs; the actual number of instances is determined by the posterior distribution of the hHMM states given the observed sequence. Also note that we have not included functional states related to gene annotation and basic promoters, but such extensions are straightforward if co-identification of CRMs and genes is desired.

Given the observed sequences, and proper (i.e., biologically meaningful) construction of the state space and its hierarchical organization, one can infer the latent state-traversal path, which correspond to a plausible annotation or segmentation of the input sequence, using a number of exact posterior inference algorithms. The original algorithms given by [5] is a variant of the inside-outside algorithm for stochastic context free grammar, and takes $O(T^3Q^D)$, where T is the length of the sequence, Q is the total number of states, and D is the depth of the hierarchy. A linear time algorithm was developed by [17] based on a transformation of hHMM into an equivalent dynamic Bayesian network. It is also possible to flatten the hHMM to an HMM with a block-structured sparse transition, and use a modified forward-backward algorithm for linear-time inference. In section 2.3 and

[2] Here, proximal-buffer refers to the background sites immediately next to the proximal-end of the motif. For consistency, orientations are defined with respect to the initial position of the input sequence. That is, the 1st position of the input sequence corresponds to the proximal end, and the last position corresponds to the distal end.

Supplementary Materials, we exploit this strategy, and develop an efficient algorithm for inference and learning under a Bayesian extension of hHMM to be described in the sequel.

Motif bigram via hHMM. An hHMM not only encodes hierarchical segmental structures in a sequence, but it can also be used to capture dependencies between sequence elements at different levels of granularity at a cost much smaller than that would be needed by a "flat" Markovian model which must resort to heavily parameterized high-order conditional probabilities. For example, we can capture the dependencies between neighboring CRMs in a TRS by modeling transitions between the CRM states. Of particular importance in this paper, we use hHMM to capture the dependencies between occurrences of motifs within a CRM. As discussed earlier, the spatial arrangement of motifs within a CRM may encode intricate combinatorial transcriptional regulatory signal. Thus modeling at least 1st-order dependencies between motifs may be beneficial to the unraveling of motifs in long TRS bearing complex regulatory function, as well-known in the case of Drosophila enhancers. Note that a direct transition between trivially defined motif states (e.g., last site of motif i and first site of motif j) would suggest that coupled motifs always occur right next to each other, which is biologically not always true. To capture possible dependencies between motifs in the vicinity of each other, we define the emission of a motif state (in $\mathbb{Q}^2$) to contain not only the motif sequence itself, but also non-motif sequences denoted as proximal and distal buffers. Such an emission can be understood as an extended instance of a motif, which we referred to as a *motif envelope*. Thus cross-background (i.e., high-order) dependencies between motifs can be captured by immediate (i.e., 1st-order) dependencies between the motif envelopes.

We write $A_2 \equiv \{a_{i,j}\}$ as the stochastic matrix for transitions among states in $\mathbb{Q}^2$, which defines a *bigram* of motifs (and their local backgrounds) within CRMs. The length of the proximal and distal buffers of a motif is geometrically distributed with mean $1/(1-\alpha_{i,i})$ and $1/(1-\beta_{i,i})$, and can be generated via self-transitions of the corresponding states at the third level (i.e., in $\mathbb{Q}^3$) with probability $\alpha_{i,i}$ and $\beta_{i,i}$, respectively. Then with equal probability $\alpha_{i,m}/2$, a proximal buffer state $b_p^{(i)}$ reaches the start states $1^{(i)}$ (resp. $L_i^{(i')}$) of motif i on the forward (resp. reverse) strand, deterministically passes through all internal sites of motif i, and transitions to the distal-buffer state $b_d^{(i)}$, thereby stochastically generating a non-empty motif envelope [3]. Each b_d^i has probability $\beta_{i,j}$ of transitioning to the proximal-buffer state of another motif j (or of the same motif when $j = i$) to concatenate another motif envelope, or it may choose to pad with some inter-cluster background before adding more envelopes, with probability $\beta_{i,c}$. All distal-buffer states also have probability $\beta_{i,g}$ of returning to the global background, terminating a CRM.

Spacer length distribution via GhHMM. A *spacer* is the interval seperating adjacent motif instances, modeled as b_c, b_p, and b_d states in BayCis. It has been suggested that the

[3] The distinction between proximal and distal buffers avoids generating empty envelops (otherwise, a single buffer state wont be able to remember if a motif has been generated beyond k positions prior to the current position under a k-th order Markov model).

range of spacer length is under selection forces according to comparative genomics data of several *Drosophila* species [13]. Empirically, we found that the distribution of spacer lengths can be approximated by a negative binomial distribution (see figure in Supplementary Materials), whereas under an hHMM, the state durations of cluster backgrounds is distributed as a goemetric distribution, which is not a good approximation of the space length distribution. In Supplementary Materials, we describe a generalized hierarchical hidden Markov model (GhHMM) which implements an approximate negative binomial distribution of spacer lengths by joining several geometrically distributed cluster background states.

The emission models: PWM and higher-order Markov background. Once the hHMM enters the motif-site states, we resort to a *motif model* to generate the nucleotides at the corresponding sites. To maintain our focus on the hHMM and relevant algorithmic issues, we only consider the scenario of searching for known motifs in this paper (although extending our model for *de novo* motif detection is straightforward based on, for example, the LOGOS framework [30]). For motif model we choose the classical product-multinomial (PM) model, which can be represented by a PWM [25].

Several previous studies have stressed the importance of using a richer background model for the non-motif sequences [26,11]. In accordance with these results, BayCis uses a standard global k-th order Markov model for the emission probability of the global background state. For the intra-CRM states, we used locally estimated Markov models. Since the models are defined to be *local*, the conditional probability of a nucleotide at position t is now estimated only from a window of length $2d$ centered at t. These probabilities can still be computed off-line and stored for subsequent uses, by using a careful bookkeeping scheme (i.e., using a "sliding-window" to compute the local Markov model of each successive position, each with a constant "update cost" based on the previous one).

2.2 Bayesian hHMM

One caveat of the standard HMM approach for CRM modeling is the difficulty of fitting the model parameters, such as the state-transition probabilities, due to rarity of fully annotated CRM-bearing genomic sequences. In principle, using the Baum-Welsh algorithm one can learn the maximal-likelihood (ML) estimates of the model parameters directly from the unannotated sequences while analyzing them. In practice, however, such a completely likelihood-driven approach tends to result in spurious results, such as overestimation of the motif and CRM frequencies and poor stringency of the learned models for potential motif patterns. Previous methods tried to overcome this by reducing the number of parameters needed as much as possible, and by setting them according to some good guesses of the motif/CRM frequencies or CRM sizes [7]. But as a result, such remedies compromise the expression power of the already simple HMM, and risk mis-representing the actual CRM structures. In the following, we propose a Bayesian approach that introduces the desired "soft constraints" and smoothing effect for an HMM of rich parameterization, using only a small number of *hyper-parameters*. This approach defines a posterior probability distribution of all possible value-assignments of the HMM parameters, given the observed un-annotated sequences and empirical prior

distributions of the parameters that reflect general knowledge of CRM structures. The resulting model allows probabilistic queries (i.e., estimating the probability of a functional state) to be answered based on the aforementioned posterior distribution rather than on fixed given values of the HMM parameters.

We assume that the self-transition probability of the global background state $\beta_{g,g}$, and the total probability mass of transitioning into a motif-buffer state $\sum_{k\in\mathbb{B}^p}\beta_{g,k}$ (note that $\beta_{g,g} = 1 - \sum_{k\in\mathbb{B}^p}\beta_{g,k}$), admit a beta distribution, $Beta(\xi_{g,1}, \xi_{g,2})$. We choose a small value for $\frac{\xi_{g,2}}{\xi_{g,1}+\xi_{g,2}}$, corresponding to a prior expectation of a low CRM frequency. Similarly, we define a beta prior $Beta(\xi_{c,1}, \xi_{c,2})$ for the self- and total motif-buffer-going transition probabilities $[\beta_{c,c},\ \sum_{k\in\mathbb{B}^p}\beta_{c,k}]$ associated with the intra-cluster background state; and another beta prior $Beta(\xi_{p,1}, \xi_{p,2})$ for the self- and motif-going transition probabilities $[\alpha_{i,i},\ \alpha_{i,m}]$ associated with the proximal-buffer state of a motif. Finally, we assume that for the distal-buffer state, the self-transition probability, the total mass of transition probabilities into a proximal-buffer state, the probability of transitioning into the intra-cluster background, and the probability of transitioning into the global background, $[\beta_{i,i},\ \sum_{k\in\mathbb{B}^p}\beta_{i,k},\ \beta_{i,c},\ \beta_{i,g}]$, admit a 4-dimensional gamma distribution, $Gamma(\xi_{d,1}, \xi_{d,2}, \xi_{d,3}, \xi_{d,4})$.

To define priors for the GhHMM parameters, the GhHMM with a single cluster background state (b_c) is considered as an HMM with several cluster background states ($\{b_c^1, \cdots, b_c^{g_{cr}}\}$) sharing the same self-transition probability $\beta_{c,c}$. Similar to other background states, we define a beta prior $Beta(\xi_{c,1}, \xi_{c,2})$ on the total probability mass of transitions into motif-buffer states $\sum_{k\in\mathbb{B}^p}\beta_{c,k}$ (note that $\beta_{c,c} = \sum_{k\in\mathbb{B}^p}\beta_{c,k}$).

Note that due to conjugacy between the prior distributions described above and the corresponding transition probabilities they model, the hyper-parameters of the above prior distributions can be understood as *pseudo-counts* of the corresponding transitioning events, which can be roughly specified according to empirical guesses of the motif and CRM frequencies. But unlike the standard HMM approach, of which the transition probabilities are fixed once specified, the hyper-parameters only lead to a soft enforcement of the empirical syntactic rules of CRM organization in terms of prior distributions, allowing controlled posterior update of the HMM transition probabilities while analyzing the un-annotated sequences. For the BayCis hHMM, we specify the hyper-parameters (i.e., the pseudo-counts) using estimated frequencies of the corresponding state-transition events, multiplied by a "prior strength" N, which corresponds to an imaginary "total number of events" from which the estimated frequencies are "derived". That is, for the beta priors, we let $[\xi_{[\cdot,1]}, \xi_{[\cdot,2]}] = [1-\omega_{[\cdot]},\ \omega_{[\cdot]}] \times N$, where the "$\cdot$" in the subscript denotes either the g, c, or p state, and $\omega_{[\cdot]}$ is the corresponding frequency. For the gamma prior, we let $[\xi_{d,1}, \xi_{d,2}, \xi_{d,3}, \xi_{d,4}] = [\omega_{d,1},\ 1-\sum_j \omega_{d,j},\ \omega_{d,2},\ \omega_{d,3}] \times N$. Overall, we need to specify 7 hyper-parameters (of course one can use different "strengths" for different priors, with a few additional parameters), a modest increase compare to, say, 3 needed in Cister [7].

2.3 Inference and Learning

We have developed an efficient algorithm called *modified FB-algorithm* for inference on a "flattened" hHMM, which reduces the time complexity of the standard forward-backward algorithm from $O(K^2\bar{L}^2T)$ to $O(K^2T)$. Identification of motifs/CRMs is

based on posterior decoding. We also developed a *variational EM algorithm* for Bayesian inference and parameter estimation under our Bayesian hHMM and GhHMM, which is much more efficient than the traditional MCMC sampling approaches. Due to space limit, details of these algorithmic innovations are given in the Supplementary Materials.

3 Results

We evaluated BayCis on both synthetic transcriptional regulatory sequences and a rich set of carefully compiled real genomic TRSs of *Drosophila melanogaster* (available at our website). The prediction performance of BayCis was compared with 5 popular published methods for supervised discovery of motifs/CRMs based on a wide spectrum of models: Cister [7], Cluster-Buster [6], MSCAN [1], Ahab [19] and Stubb [24] (all of which were applied to the real data, and two seemingly superior ones to the semi-synthetic data), which cover a wide spectrum of different models/algorithms (e.g., HMMs, windows) for motif search. We ran other methods with default parameters, specifying 500 bp CRM window where needed.

Overall, the prediction performance of BayCis is competitive or superior to all chosen benchmark methods on this quite comprehensive selection of data sets, according to a wide assortment of performance measures. By employing sound and flexible probabilistic modeling of regulatory regions, BayCis is also able to strike a good balance between precision and recall with its default MAP solution.

3.1 Semi-realistic Synthetic TRS

Synthetic TRSs are useful in that the ground truth for motif/CRM locations is known exactly. To generate semi-realistic synthetic TRSs, we planted selected TFBS from the Transfac [29] database in simulated background sequences according to model assumptions underlying the background distribution, the inter-TFBS and inter-CRM spacer length distributions for Baycis. 30 sequences of length 20,000 bp containing 0 - 3 CRMs were generated. The CRM length is uniformly distributed between 200 and 1600 bp, while the average motif spacer length is 50 bp. Each CRM contains 3 to 6 motif types and about 14 motif instances. To simulate motif co-occurrence, about 25% of the motif instances in each CRM appear as predefined pairs. The background sequences

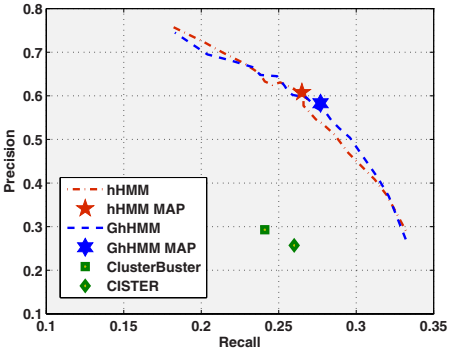

Fig. 2. The precision-recall (P/R) curves of two models of BayCis (hHMM and GhHMM) versus the P/R of default predictions by CISTER and ClusterBuster

inside/outside the CRM are simulated by a 3rd-order Markov model learnt from an intergenic region.

As shown in Fig. 2, the performance of BayCis using either hHMM or GhHMM is significantly better than CISTER and ClusterBuster in terms of the overall precision/recall (P/R) trade-off at the MAP prediction. The P/R curve of BayCis is also well above the default predictions from other methods. It also shows that GhHMM performs consistently better than hHMM in both precision and recall, although the difference is not very large. CISTER and ClusterBuster were chosen for the simulation study based on their good performance on real data (see next subsection).

3.2 Real Drosophila TRS

The dataset. The synthetic TRSs are generated partially based on the same model assumptions underlying BayCis, and thus the results cannot be interpreted as conclusive. A systematic investigation of the robustness of BayCis with respect to a wide spectrum of simulation conditions can be highly interesting but is beyond the scope of this short report; we will pursue this in a later full version of the paper. In this section we present an empirical evaluation based on a rich and carefully compiled Drosophila TRS dataset, although it is noteworthy that even though we have tried our best to gather the most complete annotations for each test sequence based on footprinting results from the literature, this "gold standard" is still possibly only a subset of the ground truth.

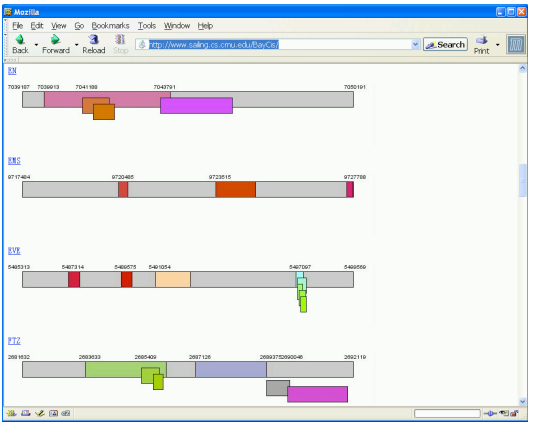

Fig. 3. Frontpage screenshot of the motif database

We created a manually curated dataset containing 97 CRMs pertaining to 35 early developmental genes (see table in Supplementary Materials for details). This collection was compiled based on a filtering of all known CRMs from a number of public databases (e.g., the REDfly CRM database [8] and the Drosophila Cis-regulatory Database at the National University of Singapore [18]), through which we only chose CRMs that are at least 200 bp long, and contain at least 5 experimentally confirmed motif instances (2 CRMs with a borderline count of 4 motif instances were also included). Each test sequence consists of the CRMs pertinent to a particular gene, all intra-CRM background inbetween, with flanking regions on either side of the extremally located CRMs such that the entire sequence is at least 10 kbp long, and the boundaries of the sequence are at least 2 kbp from the extremal CRMs. We included the exonic regions of the genes only when they fell in the aforementioned selected region, and not otherwise. This database is available at `http://www.sailing.cs.cmu.edu/BayCis`, where the BayCis software will soon be also released. A snapshot of the interface of

graphical interface of the database shown in Fig. 3, and more details are available in Supplementary Materials.

Experimental setup. BayCis is a Bayesian framework based on hHMMs and GhHMMs to model the organization and distribution of TFBS. Prior beliefs pertaining to the parameters of the model thus could be specified by the user before running on experimental data in the form of hyperparameters (i.e., pseudocounts) of the hHMM or GhHMM parameters. The PWMs of the motifs to be searched for also need to be provided because here we are interested in identifying TFBS of existing TF motifs, rather than *de novo* motif detection. As mentioned in previous sections, extending BayCis for this function is straightforward by introducing an EM step for the PWM estimation, and will be pursued in a later paper.

Hyperparameters: The choice of hyperparameters should in principle be dealt with via an "empirical Bayes scheme", which employs maximal likelihood estimates of these hyperparameters based on some fully labeled training sequences. Upon prediction on an unannotated sequence, the hHMM or GhHMM parameters themselves can be adjusted in an unsupervised fashion via the variational EM algorithm. We specify the hyperparameters as follows: for the global background, $\omega_g = 0.002$; for the inter-CRM background, $\omega_c = 0.05$; for the proximal motif buffer, $\omega_p = 0.25$; for the distal buffer hyperparameters, $\omega_{d,1} = 0.125$ (distal to global background), $\omega_{d,2} = 0.125$ (distal to clustal background), and $\omega_{d,3} = 0.25$ (distal to proximal buffer). Finally, the "strength" of the hyperparameters are set to $1/10$ of the expected counts of the transitions on a 15 kbp dataset, with the exception of ω_g which is set to $10,000$. The background probability of the nucleotide at each position was computed locally using a 2nd-order Markov model from a sliding window of 1100 bp centered at the corresponding position. For the GhHMM, based on visual inspection of spacer length distributions between motifs, we choose the parameter as $r = 2$.

Prediction scheme: BayCis provides three kinds of prediction schemes for motifs. The *maximum a posteriori* (MAP) prediction is based on the posterior probabilities of the labeling state at each site, which allows overlapping motifs. A Viterbi prediction, which gives a consistent prediction in the Bayesian setting analogous to an ML prediction under a classical setting can also be used. A third scheme is based on a simple but effective thresholding scheme where we directly predict motifs based on whether the motif states have a higher probability than the specified threshold in the posterior probabilities. For simplicity, in this paper we only present the MAP results and the P/R curve of the threshold method. Note that unlike many other scoring schemes for motif/CRM detection, such as logodds (i.e., the PSSM score) or a likelihood score regularized by word frequencies, our MAP prediction does not require a cutoff value for the scores, nor a window to measure the local concentration of motif instances, both of which are difficult to set optimally.

Evaluation measures: There is no unanimous way of evaluating the prediction performance of a motif/CRM discovery method against annotations. To avoid reliance on a single evaluation procedure and measure, we have chosen to present the performance

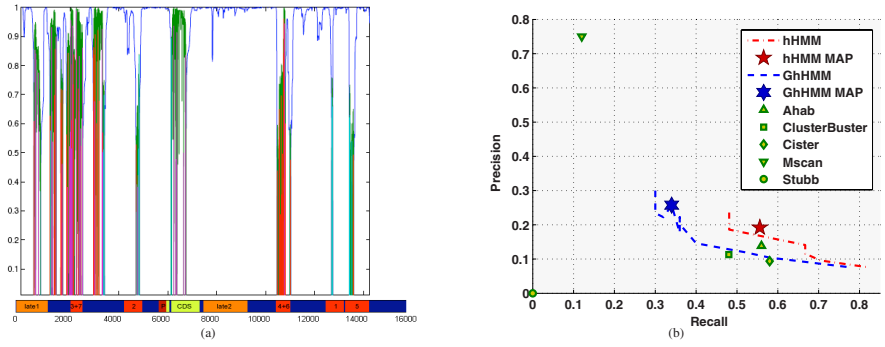

Fig. 4. Performance of BayCis (hHMM) on a representative *eve* TRS. (a) The posterior probability plot of the global background (blue), cluster background (green) and motif specific (red and other colors) states. (b) The precision versus recall performance of the MAP and thresholded predictions of the hHMM and GhHMM algorithms, as compared to those made by other methods.

of BayCis in comparison with other methods using several different evaluation procedures. This also ensures a thorough and objective presentation of results. For an overall evaluation we compare the prediction performance of BayCis with other methods using both the F1-score of precision and recall, and the coefficient of correlation (CC) score at nucleotide-level [28] as single point measures (see Supplementary Materials B.3 for detailed definitions). We do this by first summing true/false positives/negatives across datasets at the nucleotide level, and then computing F1/CC from these combined counts. To present the behavior of BayCis with respect to site-level P/R, we plot the binding-site level P/R curve from different thresholds in extracting predictions, along with the P/R at MAP predictions.

Motif prediction performance. As an illustration, Fig. 4a shows a plot of the MAP prediction along the *even-skipped* gene TRS, under a particular hyperparameter setting. As revealed in the ground-truth annotation bar bellow the plot, this region contains 5 CRMs (from left to right): *stripe3+7*, *stripe2*, *stripe4+6*, *stripe1*, and *stripe5*. BayCis picks out all of them, although the CRM boundary appears to be more stringent in most cases. We believe this can be improved by adopting a more specialized cluster background model (i.e., local higher-Markov model, better GhHMM model, etc.), which we have not fully explored yet. BayCis also identifies motif-rich regions proximal and distal to the *stripe3+7* CRM, which is not reported before, and it also finds another putative motif-rich region spanning the core promoter and the CDS of *eve*, which can be a false positive or a putative CRM. The overall MAP prediction score of BayCis, and the P/R curves resulted from applying different threshold values under BayCis, are shown in Fig. 4b, along with the scores of 5 other competing algorithms in their default configurations. The BayCis MAP predictions seem significantly better than other methods, and strike a good balance between recall and precision. It is important to realize that although the threshold method can reach high precision or recall at both extremes, in practice it is very hard to pick the optimal threshold without knowing the prediction results, and typically a threshold optimal for one sequence is not necessarily good for

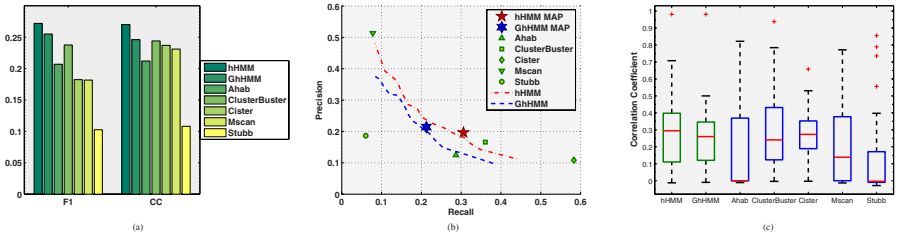

Fig. 5. (a) F1 and CC scores, and (b) P/R performances of the MAP and thresholded predictions of the hHMM and GhHMM, in comparison with other algorithms on the full *Drosophila* TRS dataset (c) A boxplot showing variation in CC across datasets

another sequence; significance-test based determination of threshold is also difficult for a complex model or large sequence. Thus, a default prediction such as MAP, which automatically finds an appropriate trade-off between precision and recall, is highly desirable.

The overall CC and F1-scores of running BayCis and five competing methods on the full set of *Drosophila melanogaster* sequences are shown in Fig. 5a. According to either measure, both the hHMM and the GhHMM version of BayCis outperforms all existing methods. The hHMM version of BayCis performs slightly better overall compared to GhHMM according to both measures. For both versions of BayCis, the MAP solution was chosen.

To look at the behavior of BayCis in the P/R landscape on our entire dataset, we plot the P/R curve resulting from different thresholds for BayCis predictions. For other methods we provide the single points in P/R landscape corresponding to their default output. As is apparent from Fig. 5b, the 5 competing methods strike different balances between precision and recall in their default output. MSCAN focuses on very high precision predictions, while Cister is geared towards high values of recall. The P/R curves of both versions of BayCis span a balanced range in the P/R landscape, with MAP estimates lying in the middle of the curves. Again, in practice the P/R values are not available for use by methods, so the balance between precision and recall has to be found based solely on the input data. Thus the ability to appropriately balance precision and recall automatically is essential.

To further investigate the prediction performance, we look at the variation of individual dataset prediction performance across all datasets. The boxplot in Fig. 5(c) shows the median CC-score for each method, as well as upper and lower quartiles and minimum/maximum values. We see that prediction scores varies much between datasets for all methods, and that the overall performance differences between methods is not very large compared to the variation of individual methods across datasets. This confirms what has long been acknowledged in the motif discovery field, that even the best performing methods will in many cases give misleading predictions (although some of the low scores may be due to lack of annotations). Among the high scoring methods (hHMM, GhHMM, Cluster-Buster and Cister), GhHMM and Cister come out as the most stable with low variance across datasets, a criterion which is useful when handling

a varied set of data. The posterior expectations of the hHMM/GhHMM parameters also carry rich architectural information of each TRS we processed, and merits systematic analyses. We defer this investigation to the full paper.

4 Discussion

BayCis uses an advanced probabilistic framework to accurately model metazoan transcriptional regulatory genomic sequences — which often consist of multiple CRMs, tandemly joined by long stretches of background DNA, each containing locally enriched occurrences of binding motifs for a certain array of transcriptional regulatory proteins. Thus, we are able to detect many TFBS while avoiding too many false positives and (slightly) outperform the best of the existing methods on a comprehensive set of *Drosophila* regulatory regions. The BayCis software will soon be released on our website.

Recently, experimental results have shown that sequences immediately flanking a TFBS may contribute to the binding energy between a TF and the TFBS [14]. This suggests that sequence composition of the proximal and distal buffers of motifs may have weak type specificity, which we would like to explore in our future work. Our current TRS database for performance evaluation is still limited in size and very diverse in terms of CRM structures and complexity, which could cause BayCis to overfit certain TRS when it is applied independently to each TRS separately (as we did in this paper), using a generic set of hyperparameters that are empirically chosen. We intend to adopt a more systematic approach to fit the hyperparameters based on a small amount of labeled TRS, e.g., using a k-fold cross validation scheme. But ultimately, we believe additional TRS data will be needed to attain further performance increase. One direction of increasing input data is to combine regulatory regions of several genes that are believed to share similar CRM structure. Such gene sets should be attainable for many real scenarios where CRM discovery methods are used, could trivially be used as input to BayCis. We speculate that this could improve predictions. The limitation lies mostly in collecting such gene sets containing rich, high-quality annotations that could serve in quantitatively measuring correspondence between computational prediction and experimental determination.

Another direction is to conjoin BayCis with a phylogenetic model of motifs across species [16,22,23], and apply the integrant to orthologous TRSs. Although this limits the applicability of the approach to species where valuable orthologous sequence is available, and to the discovery of regulatory elements shared between species, we believe it could attain considerably performance gain in the cases for which it is suited.

Acknowledgements. This material is based on work supported by the Pennsylvania Dept of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739, and by an NSF CAREER Award under Grant No. DBI-054659. The authors thank Wenjie Fu for result analysis, Jostein Johansen for help with evaluating CRM predictions, and Oznur Tastan for investigating the spacer length distributions.

References

1. Alkema, W.B., Johansson, O., Lagergren, J., Wasserman, W.W.: Mscan: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res. 32(Web Server issue), 195–198 (2004)
2. Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA 99(2), 757–762 (2002)
3. Davidson, E.H.: Genomic Regulatory Systems. Academic Press, London (2001)
4. Donaldson, I.J., Chapman, M., Gottgens, B.: Tfbscluster: a resource for the characterization of transcriptional regulatory networks. Bioinformatics 21(13), 3058–3059 (2005)
5. Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden Markov model: Analysis and applications. Mach Learning 32, 41–62 (1998)
6. Frith, M., Li, M., Weng, Z.: Clusterbuster:finding dense clusters of motifs in dna seqs. Nuc. Ac. Res. 31(13), 3666–3668 (2003)
7. Frith, M.C., Hansen, U., Weng, Z.: Detection of cis-element clusters in higher eukaryotic DNA. Bioinf. 17, 878–889 (2001)
8. Gallo, S., Li, L., Hu, Z., Halfon, M.: Redfly:a regulatory element database for drosophila. Bioinf. 22(3), 381–383 (2006)
9. Gupta, M., Liu, J.S.: De novo cis-regulatory module elicitation for eukaryotic genomes. Proc. Natl. Acad. Sci. USA 102(20), 7079–7084 (2005)
10. Huang, H., Kao, M., Zhou, X., Liu, J.S., Wong, W.H.: Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. Journal of Computational Biology 11(1) (2004)
11. Liu, X., Brutlag, D.L., Liu, J.: Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Proc. of Pac. Symp. Biocomput., 127–138 (2001)
12. Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., Rubin, E.M.: rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. 12(5), 832–839 (2002)
13. Ludwig, M.Z., Patel, N.H., Kreitman, M.: Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 125(5), 949–958 (1998)
14. Maerkl, S.J., Quake, S.R.: A systems approach to measuring the binding energy landscapes of transcription factors. Science 315, 233–237 (2007)
15. Michelson, A.: Deciphering genetic regulatory codes:a challenge for fnal genomics. Pr. Nat. Acad. Sc. USA 99, 546–548 (2002)
16. Moses, A.M., Chiang, D.Y., Eisen, M.B.: Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. Pac. Symp. Biocomput., 324–335 (2004)
17. Murphy, K., Paskin, M.: Linear time inference in hierarchical hmms. Adv. in Neural Inf. Proc. Sys. 14 (2002)
18. Narang, V., Sung, W.K., Mittal, A.: Computational annotation of transcription factor binding sites in D melanogaster developmental genes. In: Proceedings of The 17th International Conference on Genome Informatics (2006)
19. Rajewsky, N., Vergassola, M., Gaul, U., Siggia, E.D.: Computational detection of genomic cis-regulatory modules, applied to body patterning in the early Drosophila embryo. BMC Bioinformatics 3(30), 1–13 (2002)
20. Rebeiz, M., Reeves, N.L., Posakony, J.W.: Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data site clustering over random expectation. Proc. Natl. Acad. Sci. USA 99(15), 9888–9893 (2002)

21. Sharan, R., Ovcharenko, I., Ben-Hur, A., Karp, R.M.: Creme: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics 19(Suppl 1), i283–291 (2003)
22. Siddharthan, R., Siggia, E.D., van Nimwegen, E.: Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. PLoS Computational Biology 1(7), e67 (2005)
23. Sinha, S., Blanchette, B., Tompa, M.: Phyme: A probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics 5(170) (2004)
24. Sinha, S., Liang, Y., Siggia, E.: Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic Acids Res. 34(Web Server issue), W555–W559 (2006)
25. Staden, R.: Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12(1 Pt 2), 505–519 (1984)
26. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., DeMoor, B., Rouze, P., Moreau, Y.: A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. Bioinformatics 17(12), 1113–1122 (2001)
27. Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S., Lawrence, T.E.: Decoding human regulatory circuits. Genome Res. 14(10A), 1967–1974 (2004)
28. Tompa, M., Li, N., Bailey, T., Church, G., DeMoor, B., Eskin, E., Favorov, A., Frith, M., Fu, Y., Kent, W., Makeev, V., Mironov, A., Noble, A., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z.: Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23(1), 137–144 (2005)
29. Wingender, E., Dietze, P., Karas, H., Knuppel, R.: TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic. Acids. Res. 24(1), 238–241 (1996)
30. Xing, E.P., Wu, W., Jordan, M.I., Karp, R.M.: Logos: A modular Bayesian model for de novo motif detection. Journal of Bioinformatics and Computational Biology 2(1), 127–154 (2004)
31. Zhou, Q., Wong, W.H.: Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc. Natl. Acad. Sci. USA 101(33), 12114–12119 (2004)