Sequence pattern discovery with applications to understanding gene regulation and vaccine design

M. Gupta and S. Ray

1 Introduction

Genome sequencing projects have led to a rapid growth of publicly available databases of genome sequences for DNA, RNA and proteins. These data consist of sequences of letters, from an alphabet of size $d$, without punctuations or space characters; $d$ equals 4 for DNA, with the alphabet as $\{A, C, G, T\}$, while $d$ equals 20 for proteins. The task of extracting biological insight from these sequence databases represents one of the greatest scientific challenges of the twenty-first century. The recent completion of the human genome (in April 2004) makes this question a more urgent task for the scientific community.

The key to understanding many types of biological phenomena from sequence data lies in detecting the underlying patterns. For example, certain proteins that regulate genes, recognize a location on the genome sequence they must bind to, through a sequence-specific pattern. Although these patterns (or motifs) may vary from sequence to sequence across species, functionality constraints often restrain the extent of their variation. Hence pattern discovery methods can be useful in detecting functional sequences. In immune response to infectious diseases, sequence pattern recognition plays an important role in controlling disease manifestation. It is necessary for the host cell to develop a mechanism for differentiating protein (peptide) sequences of the attacking viral genome from its own proteins, to destroy the attacker before it is itself destroyed. Pattern classification methods then provide a way to design specific vaccines to combat dangerous viral infections.

In this chapter, we focus on statistical methodology for sequence pattern discovery that has been developed in recent years, in the context of both problems mentioned above. We discuss a unifying probabilistic framework of hidden Markov models for sequences which has proven a very useful tool for designing pattern discovery algorithms and describe application-oriented extensions for the specific biological problems under consideration. Specifically, Sections 2 through 4 discuss pattern discovery in the context of studying gene regulation, while Sections 5 through 7 deal with applications to vaccine design.
Sequence pattern discovery

2 Pattern discovery in studying gene regulation

Determining the mechanisms of transcriptional regulation within a cell is essential to decipher regulatory pathways of human disease. Understanding interactions within the gene regulatory network can suggest possible targets for drug development.

Biology of gene regulation. Transcription regulation is controlled by the coordinated binding of one or more transcription factors (TFs) to specific locations in the promoter regions adjacent to genes, called transcription factor binding sites (TFBSs). For example, many bacterial promoters contain a “TATA-box” binding site, with a pattern of the form TATAAT- but slight deviations from this pattern may sometimes be tolerated (Figure 1). The challenge of the motif discovery problem is to simultaneously estimate the parameters of a model describing the position-specific nucleotide type preference for the TF (or TFs) and identify the locations of these binding sites, based only on a set of DNA sequences that are expected to be regulated together.

Laboratory assays such as electrophoretic mobility shift and nuclease protection have been developed to precisely locate TF-binding sites on a gene-by-gene and site-by-site basis. DNA footprinting is a technique used to identify the location of binding sites by carrying out limited hydrolyses of the DNA with or without the protein and comparing the products. Initially, the binding of a specific TF to a site is determined. The DNA fragment and the TF are then incubated together under favorable conditions. Next, the mixture is placed on a gel and an electric current applied, so that the molecules are sorted by size— the DNA fragments with the bound TF run more slowly than the bare fragments. The next step is to identify the specific binding site within a fragment. Again, the DNA and TF are incubated together, then the DNA is chemically degraded in a controlled manner by DNase I and enzymes. The degraded DNA is sorted on a gel by size— where fragments of all sizes are represented on the gel except those that were protected from degradation by the binding of the TF. The effectiveness of these approaches however have limitations, in terms of time and expenditure, as the amount of sequence to be analyzed increases. Computational methods that assume no prior knowledge of the pattern of the binding sites then become a necessary tool for aiding in their discovery.

Statistical approaches for motif discovery. Over the last decade, many statistical methods have been developed to infer regulatory motifs for sets of co-regulated genes (Hertz et al., 1990; Bailey and Elkan, 1994; Lawrence and Reilly, 1990; Liu et al., 1995, 2001; Bussemaker et al., 2000; Liu et al., 2002a; Gupta and Liu, 2003). There have also been efforts to improve motif prediction by correlating sequence patterns with mRNA expression data (Bussemaker et al., 2001; Conlon et al., 2003) or using comparative genomics information (Wasserman et al., 2000; McCue et al., 2001; Kellis et al., 2003). Although these methods have been quite effective in bacterial and other simple genomes, they have met with limited success in mammalian genomes. Main difficulties with TFBS prediction in complex genomes include the increased volume of the sequence search space, with TFBSs occurring a few kilobases away from the gene, instead of a few hundred on average in lower organisms; the increased occurrence of
low-complexity regions—single and di-nucleotide repeats; and overall shorter and less conserved TFBSs.

Most genes in complex organisms are controlled by a combination of factors with the corresponding binding sites forming spatial clusters (Davidson, 2001), termed cis-regulatory modules (CRMs). A number of statistical methods incorporating the CRM concept into motif discovery through a hidden Markov model (HMM) framework have recently been proposed (Thompson et al., 2004; Gupta and Liu, 2005). (For a review on these and other approaches that explicitly rely on sequence information for motif discovery, see Gupta and Liu (2006).)

In many cases of practical application, however, the motif signal is too weak to be captured by completely sequence-based approaches. In such cases, the availability of auxiliary information, such as from certain related biological assays or phylogenetic comparisons, greatly enhances the performance of motif discovery algorithms, if used in an appropriate way. In this chapter we focus on such a class of approaches that make use of auxiliary data to improve prediction of motif sites.

Chromatin ImmunoPrecipitation followed by microarray hybridization (ChIP-chip) of IP enriched DNA is a recently developed technology that has been successful in localizing transcription factor binding to a resolution of about a kilobase level (Buck and Lieb, 2004). Application of this technology with high-density oligonucleotide arrays allows the scanning of the whole genome at a high resolution for sequence regions bound by TFs. However, the resolution to which ChIP-chip can actually predict TF-bound regions is still at the kilobase level, which means further refinement is necessary to predict binding sites which are usually only 10-20 bases long. The availability of such ChIP-chip data provides an opportunity to generate more accurate predictions of binding sites; but at the same time, the massive dataset size and spatial dependence structure poses great challenges for data analysis. We will describe some recent approaches that address motif discovery augmenting sequence information with data from ChIP-chip experiments.

Cross-species comparisons provide another means to identify multiple genes that are likely to be regulated similarly; the main idea underlying this being that genes that code for the same protein in related species are likely to be similarly regulated. This approach can be employed on a genome-wide scale using only the sequences from a set of related species. As TF sites across species are likely to be more conserved than random background, comparative genomics can also be used to narrow down the sequence search space. One approach is to first align orthologous sequences (Schwartz et al., 2000) and exclude highly mutated sequences between species before motif finding. By using this approach, Wasserman et al. (2000) found that 98% (74/75) of binding sites of skeletal-muscle-specific transcription factors are present in the 19% of human sequences that are most conserved in the orthologous mouse sequence. Though in recent years, cross-species alignment followed by motif discovery has showed great promise (Kellis et al., 2003; Thompson et al., 2004), explicit statistical methods which incorporate full evolutionary models (Graur and Li, 2000) into motif discovery have not yet been explored in detail.
2.1 The motif discovery problem

The motif-finding problem can be cast as discovering non-trivial words from “nature’s dictionary” and their usage frequencies. However, even when all the words in the dictionary are known, estimating the word usage frequencies by exhaustive enumeration is infeasible. For example, consider the unsegmented sentence,

“of all the words in this unsegmented phrase there are some hidden”

It is easy for someone who knows English to pick out the most probable segmentation that breaks up this sequence into a meaningful set of patterns. But were a computer to do this, it would have to parse out the sentence into all possible combinations of patterns, and then choose the one that satisfies a certain criterion (here the underlined parts represent valid English words which are not the true ones used in this context). There are two more complications in analyzing biological sequences: (1) we do not know nature’s “dictionary”, and (2) instances of the same “word” occurring at different places may not be exact replicas of the same pattern.

2.1.1 A probabilistic model for motifs

In order to initiate a probabilistic model for motif discovery, Lawrence and Reilly (1990) proposed a “block-motif” model where the unknown motif is treated as a contiguous block within a longer segment (called the background). They assumed that each sequence under consideration contained only one occurrence of the motif. The background sequence (excluding a segment of length $w$ that corresponds to the motif) is assumed to be generated by i.i.d. draws from the alphabet $\{1, 2, \ldots, b\}$ ($b = 4$ for DNA) with the frequency vector $\theta_0 = (\theta_{01}, \ldots, \theta_{0b})^T$. Each of the $w$ positions of the motif are assumed to be generated from independent draws from a multinomial distribution with parameter $\theta_i = (\theta_{i1}, \ldots, \theta_{ib})^T$, ($i = 1, \ldots, w$). The motif frequency matrix $\Theta = (\theta_1 \cdots \theta_w)$ is unknown and the motif start positions on the $N$ segments, denoted by $a = (a_1, \ldots, a_N)$ are unobserved and must be inferred. The intuitive idea is to use a missing data formulation- treating $a$ as “missing data” and then iterate between imputing the $a$ and estimating $\Theta$ using either an EM algorithm (Dempster et al., 1977) or Gibbs sampling (Gelfand and Smith, 1990).

A more realistic model is the repetitive block-motif model, which can accommodate some sequences containing multiple copies of the motif and others not containing any (Figure 2). In this model, in addition to the site locations and frequency matrix, the total number of sites is also unknown. As a simple first step, it is assumed that at any position, there is a small unknown probability $\pi$ of a site to occur– this can be modified according to the available knowledge (e.g. physical properties of DNA composition, distance to the start of the gene, etc.). The other parameters $\theta_0$ and $\Theta$ remain unchanged from the single block-motif model.

2.2 Statistical approaches to motif discovery

To generalize the Lawrence and Reilly (1990) model directly to one in which the number of sites is unrestricted is computationally expensive, though theoretically straightforward. Suppose we have a set of $N$ sequences, denoted by $S = \{S_1, \ldots, S_N\}$, where sequence $S_i$ is of length $L_i$. In order to model multiple motifs per sequence, Bailey
M. Gupta and S. Ray

and Elkan (1994) present an approximation to the “segmentation model” through a simplified model in which each sequence in the data set is broken up conceptually into all overlapping subsequences of length \( w \). Let this new data set be denoted by \( \mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) \), where \( n = \prod_{i=1}^{w} (L_i - w + 1) \). Each observation \( \mathbf{X}_i = (x_{i1}, \ldots, x_{iw}) \), \( (i = 1, \ldots, n) \) is assumed to be generated under a finite mixture model indexed by an unobserved group indicator \( Z_{ij} \), where

\[
Z_{ij} = \begin{cases} 
1 & \text{if } X_i \text{ is generated from group } j, \\
0 & \text{otherwise.} 
\end{cases}
\]

For illustration, consider a model with only one distinct motif type \( (j = 1) \) and the background denoted by \( (j = 0) \). Also, let us denote the set of parameters corresponding to the motif component and background as \( \Theta_1 = (\theta_1, \ldots, \theta_w) \) and \( \Theta_0 = (\theta_0, \ldots, \theta_0) \), where \( \theta_i = (\theta_{i1}, \ldots, \theta_{iw})^T \) (for \( i = 1, \ldots, w \)), while \( \pi_j = P(Z_{ij} = 1) \) denotes the relative proportion of motif segments (mixing proportion). Given the class indicator \( Z_{ij} = 1 \), \( X_i \) is assumed to be generated from a product multinomial model characterized by \( \Theta_j \). Let \( \Theta = (\Theta_0, \Theta_1, \pi) \). Let \( Z_{i} = (Z_{i1}, Z_{i2}) \). The complete data likelihood can then be written as:

\[
P(\mathbf{X}, \mathbf{Z} | \Theta) = \prod_{i=1}^{n} P(\mathbf{X}_i, \mathbf{Z}_i | \Theta) = \prod_{i=1}^{n} \prod_{j=0}^{1} [p(\mathbf{X}_i | \Theta_j) \pi_j]^{Z_{ij}}.
\]

Under this set-up, considering the group membership indicators \( Z_{ij} \) as missing data, it is now possible to set up a standard EM algorithm to maximize the likelihood \( P(\mathbf{X} | \Theta) \) with respect to \( \Theta \).

(i) **E-Step:** Let us denote the complete data loglikelihood as \( l(\Theta | \mathbf{X}, \mathbf{Z}) = \log P(\mathbf{X}, \mathbf{Z} | \Theta) \). Then, we have

\[
E[l(\Theta | \mathbf{X}, \mathbf{Z})] = \sum_{i=1}^{n} \sum_{j=0}^{1} Z_{ij}^{(0)} \log p(\mathbf{X}_i | \Theta_j) + \sum_{i=1}^{n} \sum_{j=0}^{1} Z_{ij}^{(0)} \pi_j, \quad (1)
\]

where

\[
Z_{ij}^{(0)} = \frac{p(\mathbf{X}_i | \Theta_j) \pi_j^{(0)}}{\sum_{k=0}^{1} p(\mathbf{X}_i | \Theta_k) \pi_k^{(0)}},
\]

and \( p(\mathbf{X}_i | \Theta_j) \) is the probability of \( X_i \) under the product multinomial distribution, i.e.,

\[
p(\mathbf{X}_i | \Theta_1) \propto \prod_{l=1}^{w} \prod_{j=1}^{4} \theta_l^{(X_{il}=j)}, \quad \text{and} \quad p(\mathbf{X}_i | \Theta_0) \propto \prod_{j=1}^{4} \theta_0^{\sum_{l=1}^{w} l[X_{il}=j]}.
\]
(ii) **M-Step**: The M-step maximizes Eq.(1) over $\Theta$ and $\pi$ to find the new estimates $\Theta^{(t+1)}$ and $\pi^{(t+1)}$ at the $(t+1)^{th}$ step, given by,

$$
\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} Z_{ij}^{(t)}}{n}, \quad j = 0, 1,
$$

$$
\theta_{lk}^{(t+1)} = \frac{c_{lk}^{(t)}}{\sum_{k=1}^{b} c_{lk}^{(t)}}, \quad l = 1, \ldots, w; \ k = 1, \ldots, b.
$$

$$
\theta_{0k}^{(t+1)} = \frac{c_{0k}^{(t)}}{\sum_{k=1}^{b} c_{0k}^{(t)}}, \quad k = 1, \ldots, b.
$$

Now let $\delta_{[A]}$ denote the indicator function taking value 1 if condition $A$ is true. Then,

$$
c_{lk}^{(t)} = \sum_{i=1}^{n} Z_{i1}^{(t)} \delta_{[X_{il}=k]}, \quad (l = 1, \ldots, w; \ k = 1, \ldots, b),
$$

which is the expected number of times letter $k$ appears in position $l$ of the motif. Similarly, under the background model,

$$
c_{0k}^{(t)} = \sum_{i=1}^{n} \sum_{l=1}^{w} Z_{i0}^{(t)} \delta_{[X_{il}=k]}.
$$

In order to avoid boundary problems when some letter frequency $\hat{\theta}_{ij}$ becomes 0 (and thus fixed), a small pseudo-count $\beta_{j}$, $(j = 1, \ldots, b)$, is added to each frequency $c_{ij}$ in the M-step.

One problem with this multiple-site approach is that it does not model the sequences exactly, overlapping subsequences being treated as independent. Thus certain restrictions need to be imposed to ensure that two overlapping subsequences are not both predicted to be generated from the same motif. Also, since the EM algorithm is susceptible to getting trapped in local modes, it is often difficult to implement this algorithm directly for a larger number of motif types, even though conceptually straightforward. A more general segmentation-based HMM approach is outlined in Section 3.

### 2.2.1 A Bayesian approach for multiple motif sites

Based on the missing data formulation as discussed in the previous section, Liu et al. (1995) developed an MCMC-based algorithm for motif discovery, referred to as the Gibbs Motif Sampler (GMS). The Bayesian framework makes it easier to generalize to multiple motif sites per sequence and multiple motif types. Again, assume a set of $N$ sequences, denoted by $S = \{S_1, \ldots, S_N\}$ where sequence $S_i$ is of length $L_i$. Also, let us denote a set of missing motif position indicators as $A = ((A_{ik}))$, $(i = 1, \ldots, N; \ k = 1, \ldots, L_i)$, where

$$
A_{ik} = \begin{cases} 
1 & \text{if } k \text{ is the start of a site in sequence } i, \\
0 & \text{otherwise}.
\end{cases}
$$
If $A$ is known, we can write the vector of letter counts in the background sequence as $c_0 = (c_{01}, \ldots, c_{0b})$, and the letter counts in the motif as a matrix $(c_{ij})$; $(l = 1, \ldots, w, j = 1, \ldots, b)$. Finally, let $\pi (0 < \pi < 1)$ denote the probability of motif occurrence at any point in the sequence. The complete data likelihood can then be written as

$$P(S, A|\theta_0, \Theta) \propto \prod_{j=1}^{b} \theta_{0j}^{c_{0j}} \prod_{l=1}^{w} \prod_{j=1}^{b} \left( \frac{\theta_{lj}}{\theta_{0j}} \right)^{c_{lj}} \pi^{M} (1 - \pi)^{\sum_{i=1}^{L} L_i - Mw},$$

where $M = \sum_{j=1}^{b} c_{lj}$ denotes the total number of motif sites. Next, assume a Dirichlet($\beta_0$) prior distribution for $\theta_0$ and a product of Dirichlet distributions with hyperparameters $\beta_l = (\beta_{l1}, \ldots, \beta_{lb})$; $(l = 1, \ldots, w)$ for $\Theta = (\theta_1, \ldots, \theta_w)$. A simple Gibbs sampling algorithm can then be constructed by initializing a random set of motif positions $A^{(0)}$, and then alternately updating $(\theta_0, \Theta)$ and $A$ by iterative sampling from their posterior conditional distributions. In the case of sampling of $A$, this would mean updating its components $A_{ik}$ one at a time while keeping the complementary set $A^c_{ik}$ fixed.

A more efficient version of the above algorithm was obtained using the idea of collapsing (Liu, 1994). For the moment, for simplicity, assume that $\theta_0$ is known. Integrating out $\Theta$ from the joint posterior distribution, we can instead sample from the new predictive conditional distribution of $A$, obtained as:

$$P(A|S) \propto P(S, A) = \int P(S, A|\theta_0, \Theta) p(\Theta) d\Theta$$

$$\propto \prod_{j=1}^{b} \Gamma \left( c_{0j} - \sum_{l=1}^{w} c_{lj} + \beta_{0j} \right) \prod_{l=1}^{w} \prod_{j=1}^{b} \Gamma(c_{lj} + \beta_{lj}) \pi^{M} (1 - \pi)^{\sum_{i=1}^{L} L_i - Mw}.$$

Using Stirling’s approximation for gamma functions, we can then arrive at a simple approximate formula for the predictive distribution, which can then be used to sample for the existence of a site at position $k$ in sequence $i$:

$$P(A_{ik} = 1 | S, A^c_{ik}) \propto \frac{\pi}{1 - \pi} \prod_{l=1}^{w} \prod_{j=1}^{b} \left( \frac{\hat{\theta}_{lj}}{\hat{\theta}_{0j}} \right)^{I(S_i, k+i-1 = j)}, \quad (2)$$

where $\hat{\theta}_{0j} = \frac{c_{0j} + \beta_{0j}}{\sum_{l=1}^{w} (c_{lj} + \beta_{lj})}$ and $\hat{\theta}_{lj} = \frac{c_{lj} + \beta_{lj}}{\sum_{l=1}^{w} (c_{lj} + \beta_{lj})}$ being the posterior estimates. Since the above probability is dependent on $\pi$, we can further use an appropriate hyperprior on $\pi$ to marginalize it out of the posterior distribution.

3 Hidden Markov models for sequence analysis

To extend the preliminary models of motif-containing DNA sequences for more complex applications, it is helpful to introduce the concept of a general mathematical abstraction, the hidden Markov model. For the moment, consider a single sequence of length $N$, with positions indexed by $i$, $(i = 1, \ldots, N)$. An HMM is a doubly stochastic probability model involving two sets of variables: the “hidden” set, say
Sequence pattern discovery

Let $h_i, i = 1, 2, \ldots, N$, be a discrete random variable taking values in a discrete space $H = \{1, \ldots, K\}$, following a first-order Markov chain; and the “observed” set $y_i, (i = 1, 2, \ldots, N)$ which are conditionally independent, given the $h_i, (i = 1, 2, \ldots, N)$. For convenience, let us assume $h_0 = 0$, a fixed state. If the “hidden” states were known, and with $\theta$ denoting an arbitrary set of parameters, the complete data likelihood could be written as:

$$
L(h, y|\theta) = \prod_{i=1}^{N} p_{y_i|h_i}(y_i | h_i; \theta)p_{h_i}(h_i | h_{i-1}; \theta).
$$

(3)

HMMs are widely used as a method of modeling discrete valued series data (MacDonald and Zucchini, 1997), including many sequence alignment and sequence segmentation models in current practice.

The HMM framework falls within the broad category of “missing data” problems in which parameter estimation can be handled through an EM algorithm or data augmentation approach (Tanner and Wong, 1987). In order to demonstrate the EM approach, we first define two entities, (i) the forward probability $f_i(\cdot)$, where

$$
f_i(k) = P(y_1, \ldots, y_i, h_i = k) = P(y_i|h_i = k) \sum_{l=1}^{K} f_{i-1}(l)P(h_i = k|h_{i-1} = l),
$$

and (ii) the backward probability $b_i(\cdot)$, defined as:

$$
b_i(k) = P(y_{i+1}, \ldots, y_n|h_i = k) = \sum_{h_{i+1}} P(y_{i+1}|h_{i+1}) b_{i+1}(h_{i+1}) P(h_{i+1}|h_i).
$$

For now, assume that $y$ is a discrete random variable taking values in the set $B = \{b_1, \ldots, b_m\}$ (e.g. for DNA, $m = 4$), with $P(y_i = b|h_i = k) = \mu_{kb}$ (emission probability) and $P(H_{i+1} = k|H_i = j) = \tau_{jk}$ (transition probability). To find the MLEs of $\theta = (\mu, \tau)$, we need to maximize $\log P(y|\theta) = \sum_{\pi} \log P(y, \pi|\theta)$ where $\pi$ denotes the set of all $K^N$ possible realizations of $\pi = (h_1, \ldots, h_N)$. Although we cannot get closed forms for the MLEs, we can make use of the EM algorithm (also known as the Baum-Welch algorithm in this context) through the following steps:

- **E-step.** Let

$$
Q(\theta|\theta^t) = \sum_{\pi} P(\pi|y, \theta^t) \log P(y, \pi|\theta).
$$

Then,

$$
P(y, \pi|\theta) = \prod_{k \in H} \prod_{b \in B} [\mu_{kb}] E_{kb}(\pi) \prod_{k} \prod_{l} T_{kl}(\pi),
$$

where $E_{kb}(\pi) = \sum_{i=1}^{N} \delta_{y_i=b, h_i=k}(\pi)$ and $T_{kl}(\pi) = \sum_{i=2}^{N} \delta_{h_{i-1}=k, h_i=l}(\pi)$ are the emission and transition frequencies corresponding to path $\pi$ ($\delta_{A}$ denotes
the indicator function which takes the value 1 if condition $A$ is true). So,

$$Q(\theta | \theta^t) = \sum_{\pi} P(\pi | y, \theta^t) \times \left[ \sum_k \sum_b E_{kb}(\pi) \log \mu_{kb} + \sum_k \sum_l T_{kl}(\pi) \log \tau_{kl} \right]$$

$$= \sum_k \sum_b \log \mu_{kb} \left[ \sum_{\pi} P(\pi | y, \theta^t) E_{kb}(\pi) \right] + \sum_k \sum_l \log(\tau_{kl}) \left[ \sum_{\pi} P(\pi | y, \theta^t) T_{kl}(\pi) \right]$$

where

$$T_{kl} = \sum_{i} P(H_i = k, H_{i+1} = l | y) = \sum_{i} f_i(k) \tau_{kl} \mu_{ly_i+b_i+1} \rho(y)$$

and

$$E_{kb} = \sum_{\{i: y_i = b\}} f_i(k) b_i(k) \rho(y).$$

- **M-step.** Using the expressions defined in the E-step, we get the parameter estimates as:

$$\hat{\mu}_{kb} = \frac{\hat{E}_{kb}}{\sum_b \hat{E}_{kb}}, \quad \hat{\tau}_{kl} = \frac{\hat{T}_{kl}}{\sum_l \hat{T}_{kl}}.$$
Sequence pattern discovery

the likelihood for the partial sequence $S_{[1:k]}$. Then

$$
L_k(\rho) = \sum_{j=1}^{D} P(S_{[k-w_j+1:k]} \mid \rho) L_{k-w_j}(\rho),
$$

(4)

where $w_j (j = 1, \ldots, D)$ denote the word lengths.

Estimating $\rho$ from this model is conceptually simple. One can directly optimize Eq.(4) via a Newton-type algorithm (Bussemaker et al., 2000). Alternatively, one can employ an EM algorithm or a Gibbs sampler, which may possibly be slower.

Based on the above idea, Gupta and Liu (2003) extended the approach to the case of non-exact words. The concept of a stochastic dictionary was introduced, consisting of a collection of "stochastic words" represented by the probabilistic weight matrices (PWM) $\Theta$, instead of the fixed words $M$. Each column of a PWM gives the probabilities of finding each letter in that position of the corresponding stochastic word. The motif-finding problem then reduces to finding the form of the stochastic word matrix (or multiple word matrices) and the likely locations of the stochastically varying words in the sequence, which is addressed through a forward-backward recursion-based data augmentation algorithm. In comparison to the site updating step Eq.(2) in the Gibbs sampler which samples from the full conditional distribution, the data augmentation step under the stochastic dictionary framework sequentially samples $A$ from its joint posterior distribution:

$$
P(A \mid \Theta, S) = P(A_L \mid \Theta, S) \prod_{j=1}^{L-1} P(A_j \mid A_{j+1}, \ldots, A_L, S, \Theta)
$$

(5)

3.2 Dependence structures in the motif model

By addressing the motif discovery problem through the HMM framework, a variety of complex structures that arise in biological applications can be addressed efficiently, for example, gapped motifs and correlated motif patterns (modules). Gapped motifs occur when not all contiguous columns of the motif are conserved across instances, which may be the case when a few but not all positions of the motif are important for transcription factor binding.

To generalize the stochastic dictionary HMM to include gapped words, let us first consider a model with only one type of ungapped motif of width $w$. An observed set of sequences containing motifs at unknown positions can then be considered to have a "hidden" indicator at every position that takes values 1 or 0 according to whether it is the start of a motif site or not. We may now view this model as an HMM by expanding state 1, which corresponds to a motif of length $w$, into a series of $w$ consecutive states $\{1, \ldots, w\}$. Hence the stochastic dictionary model with one stochastic word has an underlying state distribution characterized by the first-order Markov transition matrix (with states represented in the order $0, 1, 2, \ldots, w)$:

$$
\begin{pmatrix}
1 - \pi & \pi & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 0 & 1 & \cdots & 0 \\
\vdots & & & & & \ddots \\
1 - \pi & \pi & 0 & 0 & \cdots & 0
\end{pmatrix}
$$
M. Gupta and S. Ray

where $\pi$ denotes the probability of starting a new motif site, and $b$ the alphabet size).

We can further generalize this idea to gapped motifs. Assume a motif of minimum width $w$ with possible insertions after every position and probability of insertion $\lambda$. We can write the expanded transition matrix with $w + (w - 1) + 1$ states (the first denotes the background model) \( \{0, 1, I_1, 2, I_2, \ldots, I_{w-1}, w\} \), where the state names starting with “I” denote the insertions:

\[
\begin{pmatrix}
1 - \pi & \pi & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & \lambda & 1 - \lambda & 0 & 0 & \cdots & 0 \\
0 & 0 & \lambda & 1 - \lambda & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & \lambda & 1 - \lambda & \cdots & 0 \\
0 & 0 & 0 & 0 & \lambda & 1 - \lambda & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
1 - \pi & \pi & 0 & 0 & 0 & 0 & \cdots & 0
\end{pmatrix}
\]

The gapped motif discovery problem can be efficiently addressed by a two-stage data augmentation procedure, again based on the HMM structure (Gupta and Liu, 2003).

Regulatory modules consist of two or more sites forming a “cluster”, which are both necessary for binding but are individually often too weakly conserved to be picked up by motif discovery algorithms. In both these cases, standard motif models are not applicable and modifications must be made. The regulatory module discovery problem can also be addressed through an HMM, with successive motif sites on a sequence being modeled through a Markov process. The posterior distribution Eq.(5) for the DA procedure must then be adjusted to account for the dependence in $A$ (Gupta and Liu, 2005; Thompson et al., 2004).

4 Using auxiliary data in motif prediction

Until now, we have described motif discovery techniques based solely on sequence data. However, when dealing with large and complex genomes, such an approach is often insufficient in discovering motif patterns or it may result in false positive predictions (often due to the frequent occurrence of low complexity repeat sequences, such as ATATATATATATATAT). Fortunately, in certain cases, auxiliary information is available, which can aid motif prediction. An important question then is, how to best incorporate the additional information revealed by, say, cross-species comparisons or data from chromatin-immunoprecipitation experiments. A natural route is to build an appropriate statistical model to reflect this information and to construct a search algorithm accordingly. For example, if a sequence segment is located in a region where a cross-species comparison shows high conservation, then it is highly likely that the segment corresponds to a protein binding site. Otherwise, such a prior probability would be small. In the following sections, we describe some of our and others’ initial attempts to incorporate auxiliary data in motif discovery.

4.1 Using tiling array data in motif prediction

Recently, a new protocol called Chromatin-ImmunoPrecipitation followed by microarray hybridization (ChiP-chip) has been developed to locate sites of protein-DNA interaction in living cells (Buck and Lieb, 2004). In these experiments, DNA is cross-linked
Sequence pattern discovery

to proteins at sites of DNA-protein interactions, and sheared to 1-2kb fragments that
tile over the entire genome. The DNA-protein complexes are then precipitated by
antibodies specific to the protein of interest. The precipitated protein-bound DNA
fragments are amplified, labeled fluorescently, and hybridized to whole-genome mi-
croarrays (e.g. as probes of approximately 25-35 bp in Affymetrix arrays). DNA
fragments that are consistently enriched by the ChIP-array over repeated experi-
ments are candidates for containing the protein-DNA interacting loci. However, the
probe measurement cannot accurately indicate the location of the bound TF on the
genome as the probe can lie in a neighborhood of up to 500-600 bp of the TF site
and still show high comparative hybridization values. Hence further motif discovery
procedures must be used to exactly pinpoint the binding targets.

4.1.1 Preliminary approaches

By the nature of ChIP-chip experiments, it seems promising to direct the motif discov-
ery algorithm to search the sequences with highest comparative hybridization values
more thoroughly before searching the less likely sequences. MDscan (Liu et al., 2002b)
incorporates these features, by first ranking sequences according to scores derived from
ChIP-chip experiments, and then searching for novel motifs using an enumerative pro-
cedure. This approach has shown relative success in discovering sites in moderately
complex genomes such as yeast. With the recent development of high resolution tiling
arrays for complex genomes (including the human genome) it seems to be promising
to take the spatial nature of the array into consideration, for more accurate motif
discovery. Keles et al. (2004) propose multiple testing procedures using a scan statist-
ic over the tiling array data. For each tiled probe, a two-sample Welch t-statistic
is calculated and a “scan structure” is imposed by averaging the t-statistic over the
neighboring $w$ probes, where $w$ ($\sim 30$) is chosen to reflect the average number of
probe pairs that are likely to be covered by the average length of enriched fragments
($\sim 1$ kb) in the immunoprecipitated solution. The main motivation behind using a
“scan” structure is to make rejection of the null hypothesis (not bound by the TF)
more difficult if a probe is in the vicinity of bound probes.

The spatially dependent structure of the tiling array naturally suggests the use of
an HMM to model the binding propensities of the underlying sequence. The regions
that have a high probability of having been bound by the TF can be selected for
applying further motif discovery techniques. Below we discuss a preliminary approach
to determine TF binding locations by using the sequence and ChIP measurement
information in a HMM-based framework.

4.1.2 An HMM for ChIP-chip data

For simplicity of notation, here we assume the data are from a single contiguous
sequence. If there are gaps (due to parts of the sequence being left out of the tiled array
for ambiguous or faulty hybridization characteristics), the sequence parts separated
by them should be treated as independent segments. Let us denote the ChIP score
vector corresponding to the $N$ probe regions in the segment as $Y = (Y_1, \ldots, Y_N)$. Each $Y_i = (Y_{i1}, \ldots, Y_{ir})$ corresponds to a segment of approximately 25-35 bp in
length, the $Y_{ij}$ denoting the score from the $j^{th}$ replicate ($j = 1, \ldots, r$).

Let us now introduce a set of (hidden) indicator variables $H = (H_1, \ldots, H_N)$, where $H_i = 1$ (0) if the segment $i$ is bound (not bound) by the TF of interest. Now, under $H_i = k$, let the ChIP scores (log-ratios) be distributed as $Y_{ij} \sim g_k[\cdot; \mu_k, \sigma_k^2]$. Let $\tau = (\tau_{kl})$ $(k, l = 0, 1)$ denote the transition probability matrix for the hidden states $H$ and denote the set of all parameters as $\phi = (\mu, \sigma, \tau)$. Our first object is to find which observations are most likely to have been generated from the TF-bound state, i.e. $P(H_i = 1|Y, \phi)$, and estimate the unknown parameters in $\phi$. This is done using an EM algorithm along the lines of the procedure discussed in Section 3. The main difference in the implementation arises due to the fact that the observations $Y$ are now continuous, so the E- and M-steps are modified as follows.

- **E-step.** Let $\delta_{[A]}$ again denote the indicator function, and let us assume that for the initial observation, $H_1 = 0$. Then,

$$Q(\theta|\theta^t) = E_{H|Y, \phi} \left[ \log P(Y, H|\phi) \right]$$

$$= E \left[ \sum_{i=1}^{N} \delta_{[H_i=k]} \log g_k(Y_{i}|\phi) + \sum_{i=2}^{N} \delta_{[H_{i-1}=k, H_i=l]} \log \tau_{H_{i-1}H_i} \right]$$

$$= \sum_{k \in K} \sum_{i=1}^{N} E_{ik} \log g_k(Y_{i}|\phi) + \sum_{k \in K} \sum_{l \in K} \log \tau_{kl} \sum_{i=2}^{N} T_{kl}, \quad (6)$$

where $T_{kl} = \sum_{i=2}^{N} P(H_{i-1} = k, H_i = l|Y, \phi)$ and is calculated in similar fashion as in Section 3, while $E_{ik} = P(H_i = k|Y, \phi) = \frac{f_i(k)b_i(k)}{P(Y|\phi)}$. The forward and backward probabilities are as follows:

$$f_i(k) = P(Y_1, \ldots, Y_i, H_i = k|\phi) = g_k(Y_{i}|\phi) \sum_{l \in K} f_{i-1}(l) \tau_{ik},$$

and

$$b_i(k) = P(Y_{i+1}, \ldots, Y_N|H_i = k, \phi) = \sum_{l \in K} g_l(Y_{i+1}|\phi) b_{i+1}(l) \tau_{lk}.$$  

- **M-step.** For the time being, let us assume the functional forms for the densities of $Y_i$ are Gaussian, i.e. $g_k(\cdot) = N(\mu_k, \sigma_k^2)$, though slight alterations may be necessary if other distributions are used. The parameter estimating equations for $\tau$ remain unchanged, but the new emission parameters are estimated by:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{r} E_{ik} Y_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{r} E_{ik}}$$

$$\text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{r} E_{ik} (Y_{ij} - \hat{\mu}_k)^2}{\sum_{i=1}^{N} \sum_{j=1}^{r} E_{ik}}. \quad (7)$$

As a byproduct of the EM algorithm, we obtain the posterior state probabilities of interest, e.g. the posterior probability of probe $i$ coming from a TF-bound segment is:

$$P(H_i = 1|Y, \phi) = \frac{f_i(1)b_i(1)}{P(Y|\phi)}. \quad (8)$$
Sequence pattern discovery

Now for each $Y_i$, consider a window of width $W$, corresponding to the DNA sequence the probe was generated from. A high probability of being bound should be associated with the presence of a motif site. Hence the posterior probabilities in Eq. (8) can be used to rank sequence windows and then a motif discovery algorithm such as the Gibbs sampler or MDscan can be used to detect the motif sites.

4.1.3 Example

We applied the above method on a subset of a tiling array data set for the transcription factor PU.1. The selected region consisted of about 50,000 data points from Encode region 1 on chromosome 7 (The ENCODE Project Consortium, 2004). ChIP array measurements were taken for time points 0, 2, 8, and 32 hours, with each treatment and control value having 3-5 biological replicates. The correlation between ChIP score ranks and ranks of probes with a greater than 99.99% chance of being bound (predicted by the HMM) is quite high (Figure 3) though there is some variation that is captured by the spatial correlation. We compared the motifs predicted by (i) MDscan based only on the ChIP score rankings and (ii) highly bound regions of the HMM-based method, to the experimentally verified PU.1 binding motif (Heinemeyer et al., 1998). It seems apparent that the motif predicted by the HMM is a closer match to the known consensus, up to a single letter “shift”), especially in the more conserved positions which are likely to be functionally important. This indicates that the HMM-based method may be a more promising alternative in detecting binding sites on tiled arrays.

This model, which takes into account the tiling array dependence structure, can be further improved by more explicit modeling of the nature of the probes, as probe binding specificities are known to vary substantially. However, this is likely to lead to estimability problems due to a huge increase in the parameter space. Li et al. (2005) use an HMM with probe-specific emission probabilities; however, instead of estimating the parameters, they are fixed a priori at arbitrary values. One way to account for different binding specificities through a full statistical model is through the introducing of a sequence-specific covariate $X_i$, on which the transition probabilities $\tau$ will now depend. The model will then represent a particular case of a non-homogeneous HMM, in which estimation may be again carried out through an EM procedure; however, numerical optimization techniques may be necessary at the M-step as a closed form solution for the estimating equations may not be available. There are other directions in which the model can be improved- such as (a) using a more robust distribution rather than the normal to accommodate outliers and (b) motif parameter updating directly within the HMM framework.

4.2 Classification of motifs using cross-species comparisons

With a large number of genomes being completely sequenced, data is often available from a number of closely related species as well as distantly related ones. Since even distantly related species may have similar binding sites due to functional constraints, applying a motif discovery algorithm solely on highly aligned parts between two related genomes often succeed in finding more of the true sites. Using phylogenetic information has proved valuable in identifying novel transcription factor binding sites.
M. Gupta and S. Ray

both in lower organisms (McCue et al., 2001) and more complex genomes (Water-

man et al., 2000). Motif discovery algorithms typically yield a large number of motif

patterns among which a high proportion are likely to be false positives. While it is

not possible to say with certainty that a found pattern is “false” (it may be a previ-

ously undetected regulatory site), it is desirable to be able to judge which patterns

are more likely to be “true” as compared to others. The practical usefulness of includ-

ing phylogenetic information in motif discovery provides a motivation for applying it

within the statistical framework for “classifying” motifs to differentiate possible false

positive predictions.

4.2.1 Evolutionary classification of motifs

The usual evolutionary model assumes that, ignoring insertions and deletions, the

evolution of every position in a “neutral” (i.e., non-functional) region of the genome

follows an independent continuous-time Markov process with a matrix $Q$ as its in-

finitesimal generator. By the Kolmogorov equation, we can obtain the transition

probability matrix $P(t) = \{p_{ij}(t)\}_{4 \times 4}$ as

$$P(t) = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}.$$  \hspace{1cm} (9)

A well-known special case of this model is the Jukes-Cantor model (Jukes and Can-

tor, 1969), where all the parameters are equal. An improved two-parameter model is
given in Kimura (1980), where the transition rates ($\alpha$) and transversion rates ($\beta$) are
unequal and the infinitesimal rate matrix can be written as,

$$Q = \begin{pmatrix}
-\alpha - 2\beta & \beta & \alpha & \beta \\
\beta & -\alpha - 2\beta & \beta & \alpha \\
\alpha & \beta & -\alpha - 2\beta & \beta \\
\beta & \alpha & \beta & -\alpha - 2\beta \\
\end{pmatrix}.$$  \hspace{1cm} (10)

This means that, for example, after a small time interval $\Delta t$, there is a probability
$\beta \Delta t$ to mutate from $A$ to $C$. Under different assumptions for transition and transversion
rates, this model can be generalized further into a 6-parameter (GTR), or full 12-

parameter model. Though it may seem that using a more general form of the model
would lead to greater accuracy of predictions, this is often not the case, as for models
with a number of parameters greater than 2, we may need to assume that the common
ancestral sequence was at equilibrium in terms of nucleotide frequencies, which may
not be a feasible assumption. The Kimura (and Jukes-Cantor) model has the property
of (a) symmetric transition matrices $P$ and (b) a uniform distribution of nucleotide
frequencies at equilibrium.

Consider a short segment of DNA, $X = (x_1, \ldots, x_w)$, where every position is
subject to a point mutation pressure characterized by $Q$. Let $x_i$ be a column vector
of length 4 denoting a particular base type (i.e., “A”=(1,0,0,0)$^T$, “C”=(0,1,0,0)$^T$,
etc.). For segment $X$ to be a regulatory binding site, there must be some energetic
constraint required by the protein-DNA interaction. The simplest energy model is a
linear additive function of the form

$$U(X) = -\sum_{i=1}^{w} x_i^T v_i,$$
Sequence pattern discovery

where \( \mathbf{v}_i = (v_{i1}, \ldots, v_{i4})^T \) with the \( v_{ij} \) denoting the contribution to the overall energy if position \( i \) of the motif has base type \( j \). It is then reasonable to think that from one generation to another, the total binding energy should be “maintained” so as to make the organism viable. This type of functional conservation can be captured by using the Metropolis rejection (or, more precisely, fitness) rule (Metropolis et al., 1953) to maintain the organism’s viability. In this Metropolis-chain model, an “offspring” \( Y \) of the current segment \( X^{(t)} \) survives with probability

\[
\min \left\{ 1, \exp\left\{ U(X^{(t)}) - U(Y) \right\} \right\}.
\]

Otherwise, the mutant is replaced by the parental segment \( X^{(t)} \). That is, \( X^{(t+1)} \) is equal to either \( Y \) or \( X^{(t)} \), depending on the energetic characteristics of \( Y \).

Here we do not literally take the proposing-rejecting interpretation of the Metropolis-chain model for an individual in the population, but rather treat it as a modified Markov process for evolution of the motif. There are also other ways of enforcing energetic constraints. A biological fact that may be modeled is that the binding between the transcription factor and its binding site can be neither too tight nor too loose. This might be reflected by a survival probability of the mutant \( Y \) as

\[
\min \left\{ 1, \frac{\exp\{-U(Y)\}}{C_1}, \frac{C_2}{\exp\{-U(Y)\}} \right\},
\]

where \( \log(C_1) < \log(C_2) \) can be seen as an energy “band.”

Using this framework, we can also investigate the evolution of motif elements with more complex energy functions, e.g., those involving interactions among the positions within the site. Specifically, we simulate two Markov processes using the Kimura 2-parameter model, with the fitness rule Eq.(12), for the motif segments that are of interest.

### 4.2.2 An algorithm for differentiating potential false positives.

Assume that we have a set of confirmed (experimentally detected or otherwise) binding sites for sequences of a certain species, say A, and a set of potential binding sites in another closely related species, say B, detected computationally by one of the previously described motif-finding algorithms. Under the rules Eq.(11) and Eq.(12), we simulated the evolutionary process for each motif pattern in set B with energetic constraints specified by the sets of motifs in A. The equilibrium distribution attained by each of the starting patterns in set B represented the possible variation that may have occurred in the original motif during the evolutionary time difference between the species A and B.

For classification, we now adopt the following procedure. Using the weight matrix corresponding to each set of mouse motifs, we start with each of the human consensus segments in turn and run it through an evolutionary chain of length approximately equal to the time of divergence between the two species, based on the Kimura two-parameter model. (The term consensus refers to the segment of highest probability that can be generated from the weight matrix). The energy bounds are constructed from the lowest and highest energies for the species A sites. The evolutionary algorithm may be roughly described as follows:
(i) Start with a segment $X^{(0)} = (x_1, \ldots, x_w)$ corresponding to a chosen motif consensus for species B.

(ii) At time $t$, create a “potential” segment $Y$ it can mutate into, based on the transition matrix based on the Kimura 2-parameter model.

(iii) Calculate “binding energy” $U(Y)$ for the new segment, where $U(X) = -\sum_{i=1}^{w} x_i^T v_i$, and $v_{ij} = \log(\theta_{ij})$, $\theta_{ij}$ denotes the normalized weight matrix for the set of motifs. Select the new segment to be $Y$ with probability given as in Eq. (12), where $\log(C_1) < \log(C_2)$ represent the minimum and maximum energies for the consensus sites in the species A motifs.

(iv) Continue the process for $T$ generations, where $T$ is the approximate time of divergence between species A and B.

After this simulation, the resultant set of motifs (recorded for iterations corresponding to the time of species divergence) represent the equilibrium distribution that is achieved under this evolutionary rule and the specified binding constraints. If this equilibrium distribution accurately matches the consensus of any of the species A sites, we may speculate that such variation in the motif of species B may have occurred through an evolutionary procedure as simulated above.

4.2.3 Results

The method described above was used on a set of paired DNA sequences from human and mouse skeletal muscle (Wasserman et al., 2000). Motif sites of width 8 that have been detected experimentally in mouse sequences are also available (Table 1). The consensus sites found by Gibbs sampling in the human sequences do not match exactly with any of the true motif patterns in the mouse; the closest match being between the first motif in the mouse sequences (TAAAAATA/TATTTTTA) (second string is the reverse complement), to the human motif (TAAGAATA/TATTCTTA) but still differing in one position. Given that TFBSs need to be highly conserved to retain their functional capabilities, we now check whether the evolutionary process itself, under some binding energy constraints, can cause this segment to be mutated to an equilibrium distribution which is the same as the mouse motif sites, in about 80 million years (myr) – the believed time of divergence between the two species.

The lowest and highest binding energies of the rat/mouse sequences are 1.942187 and 8.254342 respectively, while the human segments have corresponding energy bounds of 4.082253 (TAAGAATA) and 81.328394 (TTGGCCAA). The algorithm was run starting with each of the segments and using the weight matrix and binding energy limits for the mouse motif sites. The transition matrix for 100,000 years, based on the two-parameter Kimura model (with the values of $\alpha = 5 \times 10^{-3}$ s/s/myr and $\beta = 2.5 \times 10^{-3}$ s/s/myr taken from Graur and Li (2000)- “s/s/myr” is an abbreviation for “substitutions per site per myr”) and using expression (9) is:

\[
\begin{bmatrix}
A & 0.999006872 & 0.0002498750 & 0.00004995627 & 0.00002498750 \\
C & 0.0002498750 & 0.999006872 & 0.0002498750 & 0.00004995627 \\
G & 0.0002498750 & 0.0002498750 & 0.999006872 & 0.0002498750 \\
T & 0.0002498750 & 0.0002498750 & 0.0002498750 & 0.999006872
\end{bmatrix}
\]
Discarding the first 1000 out of 1800 iterations as “burn-in” before the Markov chain reaches equilibrium, the last 800 roughly correspond to about 80 myr of divergence time between human and mouse.

If the human segments indeed could correspond to the mouse motif pattern we would expect the equilibrium distribution (Figure 5) to match the distribution of the mouse sites (Figure 4). For the fourth segment (TAAGAATA) in the human motif set, we find that the equilibrium distribution is actually very close to the mouse pattern, though somewhat flattened (i.e., more uniform). Whereas for the next closest segment (TTGGCCAA) the evolutionary algorithm under the binding energy constraint fails to move the motif pattern towards the mouse pattern distribution (its binding energy is comparatively very high). So it appears that it is extremely likely that the segment TAAGAATA is actually the corresponding regulatory binding site in the human, whereas the segment TTGGCCAA most probably corresponds to a different binding site, or is possibly spurious (further inferences would need to await confirmation from biological experiments). On comparing with the experimentally verified sites in Wasserman et al. (2000) we see that indeed TAAAAATA (or its complement) is the regulatory binding motif for the transcription factor MEF2.

Quantifying the degree of motif variation. The above example indicates that a slight variation in the motif pattern may actually be due to evolutionary processes, under the model constraints. A related question of interest may be how far the motif pattern is allowed to vary by the evolutionary process while still remaining the “same”, i.e. not losing its functional characteristics. For a simulation study, we constructed a set of mutated consensus sequences (Table 2, arranged in increasing order of binding energy levels). The binding energies of none of the segments are initially within the mouse energy bounds. The evolutionary algorithm was then run for iterations corresponding to 80 myr starting with each of the segments. The sequence logos representing the distribution of the sites at equilibrium are in Figure 6 (first 5 sequences only- the last two patterns remain unaffected under the specified evolutionary constraints). Comparing with Figure 4, it appears that patterns 1,3 and 5, though starting off extremely differently, are mutated towards the mouse consensus-while 2 and 4 are still slightly different (‘A’ changed to ‘C’ in the 2nd position, but the relative frequencies are very close) after the same length of time. Even though the seventh segment TAAGGAAA has 4 mutations compared to TAAAAATA and does not evolve to the mouse motif, TTAGGGCA has 5 mutations but still does. This indicates that mutations in the more conserved positions more seriously affect the ability to evolve toward the “same” pattern- variations in highly conserved positions of the motif are more unlikely during the evolutionary process, and if they occur, are more detrimental to functionality.

4.2.4 Further thoughts

We assumed a simple model for regulatory binding site evolution and explored the possible evolution of a binding site under certain energetic constraints. From preliminary analyses it appears that this may provide a way to filter out “interesting” motif patterns (i.e., those more likely to be binding sites) by using cross-species com-
parisons from a set of found patterns in a motif search algorithm. The simulation study shows that even apparently different patterns may actually correspond to the same mutated motif, within certain bounds, that depend on the degree of conservation of each position within the site. A useful extension of this method would be to directly incorporate phylogenetic information into the motif search algorithm. A possible method could involve weighting motif patterns by their “binding energy” (with respect to some known or approximated weight matrix) before sampling possible sites. This would increase the specificity of the algorithm while allowing for a certain degree of variation from the hypothesized motif pattern\(^1\). The flexibility of the motif model could vary based on the degree of conservation of the motif type in the known species; for less conserved sites, there would be a wider bound of variation for the “corresponding” motif in the new species under study.

5 Vaccine development using a pattern discovery approach

Now we turn our focus towards the application of pattern discovery in the field of vaccine design. Pattern discovery techniques are the most important tools in understanding the host-pathogen interaction and the immune pathway, which are the essential steps towards designing new and effective vaccines. Though the statistical challenges encountered in vaccine design are similar to motif discovery problem discussed in (Sections 2 through 4), there are a few crucial differences between the two. First of all we now have amino acid (AA) sequences instead of nucleotide sequences. Moreover, to build appropriate statistical models, now we need to choose auxiliary information which has the ability to predict the structural conformation of the AA chains. First, we provide a short introduction to the immune system and show how pattern discovery techniques are used to enhance the effectiveness of current vaccines.

5.1 Immune system and vaccine design

Foreign peptides (short fragments of proteins) are carried by the antigen presenting cells (APC) from the area of attack to the thymus gland. The recognition of these peptides by the T-cells present in the thymus triggers an appropriate (antigen-specific) immune response. Actually, at a finer scale, this recognition is governed by, peptide displayed on the groove of the the major histocompatibility complex (MHC), which is bound to the APC and the T-cell receptors (TCR) present on the T-cells. But not all peptides bind to the MHC molecule. On the other hand for developing a protective immune response, we do not require T-cell recognition of all possible peptides from the entire pathogen. In fact, T-cells specific to an ensemble of a few epitopes (MHC binding peptides), or in some cases even to a single immunodominant epitope is all that is needed to produce the most effective immune response. This is the main rationale behind the paradigm of epitope-driven vaccine, where the goal is to create vaccines using only a few epitopes of a pathogen as opposed to using live attenuated, as in MMR\(^2\) vaccine or killed whole pathogens, as in the influenza vaccine. Therefore

---

\(^1\) Also, over evolutionary time there is the possibility of binding sites changing their affinity.

\(^2\) Measles Mumps and Rubella
the design of any effective epitope-driven vaccine relies hugely on the first crucial step of finding which viral peptides bind to an MHC molecule.

5.2 Statistical approaches to peptide-binding predictions

Several techniques of pattern recognition, including the ones discussed in Section 1 (now in the context of amino acids) have been used to predict MHC binders. For a review and comparison of the available methods for binder prediction see Doytchinova et al. (2004); Yu et al. (2002); Flower (2003). But, unlike in the motif discovery setup, here we can allow overlapping sequences as binders, as the same protein can be chopped off at different positions in different occasions. Also, for this application our main source of data is the lists of experimentally verified binders and non-binders. Thus the problem of detecting peptide binders is more of a problem of pattern classification than pattern recognition, and thus can be represented by a mixture model.

Assuming a peptide $S = \{s_1, s_2, ..., s_L\}$ of length $L$ can be either a binder or a non-binder, we can write

$$P(S) = \pi P(S|\text{Binder}) + (1 - \pi) P(S|\text{NonBinder}),$$  \hspace{1cm} (13)

where $\pi$ is the proportion of binders. Note that in the context of peptides each residue $s_i \in \mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, the 20 alphabet AA set (Table 3). While scanning a new protein or the whole pathogen for possible binders, the usual technique is to first form a list of all possible peptides (usually of a fixed length) and then apply classification rules on this list of peptides. This implies that we are only interested in predicting the sequence of AA’s in the peptide and not their location in the context of the whole protein or pathogen. This is another major distinction from the motif discovery problem.

Surveying the literature one notices that the existing classification techniques have a limited focus on building a probabilistic framework describing the actual binding phenomenon. Moreover, crucial information such as physical and structural properties of peptides, which have experimentally been proven to guide the peptide binding process, have not been directly utilized at the model building stage. As an alternative to the existing methods and adhering to the central theme of pattern discovery, we now discuss how additional information can be incorporated to improve two important aspects of the probability framework describing the pattern classification rules. First, in Section 6 we show how to design an appropriate transformation of the sample space based on the bio-physio-chemical properties of the AA. In Section 7 we show how the structural knowledge of peptides can be used to provide a rigorous definition of the dependence structure of the probability models of the binders and non-binders.

First, to facilitate the understanding and motivation behind our proposed framework, we give a short description of the dataset that will be analyzed in Section 6. This dataset consists of 787 experimentally verified non-binders and 359 experimentally verified binders of MHC allele $^3$ A*0201. Binding sequences were extracted from MHCPEP (Brusic et al., 1998), a database of MHC binding peptides whereas non-binding peptides were obtained through personal communications. Examining the

$^3$MHC alleles are highly polymorphic and binders for one allele may not bind to another allele. So binders are always referred to in the context of a specific allele.
profiles (Figure 7) we observe that these binders and non-binders have very strong similarity. This is because these experimentally verified non-binders, were chosen for laboratory experiments because other binder prediction methods (mainly PWM based classifiers) classified them as possible binders. This indicates that the existing methods have a limited ability to predict binders. This motivates us to develop methods specifically designed to discover the right partition beneath this apparent similarity in the binder and non-binder profile.

6 Pattern discovery using amino acid properties

The main objective of this section is to build better predictors by defining classifiers using properties of AA rather than their usual 20 alphabet representation. First, we will outline the motivation behind the new property based approach designed by Ray et al. (2005) and then demonstrate how this new approach compares to existing sequence based methods.

Motivation behind property based prediction  Modeling the actual physical phenomenon: Recent crystallographic studies of structure of peptide-MHC binding reveal that the binders conform to a very strict configuration (See Figure 8 for a visual representation of peptide binding site). Moreover, Garboczi et al. (1994); Madden et al. (1993) , through their solution of the structure of 5 peptide-MHC complexes of allele A*0201 suggests that the binding grooves of these peptides are very similar to each other. The structure of an AA chain is in turn largely dependent on the properties of constituent residues.

Interpretability: The prediction rules obtained from the property based classifiers will be defined in terms of the properties. This gives a direct interpretation to the derived rules, which can be explored through further research. For example, biologically it makes more sense to state a rule as “the 2nd position of binders are strongly hydrophobic” rather than “binders usually have either Leucine(L) or Isoleucine(I) in position 2”.

Mathematical simplicity: There is also a strong geometric motivation for exploiting the bio-physio-chemical properties of the peptides for designing an efficient classification scheme. For numerical analysis, the 20 alphabet unordered categorical amino acid space is usually coded using a binary string of 20 bits. Let us denote this space by $\mathbb{B}^{20}$. Then a 9-mer peptide is represented as

$$S_9 = \{s_1, s_2, \ldots, s_9\} \in \mathbb{B}^{180} \text{ as each } s_i \in \mathbb{B}^{20}$$

Using this we rewrite the mixture model in (13) as

$$P(S) = \pi P(S|\text{Binder}) + (1-\pi)P(S|\text{NonBinder}), S \in \mathbb{B}^{180}.$$ 

But $\mathbb{B}^{180}$ is not a regular topological space. That is, strict definition of “direction” and “distance” is lagging in this sample space. Moreover, there is the problem of high dimension. To control this over-parametrization, the most common techniques used are: geometric transformation of original data vectors (in our case the AA sequence), variable selection and introduction of statistical models. And in the context of AA’s,
the most natural way of inducing geometric transformation is through their known bio-physio-chemical properties.

In contrast to the huge literature and available software for sequence-based prediction algorithms, there have been a very limited focus on predictions based on the properties of the AA constituting the peptides. Starting with Segal et al. (2001)’s proposal of using AA properties as a means of decreasing the number of variables in a regression setup, an exhaustive list of references on property based prediction can be found in Supper et al. (2005). But, instead of building an integrated probability framework, most of these approaches use the AA properties in a “step-wise” fashion. In contrast we use these properties to design a well defined probability space on which we can build any classifier. At the moment we do not focus on comparing classification schemes, rather, we demonstrate the advantage of using property based methods over sequence based methods using a number of classifiers.

6.1 Transformation and dimension reduction

Representing an amino acid with $k$ properties is a many-to-one transformation, $T$ which can be represented by

$$T : \mathbb{R}^{20} \rightarrow \mathbb{P}_1 \times \mathbb{P}_2 \times \cdots \times \mathbb{P}_k = \mathbb{P}^k,$$

where $\mathbb{P}_j$ defines the vector space spanned by property $j$. Thus $\mathbb{P}_j = \mathbb{R}$ (real line) for continuous properties, e.g. hydrophobicity and molecular weight, whereas $\mathbb{P}_j = \mathbb{B}$ for ordinal properties, e.g. indicators of polar and aromatic. Extending the notion of this transformation to the 9-mer peptides we have

$$T : \mathbb{B}^{180} \rightarrow \mathbb{P}^{9k}.$$

In other words $T(S_9) = Y = \{Y_{11}, Y_{21}, \ldots, Y_{k1}, Y_{12}, \ldots, Y_{ji}, \ldots, Y_{19}, Y_{29}, \ldots, Y_{k9}\}$, where $Y_{ji}$, is the value of the $j^{th}$ property in the $i^{th}$ position. For a clearer presentation we can write $Y$ as a matrix

$$Y = \begin{pmatrix}
Y_{11} & Y_{12} & \cdots & Y_{19} \\
Y_{21} & Y_{22} & \cdots & Y_{29} \\
\vdots & \vdots & \ddots & \vdots \\
Y_{k1} & Y_{12} & \cdots & Y_{k9}
\end{pmatrix} = \begin{pmatrix}
Y(1) \\
Y(2) \\
\vdots \\
Y(k)
\end{pmatrix}, \quad (14)$$

where the $j^{th}$ row, $Y(j)$, represents the value of the $j^{th}$ property corresponding to the 9 amino acids (AA) in the sequence. Note that $\mathbb{P}^{9k}$ is an usual euclidean space. For example, the categorical AA can be represented in a 3 dimensional space with 3 continuous properties, selected from Table 3, as the basis vectors (Figure 9). Additionally, carefully chosen properties may provide desired dimension reduction and significant increment in signal extraction. But these properties should be chosen cautiously, as we may perform less optimally when the chosen properties do not capture the distinction between binder and non-binders.

6.2 Property based classification

Now we apply the property-based analysis on the dataset described in Section 5. We demonstrate the advantage of using property based method using three machine
learning approaches, Support Vector Machine (SVM) (Vapnik, 1995), Random Forest (RF) (Breiman, 2001), and Bagging (BAG) (Breiman, 1996).

6.2.1 Variable selection

The first step towards building a property based classifier is to define the appropriate transformation, i.e. to select the properties on which the projection will be defined. We have used a simple forward selection method to select the important variables, selecting the same property for all the 9 positions at each stage. Using the matrix representation in Eq.(14) this reduces to selecting only row labels, instead of the individual entries. We start with the single best property corresponding to the minimum cross-validation based misclassification rate (CVBM), and sequentially add one property at a time, until there is no gain in the CVBM. If \( k_0 \) properties are chosen, we can rewrite Eq.(13) as

\[
P(Y) = \pi P(Y|Binder) + (1-\pi)P(Y|NonBinder), \ Y \in \mathbb{R}^{9k_0} \quad (15)
\]

Starting with the set of properties listed in Table 3 our classifiers selected hydrophobicity, volume and isoelectric point (denoted as Hydro, Vol, Iso in Table 3) as being the three most important properties. Though BAG did not select the variable Volume, it selected Area, which is highly correlated with Volume. Moreover, within the given set of properties SVM achieves the lowest misclassification rate with 5 selected properties (hydrophobicity, volume, isoelec, branch, aromatic). Also, among all classifiers, based on a maximum of three properties, SVM achieves the lowest misclassification rate.

6.2.2 Global comparison using AROC

One common criticism of using misclassification rate as a measure of comparison is its dependence on tuning parameters of the specific classifiers. So, for comparing the performance of the property based and sequence based classifiers we use the criterion of Area under the Receiver Operating Characteristic curve (AROC). This will also enable us to compare our results with other competitive methods.

To draw the comparison we calculate the AROC values using

(i) Original amino acid Coding (Categorical)
(ii) Hydrophobicity (Hyd)
(iii) Volume (Vol)
(iv) Isoelectric Point (Iso)
(v) All properties (ii)-(iv) (3 prop)

A comparison of the AROC values (Figure 10) clearly demonstrates that the property-based methods perform considerably better than the sequence based method (AROC 0.90-0.94 vs 0.81-0.84). Note that our classifiers are not especially designed for sequence analysis. We found that the AROC values of the 6 widely used classifiers, especially
designed for MHC-peptide analysis actually range from 0.81-0.87 (see Table 1 of Yu et al., 2002), which is very close to the sequence based AROC values. We can draw two immediate conclusions from the last comparison. First, property-based classifiers outperform especially designed sequence based classifiers. Moreover, this study provides enough evidence to suggest that other widely used classifiers can be improved upon by using the property based approach.

7 Using HMMs to classify binders and non-binders

Now we concentrate on building a probabilistic model of the peptide sequence. As mentioned in Section 3, HMMs are a natural way of modeling conserved patterns of sequence data. But the use of HMM for binder-prediction has mainly focused on building profiles which incorporate a dependence structure between adjacent residues (Yu et al., 2002; Mamitsuka, 1996). On the other hand, Zhang et al. (1998), Schueler-Furman et al. (2000), and Altuvia and Margalit (2004) use the structural information of peptides solely as explanatory variables or predictors. In this section, we will demonstrate how structural knowledge of peptides and MHC-molecules can be incorporated in the HMM framework to model the MHC-binding phenomenon. Although our initial discussion and model development will be based on AA sequence, $S$, we will later show how the property based data $Y$ can be used for building profile HMMs.

7.1 Structural information based HMM

Although the underlying principle of HMMs used in predicting motif sites and MHC-binders is essentially the same, there are a few basic differences in the implementation and objectives of the two problems. On one hand, in the case of MHC-I molecule the length of the peptides are predictable, so the issue of alignment is less important. On the other hand there is a huge increase in number of parameters as we are now dealing with AAs instead of nucleotides.

To build the HMM we start with a 9-mer peptide and denote the 9 positions by $P_1, \ldots , P_9$. Motif finding and classification algorithms have clearly shown that for allele A*0201, the positions $P_2$ and $P_9$ are the most conserved positions and X-ray crystallographic studies have strongly supported this conclusion. But, several researchers in this field and our own observations (Figure 7) indicate that other positions play a very important role in discriminating binders from non-binders. Adjacent positions to the binding pockets may interfere with the binding of the whole peptide. Zhihua et al. (2004) states that there are 6 binding pockets, (we label them A through F) in the MHC molecules (Figure 11 gives a schematic representation), of which $P_2$ and $P_9$ are accommodated by the binding pockets B and F (studied by Ruppert et al., 1993; Kubo et al., 1994; Parker et al., 1994). But not much is known about how the 7 remaining positions $P_1, P_3, P_4, P_5, P_6, \text{ and } P_8$ bind into the 4 remaining pockets A, C, D, E. Each of these positions may be occupied by a peptide residue or it may remain empty.

Here we build an HMM which requires a peptide to bind to all 6 pockets, with $P_2$ and $P_9$ fitting pockets B and F. First we define our state space

$$H = \{B, E, M_1, M_2, M_3, M_4, M_5, M_6, I_2, I_3, I_4, I_5\}.$$
where $B$ and $E$ denotes the Begin and End States, $M_j$ denotes the “state” of a peptide residue fitting into the $j^{th}$ binding pocket of the MHC (match), while $I_j$’s denote the different intermediate states of the residues not fitting in any of the binding pockets (insertion). Here we use several $I_j$’s instead of one single insert state as the transition probabilities from each insert state to a particular match states may be different.

To completely specify the HMM, we need to define the emission and transition probabilities. We use the underlying structure of the proposed HMM (Figure 12) and the knowledge of amino acids to define these probabilities. First, we will assume that the emission probabilities of each insert state is the same, i.e.

$$
\mu_{ak} = P(s_i = a|h_i = I_k) = P(s_i = a|h_i = I_l) = \mu_{al} \quad \forall \ l, k, \text{ and } a \in A.
$$

For the match states, however, $P(s_i = a|h_i = M_k) (k = 1, \ldots , 6)$ may vary for different $k$. This leaves us with the estimation of $19 \times 7$ emission probabilities. Although the dimension of the transition probability matrix is $13 \times 13$, we need to estimate only 13 nonzero entries to completely specify $\tau$. They are

$$
P(I_i|I_i) = 1 - P(M_{i+1}|I_i) \quad i = 2, \ldots , 5 \text{ and } P(I_i|M_i) = 1 - P(M_{i+1}|M_i) \quad 1 = 0, \ldots , 5, \text{ where } M_0 = B \text{ and } M_7 = E.
$$

Without going into the details of parameter estimation of HMMs (but following the overall scheme of Section 3), we demonstrate how a new 9-mer is scored and classified into a binder or a non-binder. Here, we define two sets of scoring functions, one solely based on the binder data while the other includes both binder and non-binder data.

Let $\hat{\theta}^b$ denote the estimate of the proposed profile HMM. The probability of a 9-mer $S$ is generated from this profile HMM is given by

$$
P(S|\hat{\theta}^b) = \prod_{i=1}^{9} \sum_{h \in H} P(s_i = h_i = k) P(h_i = k|h_{i-1} = l), \text{ where } h = \{h_1, \ldots , h_9\} \quad (16)
$$

This probability itself can be used as a score. But, a general criticism of this likelihood based method is the dependence of this probability on the sequence length (Durbin et al., 2000). We may alternatively calculate the length independent score using the ratio of posterior probabilities, given by

$$
E_{lo}(S) = \log \frac{P(Binder|S)}{P(Binder|\hat{\theta}^b)} = \log \frac{P(S|\hat{\theta}^b) \hat{\pi}}{P(S|\hat{\theta}^b)(1 - \hat{\pi})}
\quad = \log \frac{\hat{\pi}}{(1 - \hat{\pi})} + \log \frac{P(S|\hat{\theta}^b)}{P(S|\hat{\theta}^0)}, \quad (17)
$$

where $\theta^0$ denotes the parameters of a background model and $\pi$ denotes the proportion of samples coming from the binder data. Also, equation (17) provides the decomposition of the score $E_{lo}$ into the log odds of the prior and the usual log-odds score. This score, or its approximate values, can be calculated using available dynamic programming techniques (see Section 5.4 of Durbin et al., 2000). Usually, $\theta^0$ is an insertion-only
model and thus we have the following simplification

\[ P(S|\theta^0) = \prod_{i=1}^{9} P(s_i|h_i = I) = \prod_{i=1}^{9} MN(s_i; \alpha^0) \propto \prod_{j=1}^{20} (\alpha_j^0)^{c_j}, \tag{18} \]

where \( c_j \) is the frequency of amino acid \( a_j \) in the 9 mer, \( MN \) represents a multinomial distribution, and \( \alpha_i^0 \) denotes the probability of a amino acid \( a_i \) in the protein pool (all proteins produces by the pathogen).

As the experimentally verified non-binders are quite similar to the binders, \( E_{b0} \), which is scored against the background, may not have the ability to discriminate between potential binders and the non-binders. So we propose the use of a scoring system which uses the non-binder data to build a mixture model as proposed in Eq.(13). Denoting \( \theta^b \) as the parameter of the profile HMM corresponding to the non-binders, we may define the new score

\[ E_{bn}(S) = \log \frac{P(Binder|S)}{P(Nonbinder|S)} = \frac{P(S|\theta^b)\pi_{bn}}{P(S|\theta^0)(1 - \pi_{bn})} = \log \frac{\pi_{bn}}{(1 - \pi_{bn})} + \log \frac{P(S|\theta^b)}{P(S|\theta^0)}, \]

where \( \pi_{bn} \) denotes the proportion of binders in the training set. However, the non-binder data is distinctly different from the background (see Figure 7); in fact these non-binders are actually false positives selected by PWM classifiers (selected on the basis of conservation of P2 and P9). So we propose a two-pocket model, which are filled by the P2 and P9 and the rest of residues are assumed to be coming from insert states. Using \( M_1 \) and \( M_2 \) for these two pockets, we can write

\[ P(S|\theta^a) = P(s_2|M_1)P(s_9|M_2) \prod_{i \in \{1,3,4,5,7,8\}} P(s_i|h_i = I) = MN(s_1; \alpha^{(2)})MN(s_9; \alpha^{(9)}) \prod_{i \in \{1,3,4,5,7,8\}} MN(s_i; \alpha^0)MN, \tag{19} \]

where \( \alpha^{(2)} \) and \( \alpha^{(9)} \) are the relative frequency of AA at 2nd and 9th positions, respectively. Note that here we are assuming independence of positions (similar to a PWM model), but a more complicated model can be built using dependence between the positions.

**Note:** Rather than treating the HMMs designed in this chapter as the perfect solution to our classification problem they should be viewed as a general template for building profile HMMs based on structural information of peptide-MHC interaction. Richer structural information together with any other auxiliary data can be easily incorporated in this framework to build more realistic models.

### 7.2 Property based profile HMM

Now we outline the steps behind using the property-based transformation we obtained in Section 6 in building an HMM based classifier. Keeping the basic structure of the hidden states unaltered, the HMM may be built by defining the emission probabilities
using the transformed probability space defined on $\mathbb{R}^{9k}$, spanned by $k$ selected properties. For simplicity we assume $\mathbb{R}^{9k} = \mathbb{R}^k$, i.e. all properties selected are continuous in nature.

For position $i$, instead of an AA we now observe a multivariate property vector $Y_i = (Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{ik}) \in \mathbb{R}^k$, where $Y_{ij}$ represents the $j^{th}$ property in the $i^{th}$ position. Keeping the underlying transition probability structure of the profile HMM unaltered, we now assume $Y_i \sim g_k[\cdot; \theta_k]$, where $g_k$ is a continuous multivariate distribution with parameter $\theta_k$. Assuming a specific functional form of $g_k$ (e.g. multivariate normal), an EM algorithm can be formulated to estimate $\tau_j$ and $\theta_j$ (Section 4.1).

This property-based HMM has two clear advantages over the sequence profile-based HMMs. First, the parameter space (corresponding to the emission probabilities) is drastically reduced. Moreover, the property space, rather than the actual sequence, may provide a more biologically interpretable binding model.

8 Concluding remarks

In this chapter our overall objective was to introduce the ideas of sequence pattern discovery and pattern recognition as relevant to the fields of gene regulation and understanding immunological response to infectious disease. We have discussed models and methods ranging from the simple weight matrix-based approach to generalizations based on more complex dependence structures through a hidden Markov model framework. The general theory and techniques are applicable in a variety of sequence analysis problems including the two specific ones discussed here. However, it cannot be over-emphasized that using context-specific auxiliary information relevant to the underlying biology has a drastic power to increase the accuracy of these methods. A current focus in the computational biology field is thus tending towards developing the most effective ways to combine multiple data sources and data types with the goal of making the most accurate biological inference.

References


Sequence pattern discovery


M. Gupta and S. Ray


Sequence pattern discovery


Table 1: Consensus motifs for sites found in the mouse and human skeletal muscle regulatory sequences.

<table>
<thead>
<tr>
<th>Mouse motif consensus</th>
<th>Number of sites</th>
<th>Human motif consensus</th>
<th>Number of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAAAAATA/TATTTTA</td>
<td>38</td>
<td>GTGGGTGG/CCACCCAC</td>
<td>21</td>
</tr>
<tr>
<td>GTGACTG/CAGTCACG</td>
<td>17</td>
<td>TCGGGGAC/GTCCGCGA</td>
<td>18</td>
</tr>
<tr>
<td>GCATTGTG/CACAAATGC</td>
<td>15</td>
<td>CAGCTGTC/GACAGCTG</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TAAGAATA/TATTTTA</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTGGCCAA/TTGGCCAA</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CACCCAGC/GCTGGGTG</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2: Mutated consensus sequences for the simulation study.

<table>
<thead>
<tr>
<th>Mutated pattern</th>
<th>Binding energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAAGAAAA</td>
<td>23.003709</td>
</tr>
<tr>
<td>TAAGGACA</td>
<td>24.890779</td>
</tr>
<tr>
<td>TTTGAAAA</td>
<td>26.874910</td>
</tr>
<tr>
<td>TAAGGGCA</td>
<td>27.781150</td>
</tr>
<tr>
<td>TTAGGGCA</td>
<td>29.372910</td>
</tr>
<tr>
<td>TAACCATA</td>
<td>39.814952</td>
</tr>
<tr>
<td>TAAGGAAA</td>
<td>41.925165</td>
</tr>
</tbody>
</table>
Sequence pattern discovery

(a) (b)

<table>
<thead>
<tr>
<th>TAGAAT</th>
<th>TATACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATTAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TAGAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TATAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TATAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TAGACT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TATAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TAGAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TATAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TATAAT</td>
<td>TATAAT</td>
</tr>
<tr>
<td>TATAAT</td>
<td>TATAAT</td>
</tr>
</tbody>
</table>

Figure 1: (a) The alignment of 20 TATA-box binding sites in a B. subtilis data set; (b) the corresponding position specific weight matrix for the alignment in (a); and (c) the corresponding sequence logo (Stephens and Schneider, 1990). The height of each letter in the stack is proportional to its frequency at the site; the height of the stack is adjusted to reflect the total information content at the site. Inverted letters represent the ones observed less often than expected under the background model.

(a) (b)

Sequence $i$

<table>
<thead>
<tr>
<th>w</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{ij} ) = 1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: (a) The repetitive block-motif model and (b) the mixture model approximation.
Figure 3: (a) Correlation of ranks of top 99.99\textsuperscript{th} percentile of predicted sites with their ChIP measurement-based ranks, and Pu.1 motif logos from (b) MDscan (c) HMM-based method (d) TRANSFAC database (Heinemeyer et al., 1998).

Figure 4: Sequence logo for 38 mouse motif sites.

Figure 5: Equilibrium distribution of human motif sites after 80 myr. left panel: starting from segment 4 (TAAGAAAA); right: starting from segment 5 (TTGGCCAA).

Figure 6: Equilibrium distribution of constructed motif sites after 80 myr. From top corner (left to right), corresponding starting segments are (1) TAAGAAAA, (2) TAAGGACA, (3) TTTGAAAA, (4) TAAGGCA, (5) TTAGGCA. The last two sites are not shown as they exhibit no change from the start point under the used evolutionary rule.
Table 3: Selected amino acid properties extracted from Kawashima et al. (1999)

<table>
<thead>
<tr>
<th>AA</th>
<th>Full Name</th>
<th>Mol. Weight</th>
<th>Vol</th>
<th>Area</th>
<th>Hydro</th>
<th>Iso</th>
<th>Aliphatic</th>
<th>Aromatic</th>
<th>Branch</th>
<th>Sulfur</th>
<th>Cyclic</th>
<th>Hydro.bond</th>
<th>Nonpolar</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>alanine</td>
<td>89.09</td>
<td>88.6</td>
<td>115</td>
<td>1.8</td>
<td>6.00</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>cysteine</td>
<td>121.16</td>
<td>108.5</td>
<td>135</td>
<td>2.5</td>
<td>5.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>aspartate</td>
<td>133.10</td>
<td>111.1</td>
<td>150</td>
<td>−3.5</td>
<td>2.77</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>glutamate</td>
<td>147.13</td>
<td>138.4</td>
<td>190</td>
<td>−3.5</td>
<td>3.22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>phenylalanine</td>
<td>165.19</td>
<td>189.9</td>
<td>210</td>
<td>2.8</td>
<td>5.48</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>glycine</td>
<td>75.07</td>
<td>60.1</td>
<td>75</td>
<td>−0.4</td>
<td>5.97</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>histidine</td>
<td>155.15</td>
<td>153.2</td>
<td>195</td>
<td>−3.2</td>
<td>7.47</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>isoleucine</td>
<td>131.17</td>
<td>166.7</td>
<td>175</td>
<td>3.8</td>
<td>5.94</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td>lysine</td>
<td>146.12</td>
<td>168.6</td>
<td>200</td>
<td>−3.9</td>
<td>9.59</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>leucine</td>
<td>131.17</td>
<td>166.7</td>
<td>170</td>
<td>3.8</td>
<td>5.98</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>methionine</td>
<td>149.21</td>
<td>162.9</td>
<td>185</td>
<td>1.9</td>
<td>5.74</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>asparagine</td>
<td>132.12</td>
<td>114.1</td>
<td>160</td>
<td>−3.5</td>
<td>5.41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>proline</td>
<td>115.13</td>
<td>112.7</td>
<td>145</td>
<td>−1.6</td>
<td>6.30</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Q</td>
<td>glutamine</td>
<td>146.14</td>
<td>143.8</td>
<td>180</td>
<td>−3.5</td>
<td>5.65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>arginine</td>
<td>174.20</td>
<td>173.4</td>
<td>225</td>
<td>−4.5</td>
<td>11.15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>serine</td>
<td>105.09</td>
<td>89.0</td>
<td>115</td>
<td>−0.8</td>
<td>5.68</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>threonine</td>
<td>119.12</td>
<td>116.1</td>
<td>140</td>
<td>−0.7</td>
<td>5.64</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>valine</td>
<td>117.15</td>
<td>140.0</td>
<td>155</td>
<td>4.2</td>
<td>5.96</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>tryptophan</td>
<td>204.23</td>
<td>227.8</td>
<td>255</td>
<td>−0.9</td>
<td>5.89</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>tyrosine</td>
<td>181.19</td>
<td>193.6</td>
<td>230</td>
<td>−1.3</td>
<td>5.66</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 7: Sequence logo (Stephens and Schneider, 1990) of position specific conservation of (a) Binders in database (b) Non-binders in database (c) Background (random 9 mer)

Figure 8: Rasmol (Sayle and Milner-White, 1995) plot of MHC I molecule B*5301 complexed with peptide Ls6 (KPIVQYDNF) from the malaria parasite *P. falciparum* (PDB ID:1A1O et al., 1996) (viewed from two different angles). The MHC-molecule is represented by a ribbon-like strand, while the peptide is shown as the dark *space fill* structure.
Figure 9: Representation of amino acids in $\mathbb{R}^3$ described by three protein properties: hydrophobicity, volume, and isoelectric point.
Figure 10: AROC values for allele A*0201 categorized by (a) variables used (b) classifiers.

Figure 11: Schematic diagram showing 6 binding pockets and 9 peptides for describing peptides binding to MHC-allele A*0201.

Figure 12: Proposed profile HMM for peptide binding to MHC-allele A*0201.