Sequential Pattern Discovery under a Markov Assumption

Technical Report No. 02-08
Information and Computer Science Department,
University of California, Irvine

Darya Chudova, Padhraic Smyth

Information and Computer Science
University of California, Irvine
CA 92697-3425
{dchudova,smyth}@ics.uci.edu

Abstract

In this paper we investigate the general problem of discovering recurrent patterns that are embedded in categorical sequences. An important real-world problem of this nature is motif discovery in DNA sequences. There are a number of fundamental aspects of this data mining problem that can make discovery "easy" or "hard"—we characterize the difficulty of learning in this context using an analysis based on the Bayes error rate under a Markov assumption. The Bayes error framework demonstrates why certain patterns are much harder to discover than others. It also explains the role of different parameters such as pattern length and pattern frequency in sequential discovery. We demonstrate how the Bayes error can be used to calibrate existing discovery algorithms, providing a lower bound on achievable performance. We discuss a number of fundamental issues that characterize sequential pattern discovery in this context, present a variety of empirical results to complement and verify the theoretical analysis, and apply our methodology to real-world motif-discovery problems in computational biology.

1 Introduction

Large data sets in the form of categorical sequences (defined on a finite alphabet of symbols) frequently occur in a variety of real-world applications. Examples of such data sets include DNA and protein sequences in computational biology, alarm message sequences in telecommunication networks, and sequences of Web-page requests or actions in user modeling. An important problem in this context is the unsupervised identification of recurrent patterns in such sequences, i.e., the detection and discovery of relatively short, relatively rare, possibly noisy, repeated substrings in the data.

For example, in computational biology such patterns are known as motifs. Motifs can be thought of as short highly-conserved regions in a DNA sequence (Stormo and Hartzell (1989); Lawrence et al. (1993); Bailey and Elkan (1995); Pevzner and Sze (2000)). DNA has a 4-letter alphabet and typical motifs range from 5 to 45 base-pairs (symbols) long. In motif discovery there may be some prior knowledge on the number of motifs (e.g., one per sequence) and their exact or expected lengths, but there is typically little or no knowledge on where the motifs occur in each sequence or what symbols they contain. Motif-discovery problem is an important motivator for the work we will describe in this paper; however our goal will be on understanding the general nature of pattern discovery in sequences.

As an example consider a 4-letter alphabet `A, B, C, D` and a pattern `ADDABB` embedded in some background process, e.g.,

```
....BACADBADBBC[ADDABB]BACDBDBA[ADDACB]DAC...
```

The noisy occurrences of the true patterns in this sequence are enclosed in brackets. Also note that there are other locations in the sequence where the true pattern `ADDABB` can have a partial match to the background, e.g., the subsequence `ADBADB` that starts 3 symbols into the sequence. In a long sequence there may be many such spurious "background matches," leading to false detections, and making both detection and discovery quite difficult.

In particular we are interested in determining what makes this problem hard from a learning viewpoint. What is the effect on pattern discovery of alphabet size? of sequence length? of pattern frequency? There is a long tradition in statistical pattern recognition and machine learning of providing mathematical bounds on the difficulty of learning problems as a function of fundamental problem characteristics.

There are three different types of "pattern models" of interest, ranging from the case where we have perfect knowledge of the pattern to where we have very little knowledge:

1. **Model 1, True Model:** in this case the true probabilistic model for generating patterns and background is known, and we can use this model to optimally classify each symbol in a sequence into either the pattern or background class. The lowest achievable error rate associated with the optimal decision rule is known as the Bayes error rate (Duda, Hart, and Stork, 2001).

2. **Model 2, Supervised Training:** we know the general form of the pattern and background models, and the locations of the patterns within training sequences. We can estimate the parameters of the models and solve the *pattern detection* problem by using the estimated model to classify symbols in the new sequence into pattern and background classes.

3. **Model 3, Unsupervised Training:** we know the general form of the pattern and background models, but both the parameters of the model and the locations of the pattern-background boundaries are unknown (the *pattern discovery* problem).

The pattern discovery problem associated with Model 3 is the problem of primary interest in this paper. The objective function that we use to measure the quality of a learned pattern-background model is the average number of errors a model makes when classifying symbols from a new test sequence. Clearly, the classification error rate of a model where the parameters are fixed to their true values (Model 1) will on average be lower than that of a model where the parameters are learned in a supervised fashion (Model 2). In turn, Model 2 will on average have a lower error rate than a model learned in an unsupervised fashion (Model 3, the model we are primarily interested in).

Characterizing the error rate of Model 3 directly is potentially a rather complex task. However, we can use the Bayes error rate of Models 1 and 2 as lower bounds on the error rate of Model 3 (the model provided with the least information). This is the general strategy we pursue in this paper. The motivation is that if the Bayes error rate for a problem is high (the error rate of Model 1), then pattern discovery for Model 3 will also be difficult, since the best possible model we can learn via pattern discovery can never do better than the Bayes error rate.

The outline of the paper is as follows. We introduce a general hidden Markov model framework for pattern discovery in Section 2, and Section 3 follows with a review of related work. In Section 4 we derive an approximate analytical expression for the Bayes error under the hidden Markov model assumption and illustrate the insights that it provides into the problem. Section 5 analyzes the effect of autocorrelation structure in a pattern. In Section 6 we describe and compare three different well-known probabilistic algorithms for pattern discovery. In Section 7 we apply the Bayes error rate analysis to a set of real motif finding problems. Section 8 contains discussion and conclusions.

The primary novel contributions of this paper are as follows:

- We provide an accurate approximate expression for the Bayes error rate for pattern discovery under a Markov assumption. To our knowledge there has been no previous investigation of Bayes error for such problems.
- We illustrate how different factors such as alphabet size, pattern length, pattern frequency, and pattern autocorrelation can directly affect the Bayes error rate.
- We empirically investigate several well-known algorithms for pattern discovery in a Markov context and examine their sensitivity to training data size, accuracy and strength of prior knowledge, and difficulty of the underlying discovery problem.
- We apply these techniques to motif-finding problems and demonstrate how the theoretical framework of the Bayes error rate can shed light on the reliability of automated algorithms for real data.

2 A Markov Model for Sequential Patterns

We use the following notation throughout the rest of the paper:

- n_A denotes the size of the observable alphabet.
- L denotes the length of a pattern, for fixed-length patterns.
- The *consensus* pattern is a subsequence of L symbols, where the ith symbol is the most likely symbol to appear in position i of the pattern, $1 \leq i \leq L$.
- ε denotes the "noise" probability of a substitution error in each of the pattern positions, i.e. $(1 - \varepsilon)$ is the probability of the consensus symbol appearing in each position.

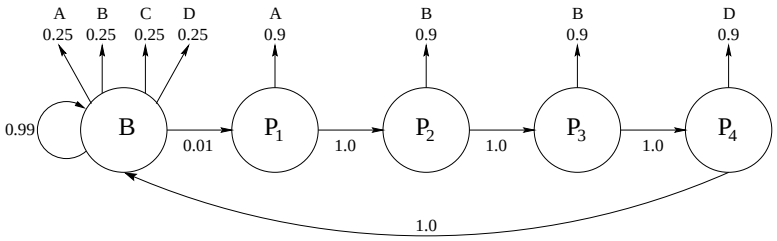

Figure 1: Example of the HMM state transitions for a pattern whose most likely instantiation is $ABBD$ embedded in a uniform background sequence.

- n_s denotes the expected number of substitutions in a pattern, $n_s = L \times \varepsilon$.
- F denotes the frequency of pattern occurrence in the sequences, so that the expected number of patterns in a sequence of length N is given by $F \times N$.

We will first consider the model for noisy fixed-length patterns: two occurrences of the pattern can differ only due to substitution errors. We will assume that the substitutions occur independently in each position. Although these assumptions on pattern structure are quite simple it has nonetheless proven to be very useful as a model for motif discovery in computational biology (see Stormo and Hartzell (1989); Liu et al. (1995); Pevzner and Sze (2000); Buhler and Tompa (2001) for detailed discussions). We use a hidden Markov model (HMM) framework to provide a generative model for simulating sequences with embedded patterns. The HMM generates state sequences with Markov dependence where the states are not directly observed. In each of the states, an observed symbol is produced according to a state-specific multinomial distribution.

We use an HMM with $(L + 1)$ states for generating patterns of length L. A single state B models the background process, while L pattern states, P_1 to P_L, correspond one-to-one to L pattern positions. For the background state B the multinomial distribution on symbols corresponds to the frequency with which each symbol appears in the background. For the pattern states P_i, $1 \leq i \leq L$, the multinomial distribution is "tuned" to a specific symbol for that position. We assume a strictly "linear" state transition matrix, where

- the background state B can only transition to either itself or the first pattern state P_1,
- each pattern state P_i can only transition to state P_{i+1}, $1 \leq i \leq L - 1$,
- the last pattern state P_L can only transition back to the background state B.

Figure 1 shows a model for generating patterns of length 4 for a consensus pattern `ABBD`, using a 5-state HMM.

More generally, it is straightforward to extend the model to handle variable length patterns by including specialized insertion and deletion states, as is often done in computational biology for multiple sequence alignment (Baldi et al. (1994); Eddy (1995)). In general the pattern and background can be modeled as being higher-order Markov, or entirely non-Markov (e.g., using stochastic context-free grammars). We focus our attention in this paper on the first-order hidden Markov model since it can be viewed as a baseline for sequential pattern learning.

The main focus of this paper is on characterizing how easy or hard pattern discovery tasks are for this form of HMM by investigating the Bayes error rate for problems of this nature.

3 Related Work

Relevant prior work for the results reported in this paper can be categorized into two general classes: (1) earlier efforts to derive analytical expressions for the Bayes error rate under a Markov assumption, and (2) more recent work on specific algorithms for motif discovery in computational biology.

3.1 Related Work on the Bayes Error Rate

First we consider prior work on the Bayes error rate, P_e^*, in a Markov context, where the Bayes error is defined as the minimum achievable error rate in classifying symbols into classes, and the classes have Markov dependence. In our case, the classification problem is equivalent to performing inference in a hidden Markov model: recovering a sequence of the hidden states (classes) that have generated an observed sequence.

There is a significant amount of prior work in the classification literature on analyzing P_e^* in the memoryless case. The concept of the Bayes error, the error rate of the optimal classifier for a given problem, has led to fundamental and important insights into the nature of classification in multivariate feature spaces (e.g., Chow (1962); Duda et al. (2001); Fukunaga (1990); McLachlan (1992); Ripley (1996)).

When the classes have Markov dependence, Raviv (1967) demonstrated how this dependence can be used to improve the quality of classification of printed text. However, there has been relatively little prior work on deriving closed-form expressions for P_e^* in HMMs, perhaps not surprisingly since deriving such an expression as a function of the hidden Markov model parameters appears difficult if not impossible for the general case.

Earlier efforts have concentrated on the development of lower and upper bounds on the probability of error as a function of the separation between the emission distributions for each state (or class), without using any assumptions on the functional form of the distributions. For the simplest case of a two-state process, Chu (1974) and Lee (1974) provide bounds on the Bayes error in a Markov context using *sub-optimal* decision rules that only take into account one or two preceding symbols rather than the whole observed sequence. Even though these bounds are tight for the *sub-optimal* procedure and can handle arbitrary emission distributions, they can not be extended to take into account memory beyond the near-neighbors of the current state, larger contexts, or models with more than 2 hidden states. However, it is relatively easy (for example) to create Markov chains where the information beyond the near-neighbors can greatly help in reducing the error rate (such as those that will be investigated in this paper). Hence, the bounds produced by these approaches will not be tight for many problems.

3.2 Related Work on Motif Discovery

A variety of different motif discovery algorithms have previously been proposed and successfully used in practice in computational biology. Some of the best known algorithms are based on probabilistic models of the data similar to the one discussed in this paper — the Motif sampler algorithm of Liu et al. (1995) and the MEME algorithm of Bailey and Elkan (1995). These algorithms are covered in detail in Section 6, where we empirically compare the performance of these algorithms on simulated problems. Here we very briefly mention some of the other recently proposed algorithms for discovery of fixed length motifs in DNA sequences: combinatorial search techniques are employed in Pevzner and Sze (2000); a hill climbing optimization algorithm was proposed in Hu et al. (1999); an algorithm for detection of over-represented exact L-mers was proposed in Helden

et al. (1998); an algorithm for finding over-represented non-exact L-mers was proposed in Buhler and Tompa (2001).

This list illustrates the wide variety of approaches that have been used for the motif discovery problem. In what follows we provide a theoretical bound on the performance of any algorithm that adopts a probabilistic model of the data similar to the HMM described above. The commonly made assumption of independent substitutions allows us to cast many of the proposed models and algorithms into probabilistic models that are either special cases of, or very similar to, the HMM model proposed in this paper.

4 Analysis of the Bayes Error Rate

4.1 Motivation

The difficulty of discovering or learning a pattern can be characterized along multiple dimensions. For example, intuitively we expect that it will be affected by the size of the observable alphabet n_A, the length of the pattern L, the variability within the pattern as characterized by the substitution error ε, the frequency of pattern occurrence F, the similarity of the pattern and the background distributions, and the amount of available training data. Rather than characterizing learnability along each of these dimensions, we instead look at a single characteristic, the Bayes error rate, that fundamentally quantifies the difficulty of detecting a pattern. The Bayes error rate provides a tight lower bound on the error that can be obtained with any learning algorithm for a given data-generating process and can be used to compare different problems on a single scale in terms of their complexity. The Bayes error can only be computed exactly if we know the true conditional distributions and class probabilities for the classification problem. While for most practical problems the true distributions are not known, these theoretical results nonetheless provide fundamental insight into the nature of multivariate classification problems (clarifying the roles of marginal class probabilities and separability of class-conditional distributions for example, see Fukunaga (1990) and McLachlan (1992)).

Let $\mathbf{O}$ denote the sequence of observed symbols, and let o_j be the jth element of the observed sequence; let $\mathbf{H}$ denote the sequence of hidden states, and let h_j be the jth element of the state sequence. We can think of the pattern detection problem as a two-class classification problem where we wish to classify each location j in the sequence as either coming from the background $h_j = B$ (class 1) or the pattern $h_j = P_{1\ldots L} = P_1 \vee P_2 \vee \ldots \vee P_L$ (class 2).

If the state sequence were memoryless (independent draws from a distribution on state values, rather than being Markov) then we have a standard non-sequential classification problem and the Bayes error rate is defined as:

$$P_e^* = \sum_o \min_h \{\ p(h = B|o),\ p(h = P_{1\ldots L}|o)\ \}\, p(o) \tag{1}$$

where h is the state value and the sum is over the n_A different values that the symbol o can take in the alphabet. For each value o, the optimal Bayes decision is to choose the more likely (maximum probability) value of B or $P_{1\ldots L}$. The probability of making an error is the probability that the minimum probability event actually occurred, averaged over the different possible values of o.

When we add Markov dependence, to make an optimal decision on the class at location j we must now consider not only the observation o_j but the *whole sequence* of observed values $\mathbf{O}$. We can define an expression for the Bayes error rate for this Markov case that is analogous to that in Equation 1 for the IID case. The Bayes error is now the average per symbol error rate. We can

define this as:

$$P_e^* = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0} N \min_h \{\ p(h_i = B|\mathbf{O}),\ p(h_i = P_{1\ldots L}|\mathbf{O})\ \} \tag{2}$$

Intuitively, this "per-symbol" Bayes error rate corresponds to the fraction of symbols in an infinitely long realization $\mathbf{O}$ of the hidden Markov process that would be misclassified using an optimal Bayes decision rule. The optimal Bayes decision rule for any location in the sequence (i.e., the calculation of $p(h_i = B|\mathbf{O})$ and $p(h_i = P_{1\ldots L}|\mathbf{O})$) can be computed in time linear in the length of the sequence using the well-known forward-backward algorithm for HMMs (Rabiner, 1989). The classification mistakes occur whenever a background symbol appears more likely to have been produced by the pattern state rather than the background state, given the context, or vice-versa. The Bayes error rate in principle indicates how difficult it is (even in the case of Model 1 with "perfect knowledge") to isolate occurrences of the pattern from the sea of background.

4.2 Analytical Expressions for the Bayes Error Rate

Given a particular HMM structure, we can estimate the Bayes error rate in two ways. We can seek a closed-form expression for P_e^* as defined by Equation 2, or we can empirically estimate P_e^* on long sequences (where the true states are known) by counting the average number of per-symbol classification errors made by the forward-backward algorithm. The disadvantage of the empirical approach is that it does not reveal the functional form of dependence of the Bayes error on the parameters of the model. Thus, it is useful to pursue closed-form analytical expressions for P_e^*. We derive closed-form expression under assumption of uniform background probabilities and uniform substitution probabilities.

The Bayes optimal decision that classifies each symbol into the background or pattern class depends on the *whole* observed sequence $\mathbf{O}$. This induces a dependence between the decisions that are made with respect to each symbol. A useful approximation is to ignore this dependence and to consider the following simplified IID problem: given the observed sequence $o_1, \ldots, o_N$, classify each position i independently as being either the first pattern position or not, given the current symbol and the next $L-1$ symbols $o_i, \ldots, o_{i+L-1}$. This simplifies the problem because of the reduction of context in each position from the whole observable sequence to an adjacent set of L symbols. This simplification was also used by Lawrence et al. (1993) and Liu et al. (1995). In practice, for motif applications for example, we have found that the posterior class probabilities obtained from the IID assumption are often very close to the ones obtained from an HMM decision rule based on the forward-backward algorithm. The Bayes error rate for this IID decision rule is a tight upper bound on the original Bayes error rate, i.e., $P_e^* \leq P_e^{IID}$. The IID approximation breaks down (i.e., it is not close to the true Bayes error) for patterns with periodic internal structure (for example, the pattern $ABABABAB$). In such cases the optimal decision depends on symbols outside the L-context—we study this effect in more detail in section 5.

In Appendix 1 we provide the derivation of a closed-form expression for P_e^{IID}. However, the resulting expression is rather complex to evaluate and interpret, particularly for large alphabet sizes. In order to obtain a simpler (but approximate) functional form for the Bayes error rate we can ignore the fact that the state sequence that generates any particular substring of length L can contain both background and pattern states. Instead, we assume that each substring of length L, starting at position i, is generated either by a run of L background states or a run of L pattern states, and make a decision independently at each position i based on this assumption. We call this model *IID-pure* and the associated error rate $\widetilde{P}_e^{IID}$. This approximation is precise everywhere except the vicinity of the pattern. We will see later in the paper that this approximation is accurate as long as the marginal probability of pattern states $F \times L$ is small.

Under the *IID-pure* assumption the data in effect consists of vectors of length L of observed substrings $O = (o_1, \ldots, o_L)$, which are generated by an unknown sequence of hidden states $H = (h_1, \ldots, h_L)$. The hidden state sequence under the *IID-pure* assumption can only take two values: $\mathbf{P} = (P_1, \ldots, P_L)$ with probability F, corresponding to the pattern sequence, or $\mathbf{B} = (B, \ldots, B)$ with probability $(1 - F)$ corresponding to a pure background run. Let $\mathbf{P}$ denote a sequence of pattern states, and $\mathbf{B}$ denote a sequence of pure background.

By definition, the Bayes error rate for this problem is given by

$$\begin{aligned} \widetilde{P}_e^{IID} &= \sum_{O=(o_1,\ldots,o_L)} \min\{p(O|H=\mathbf{P})\ p(H=\mathbf{P}), \quad p(O|H=\mathbf{B})\ p(H=\mathbf{B})\} \\ &= \sum_{O=(o_1,\ldots,o_L)} \min\{\ p(O|H=\mathbf{P})\ F, \quad p(O|H=\mathbf{B})\ (1-F)\ \} \end{aligned}$$

Denote by C_l the set of all L-mers that differ from the consensus pattern in exactly l positions, denote the number of elements in each such set by N_l, and let an individual element of each such set be O_l: thus,

$$\widetilde{P}_e^{IID} = \sum_{l=0}^{L} \sum_{O_l \in C_l} \min\{\ p(O_l|H=\mathbf{P})\ F, \quad p(O_l|H=\mathbf{B})\ (1-F)\ \}$$

The probability that a sequence of pattern states generated any L-mer O_l from C_l is the same for all elements O_l of C_l. Hence, we can group the corresponding terms in the sum and obtain

$$\widetilde{P}_e^{IID} = \sum_{l=0}^{L} N_l\ \min[\ p(O_l|H=\mathbf{P})\ F, \quad p(O_l|H=\mathbf{B})\ (1-F)\ ]$$

The number of L-mers with l substitutions relative to the consensus pattern is $N_l = \binom{L}{l}(n_A-1)^l$. The probability of producing such an L-mer starting from the first pattern state is $p(O_l|H=\mathbf{P}) = (1-\varepsilon)^{(L-l)}\left(\frac{\varepsilon}{n_A-1}\right)^l$. All L-mers produced by the background have equal probability, $p(O_l|H=\mathbf{B}) = \left(\frac{1}{n_A}\right)^L$. Finally, we obtain

$$\widetilde{P}_e^{IID} = \sum_{l=0}^{L} \binom{L}{l} (n_A-1)^l \min\left\{(1-\varepsilon)^{(L-l)}\left(\frac{\varepsilon}{n_A-1}\right\}^l F, \quad \left(\frac{1}{n_A}\right)^L (1-F)\right\} \tag{3}$$

where

$$P_e^* \leq P_e^{IID} \approx \widetilde{P}_e^{IID}$$

We look at various interpretations and limiting cases of this expression for $\widetilde{P}_e^{IID}$ later in this section. In general, the Bayes error varies between zero and the probability of the minority class, which for our purposes will be the pattern frequency F. In practice, it is useful to bring problems with different pattern frequencies onto the same scale. This can be done by considering the normalized Bayes error rate: the fraction of *all patterns* that are misclassified rather than the fraction of *all symbols*, i.e., $P_{Ne}^* = \frac{P_e^*}{F}$. This normalized Bayes error varies between 0 and 1, where the value 1 corresponds to the decision rule of always classifying each symbol as the background.

Note that one can extend the above analysis in a straightforward manner to the Bayes risk for loss functions other than classification, e.g., when false alarms and missed detections for patterns have different associated costs.

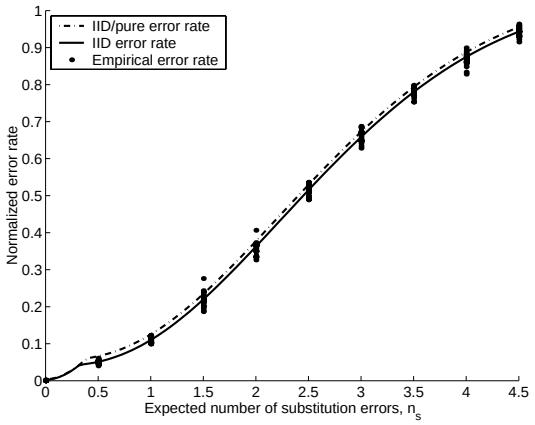

Figure 2: Analytical and empirical estimates of the normalized probability of error as the symbol substitution probability varies, $L = 10$, $F = 0.01$.

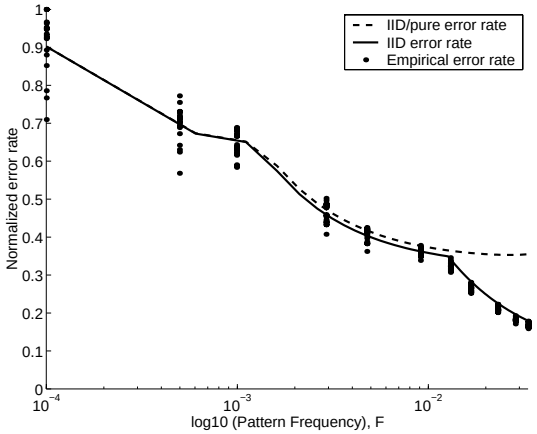

Figure 3: Analytical and empirical estimates of the normalized probability of error as the pattern frequency varies, $L = 10$, $n_s = 1$.

Figure 2 shows how both empirical and analytical estimates of the normalized probability of error change as we vary the expected number of substitution errors. The dots correspond to empirical evaluations of the Bayes error on different sequences of length 10^5, and the dotted and solid lines plotted next to each other correspond to the analytical approximation under the IID and IID-pure assumptions respectively.

Figure 3 shows the normalized Bayes error as we vary the pattern frequency while the pattern length and substitution probability are fixed. The solid line that corresponds to the analytical approximation correctly captures the non-linearity of the dependence on F. The "switching" that is seen on these plots occurs whenever substrings with one more substitution relative to the consensus pattern become recognized as patterns. We also see that the IID-pure approximation (dotted line) starts to deviate from the empirical results only when the marginal pattern probability $F \times L$ increases above 0.1, consistent with the theory, since this is where we expect the assumptions to break down.

Figures 2 and 3 clearly demonstrate that both of the analytical approximations, P_e^{IID} and $\tilde{P}_e^{IID}$, are accurate over a wide range of parameter values.

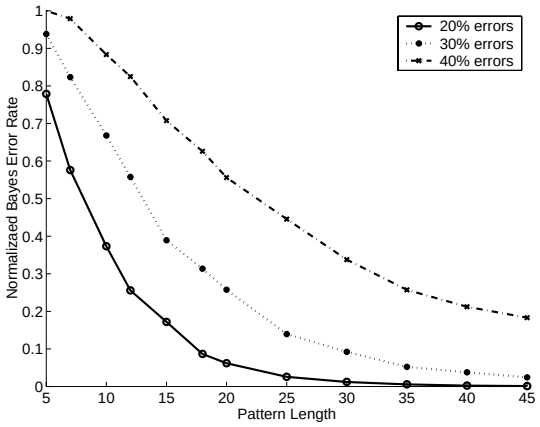

Figure 4: The normalized Bayes error rate for three subsets of problems; each subset is characterized by exactly the same expected percentage of substitution errors.

Other heuristic measures have also been proposed in computational biology to characterize the difficulty of particular pattern-discovery problems. For example, Sze et al. (2002) use a metric defined as the expected percentage of symbols with substitution errors in a pattern ("the error percentage metric"). While this metric works for patterns of a given fixed length, there are cases when it fails to differentiate between easy and hard problems. For example, consider a set of problems defined over the same alphabet and all having the same substitution probability ε and pattern frequency F. We can define subsets of these problems such that in each subset the error percentage metric remains constant but the lengths of the patterns can vary within the subsets. We illustrate in Figure 4 that for such subsets (each of the three lines corresponds to a different subset) the normalized error P^*_{Ne} can vary between 0.2 and 1 while the error percentage metric remains the same.

4.3 Insights Provided by the Analytical Expression

The analytical expression of Equation 3 shows the functional dependence of the Bayes error rate on the parameters of the model, and allows one to gain insight into the behavior of the Bayes error as we vary different parameters of the problem. In particular, if we "plug in" typical parameters for motif discovery problems in computational biology, the Bayes error can be quite high.

In this section, we will use a set of typical parameters from motif discovery applications to illustrate qualitatively the general difficulty of the pattern discovery problem. Suppose we are trying to discover patterns (motifs) of length $L = 5$ in a DNA sequence with alphabet size $n_A = 4$. Assume that patterns appear with probability $F = 0.005$. We also assume wherever needed that the average number of substitutions per pattern is 1. Equation 3 allows one to address the following questions directly:

- How much error is associated with the optimal recognition procedure if the probability of substitutions ε (the symbol noise in the pattern) goes to zero? The limit for the normalized Bayes error can be written as

$$P^*_{Ne} \to \min\left[1, \frac{1-F}{F}\left(\frac{1}{n_A}\right)^L\right].$$

  Plugging in the specific values for our hypothetical problem, even the optimal detection algorithm will incorrectly misclassify on the order of 20% of all patterns even if zero substitutions

are present in the pattern model. Naturally, allowing substitutions can only increase this error rate.

- Given the pattern length and pattern frequency, what is the substitution probability ε such that the optimal procedure misses all of the patterns and classifies them as background? All patterns will be misclassified as background if

$$\varepsilon > 1 - \left(\frac{1-F}{F}\right)^{\frac{1}{L}} \frac{1}{n_A}.$$

  In our hypothetical problem this corresponds to a substitution probability of $\varepsilon = 0.28$, or equivalently, the average number of substitutions is greater than $n_s = 1.39$. In this case the optimal procedure will miss *all* of the pattern symbols, and classify them all as background.

- Given a particular L and ε, what is the value of the pattern frequency such that the optimal procedure classifies all of the pattern symbols as background? All occurrences of a pattern will be classified as the background if

$$F < \frac{1}{\left(n_A\ (1-\varepsilon)\right)^L + 1}.$$

  In our example, if the pattern frequency is less than 3 in a thousand, the optimal procedure misses all the patterns, and classifies them as background (with $\varepsilon = 0.28$).

- If we fix the pattern frequency, pattern length and probability of substitution, we can find the maximum number of substitutions k^*, such that substrings with up to k^* substitutions are recognized as patterns, and all others are classified as background. The number k^* is given by the following expression:

$$k^* = \left\lfloor \frac{\ln(1-F) - \ln(F) - L\ln\left((1-\varepsilon)n_A\right)}{\ln\left(\frac{\varepsilon}{1-\varepsilon}\frac{1}{n_A - 1}\right)} \right\rfloor.$$

  In the hypothetical example above, occurrences of the consensus pattern with even a *single* substitution error will always be classified as background by the optimal procedure.

- What is the expected false positive/false negative rate of the optimal procedure? This quantity can be computed using k^* from the calculations above. In our toy problem, the normalized Bayes error rate is equal to 0.87. Approximately 77% of these errors will be made due to false negatives, and 23% due to false positives, i.e., the optimal detection procedure is going to miss about $77\% \times 0.87 \approx 67\%$ of the patterns even when the true model is known.

To extend the model to handle arbitrary insertions and deletions between consecutive pattern positions, one can introduce an additional $(L-1)$ insertion states and L deletion states "between" the original L pattern states. This is a standard way of modeling insertions and deletions in the HMM models that are used to model biological sequences (see Baldi et al., 1994). In the experiment in Figure 5 we fixed the parameters of a particular model and empirically evaluated the Bayes error rate of problems where we vary the probability of insertions from 0.005 to 0.2. The horizontal line on the bottom of the plot indicates the probability of error for the model with no insertions. In both this experiment and related experiments (not shown) we have found that introducing insertions can increase the Bayes error of a problem significantly.

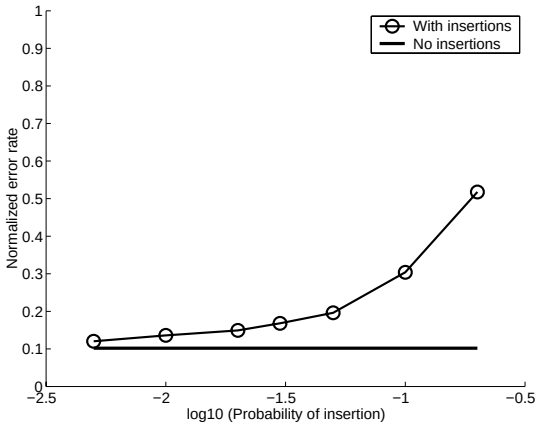

Figure 5: Normalized probability of error as a function of the probability of insertion, $L = 10$, $n_s = 1$, $F = 0.01$.

The Bayes error rate analysis above can be modified to handle multiple consensus patterns (potentially of different lengths, frequencies, etc.) embedded in a common background. Detecting and discovering multiple different motif patterns is an important problem in computational biology (see Bailey and Elkan (1995) and the fragmentation model of Liu et al. (1995)). In the simplest case, we have two patterns that are rather distinct, namely, they are more likely to be confused with the background than with one another. If we denote the model that contains a single pattern P_i with frequency F_i by M_i $(i = 1, 2)$, and the model that contains both P_1 and P_2 with frequencies F_1 and F_2 by M_{12}, then the Bayes error can be shown to satisfy

$$P_e^{M_{12}} = P_e^{M_1} + P_e^{M_2} - (S_1 F_2 + S_2 F_1) P(O_L|B),$$

where $P_e^{M_1}$ and $P_e^{M_2}$ are the error rates of models M_1 and M_2, S_i is the number of distinct strings of length L that would be recognized as pattern P_i by model M_i (a rather small number compared to the overall number of strings of length L), and where $P(O_L|B)$ is the probability of a random L-mer being produced by the background. The second term tends to be relatively small in practice. Thus, if the patterns in M_1 and M_2 are dissimilar then the superposition of the 2 pattern processes leads to a Bayes error which is approximately the sum of the 2 Bayes errors for the individual processes. Appendix 2 contains further details on the derivation of Equation 4.3.

5 The Effect of Pattern Structure

The analytical analysis of the Bayes error rate suggests that the IID assumption does not hold true for patterns with a periodic structure, for example, patterns such as $BCBCBCBCBC$ or $BBBBBBBBBB$. Even though it may seem counter-intuitive, periodic patterns will have a systematically higher Bayes error rate. To quantify exactly the type of pattern structure that violates the IID assumption (and leads to a higher Bayes error), we use the notion of a pattern's autocorrelation vector. The autocorrelation vector is a binary vector that has the same length as the pattern, with the i-th position equal to 1 whenever the $(L-i+1)$-th prefix of the pattern coincides with the $(L-i+1)$-th suffix (see Régnier and Szpankowski (1998) and Pevzner (2000)).

We demonstrate how the Bayes error increases for periodic patterns in Table 1 on a set of four simulated problems. The pattern structure is varied from "random" patterns (zero autocorrelation) to patterns with period 1 (the autocorrelation vector has 1's in all positions). All other parameters

Table 1: Empirical estimates of the normalized Bayes error rate for 4 types of pattern structure: random, period= 3, period= 2, period= 1.

n_s	random	BCCBCCBCCB	BCBCBCBCBC	BBBBBBBBBB
0	0	0.0092	0.0245	0.0636
1	0.1038	0.1478	0.1462	0.1810
2	0.3840	0.3899	0.4134	0.4225
3	0.6539	0.6395	0.6591	0.6598

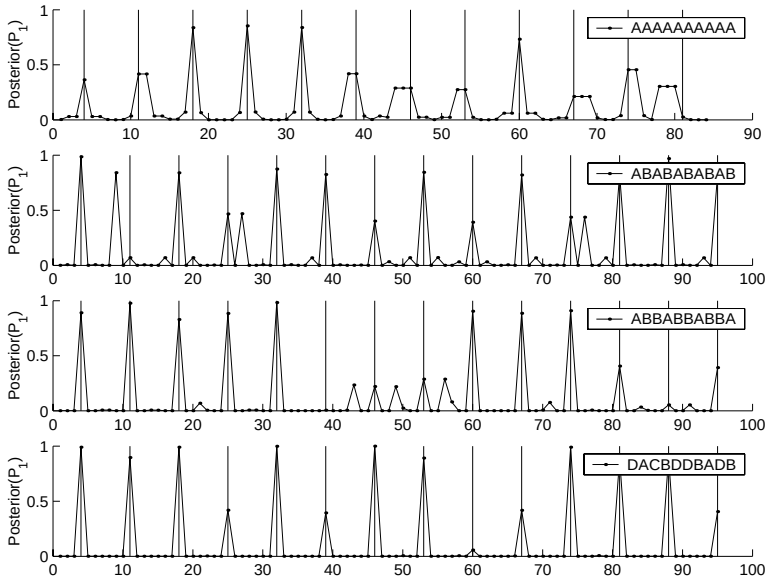

Figure 6: Posterior probability of the first pattern state using the true parameters of the HMM, with $L = 10$, $n_s = 1$, $F = 0.01$. The X-axis represents position in the sequence. Long runs of background have been removed from the plot.

except for the symbols comprising the pattern were the same in each problem: pattern length L, frequency F, and substitution probability ε. From the Table we see that with one expected error (for example) the normalized Bayes error rate for the random pattern is 0.10 and increases to 0.18 for the pattern with the highest autocorrelation (all B's). As more substitution noise is added (higher numbers of expected substitution errors) this noise in effect starts to dominate the Bayes error, and the difference between random and periodic patterns is much smaller (although the overall Bayes error rates are much higher in this case). Thus, in general, the detection of structured patterns in a Markov context presents a more difficult learning problem than the detection of random patterns. These findings provide a direct explanation for the results observed in Helden et al. (1998) where only patterns with clear periodic structure were found to present practical complications.

Intuitively, with unlabeled data, the true boundaries of periodic patterns are harder to determine, and this is precisely what makes learning more difficult. This characteristic behavior for periodic patterns can be visualized in plots of the posterior pattern probability for a given sequence. Figures 6 and 7 illustrate how the posterior probability of the first pattern state changes in the neighborhood of the true (known) boundary of the pattern (as marked by the vertical line) for four different patterns. Only the pattern structure is being varied with respect to the autocorrelation,

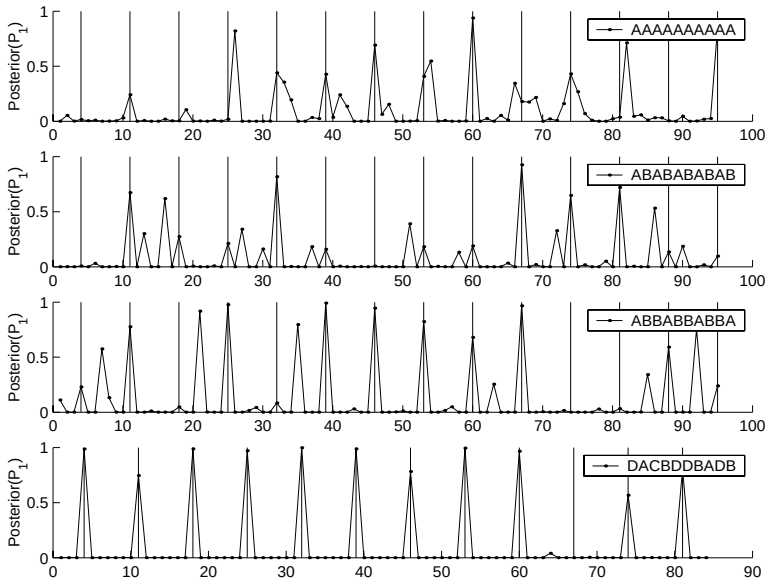

Figure 7: Posterior probability of the first pattern state according to the estimated HMM model, with $L = 10$, $n_s = 1$, $F = 0.01$. The X-axis represents position in the sequence. Long runs of background have been removed from the plot.

across the different plots. The top curve corresponds to a pattern with period 1, and the last one to a pattern with no autocorrelation. Figure 6 was obtained using the *true* parameters of the model, and Figure 7 shows decoding results based on parameters *estimated* by the HMM learning algorithm.

Note that even knowing the true model parameters, it is harder to detect a highly autocorrelated pattern precisely than it is to detect an uncorrelated pattern. For autocorrelated patterns the posterior symbol probabilities produced by the models are further away from 0 and 1, indicating more uncertainty in the model and (generally speaking) a higher associated Bayes error rate.

The difference between aperiodic and periodic patterns is quite dramatic, particularly in the case where the algorithm also must learn the model (Figure 7). In this case, the higher Bayes error rate for the periodic patterns compounds the difficulty of learning; detecting the patterns is harder. For example, we can clearly see from the third plot in Figure 7, corresponding to the pattern $ABBABBABBA$, that the model has completely failed to locate the correct boundaries of the pattern, while in the last plot for the aperiodic (random) pattern $DACBDDBADB$ it has successfully "latched onto" most of the pattern locations.

6 Experimental Results for Specific Pattern Discovery Algorithms

6.1 Three Pattern Discovery Algorithms

In this section we use the Bayes error framework to help us better understand the characteristics of specific learning algorithms and data sets. We analyze the performance of a number of standard motif-discovery algorithms on a set of simulation-based motif-finding problems (similar to the "challenge problems" of Pevzner and Sze (2000)).

Two of the most-widely used algorithms for motif discovery are the Motif sampler proposed by Liu et al. (1995) and the MEME algorithm by Bailey and Elkan (1995). Both use an underlying

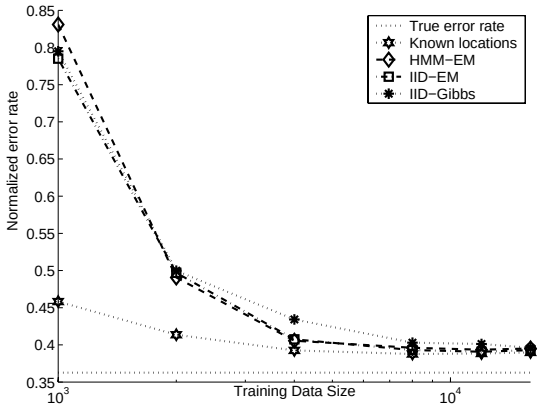

Figure 8: Normalized probability of error for the learned models as a function of training set size with $L = 10$, $F = 0.01$, $n_s = 2$. Each point on the plots was obtained by averaging predictions from models trained on 20 different training sets of the same size. The top 3 curves are the 3 algorithms discussed in the text, the next curve is an HMM trained with labeled pattern locations, and the dotted line at the bottom is the Bayes error rate.

generative model that is similar to the one discussed in this paper. The background is modeled with a single multinomial distribution and the pattern is represented by a product multinomial distribution. Both algorithms assume a simplified IID version of the problem, where the memberships of each substring of length L in the background or pattern are treated as IID variables. This simplification allows for faster inference and learning by ignoring the dependence between consecutive symbols beyond the context of fixed neighborhood of size L.

Learning (or pattern discovery) is, however, carried out differently in these algorithms: MEME uses EM with restarts and clever heuristic initialization techniques to find the unknown parameter values, while the Motif sampler uses Gibbs sampling to estimate the parameters of the model in a Bayesian setting. More recent versions of the algorithms (see Liu et al. (2001) for example) include extensions that allow them to handle multiple occurrences of multiple patterns, higher-order background models, etc.

In our experiments we used the publicly available Motif sampler code described in Liu et al. (1995) and implemented our own version of the EM algorithms for the IID problem that is somewhat similar to the MEME algorithm. We will refer to these as IID-Gibbs and IID-EM algorithms, respectively. In addition, for comparison, we also include the performance of the HMM model (as described in Section 2) trained with the standard EM algorithm for HMMs, to be referred to as the HMM-EM algorithm. The IID-EM and the HMM-EM both use the EM algorithm for pattern discovery, but differ in the nature of their underlying model. The IID-Gibbs algorithm uses the same IID model for the problem as the IID-EM algorithm, but uses stochastic Gibbs sampling to locate patterns and learn their parameters. Appendix 3 describes the three learning algorithms and their associated parameters in detail.

¿From a practical viewpoint it is interesting to note that the different algorithms require significantly different computation times for learning. While the IID-Gibbs can fit a model in seconds for the sizes of learning problems discussed in this section, the HMM-EM algorithm can take several minutes due to the more complex forward-backward calculations.

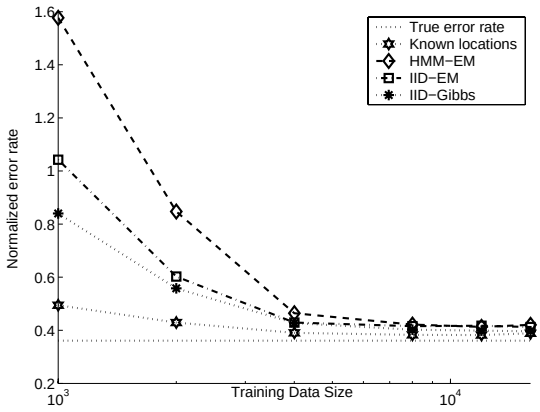

Figure 9: Normalized probability of error for the learned models as a function of training set size with $L = 10$, $F = 0.01$, $n_s = 2$, using a weak prior on pattern frequency. Each point on the plots was obtained by averaging predictions from models trained on 20 different training sets of the same size. The top 3 curves are the 3 algorithms discussed in the text, the next curve is an HMM trained with labeled pattern locations, and the dotted line at the bottom is the Bayes error rate.

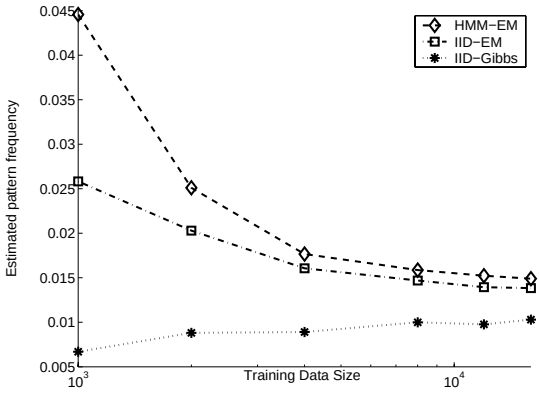

Figure 10: Estimated pattern frequency for the learned models as a function of training set size with $L = 10$ and $n_s = 2$, using a weak prior on pattern frequency. The true pattern frequency F is 0.01. Each point on the plots was obtained by averaging predictions from models trained on 20 different training sets of the same size.

6.2 Experimental Methodology

In the experiments described below we ran each of the IID-EM, IID-Gibbs and HMM-EM algorithms on the same sets of simulated problems. Sometimes, multiple local maxima on the likelihood surface cause these algorithms to find partial or shifted solutions. To avoid these partial solutions, some form of a shifting heuristic is often used, see, for example, Liu et al. (1995); Bailey and Elkan (1995). For the IID-EM and HMM-EM algorithms we shift the predicted pattern locations by 1 to the left or right once the likelihood stops increasing. We then use these locations as the initial conditions for the next round of EM. We continue to shift the solutions and run from the corresponding initial conditions as long as the value of the likelihood at convergence of EM keeps increasing.

Bailey and Elkan (1995) suggested to use the L-mers found in the original sequences to initialize the emissions in the pattern positions, and run EM from these conditions for just a single iteration. The model with the highest value of the likelihood is then trained to convergence of EM. We tried this method with all three algorithms, and while it sometimes improved the running time, it did not improve the mean performance curves of the algorithms. We terminated each run of EM algorithm whenever a relative increase in the likelihood between two consecutive iterations was smaller than 10^{-5}. We allowed multiple random restarts of EM as long as the total number of EM iterations for a given problem did not exceed 2000 (on average, more than 10 random restarts).

Unless stated otherwise we used the default priors of the Motif Sampler algorithm by Liu et al. (1995) in the experiments below. For the pattern frequency prior we used a Beta prior with mean set to the true pattern frequency F and an equivalent sample size $(\alpha + \beta)$ of 4 times the expected number of background symbols in the training data for the strong prior, or alternatively 0.1 times the expected number of background symbols for the weak prior. For the emission probabilities we used a Dirichlet prior with parameters proportional to the overall frequency of symbols in the training data. The equivalent sample size was set to 0.1 times expected number of patterns in the training data. The transition matrix in the pattern was set to a linear structure, i.e., each state has probability 1 of transition to the next state in the pattern and all other transition probabilities were set to 0.

We performed experiments with data simulated from known true HMM models in the class of challenge problems described by Pevzner and Sze (2000). These challenge problems are designed to simulate realistic motif-finding problems in a controlled manner.

The performance of different models was measured by the normalized error rate on relatively long sequences of out-of-sample test data to provide stable estimates of each model's error rate. Note that this error rate can exceed a value of 1 for some of the estimated models, i.e., the error rate of some models is higher than simply assigning all symbols to the background class, which has a normalized error rate of 1 by definition.

6.3 Error Rates as a Function of Training Data Set Size

We studied the effect of varying the size of the training data from 1000 to 16,000 symbols on a given challenge problem. The challenge problem had parameters $L = 10, F = 0.01$, and $n_s = 2$, with consensus pattern $ACBDBBDCAC$. For each fixed training set size, we generated 20 training sequences of the appropriate length from this model and trained each of the 3 algorithms on each training data sequence. For each algorithm, the resulting 20 models were then evaluated in terms of per-symbol classification accuracy on a test sequence of length 10^5 symbols, using the optimal decoding procedure for that model (forward-backward for the HMM, and posterior class probabilities for the other two models). The classification accuracy was then measured relative to the true known class labels (pattern or background) on the test sequences, and averaged over the

20 learned models for each algorithm.

The estimated models can differ from the original true model due to noise (substitution errors) in the actual training patterns and ambiguity in locating the boundaries of the patterns within the training data. The results below characterize the contribution of each of these two factors to the overall algorithm performance.

Figures 8 and 9 illustrate the results of these experiments. On both plots, the lowest curve shows the normalized Bayes error rate of the problem (given by the true model, Model 1)—the lower bound on the probability of error of *any* classifier. The intermediate curve shows the performance of the linear HMM model that was given the *known, true* locations of the patterns (i.e., the supervised learning problem or parameter estimation problem, Model 2). The uppermost curves show the normalized error rate for three different algorithms: IID-EM, HMM-EM, and IID-Gibbs (in effect, three different methods to estimate Model 3).

The plots in Figure 8 were generated using a strong prior on the pattern frequency—in a sense, the true pattern frequency was given to the algorithms. In this case, the learning algorithms all have roughly the same performance as a function of training set size. We see a different behavior in Figure 9 where we specified a correct, but weak prior on the pattern frequency F (weak in the sense of smaller equivalent sample size). In this case, the Gibbs strategy remains about as accurate (as a function of training sample size) as it was for the strong prior.

However, the EM-based algorithms are much less accurate then before, and much less accurate than the IID-Gibbs algorithm, particularly for smaller training set sizes. The EM algorithms greatly overestimate the frequency of pattern occurrences when allowed to deviate from the prior. Figure 10 provides an illustration of this phenomenon. We conjecture that this effect is due to the "batch" nature of updates in EM: the algorithm accumulates the pattern probability from the whole observed sequence before making an update. In contrast, the IDD-Gibbs makes the changes "on-line", for a single occurrence of the pattern at each step.

Furthermore, if an *incorrect* pattern frequency (mismatched to the data) is specified in the prior (i.e., the mean of the Beta prior does not match the true frequency F), the IID-Gibbs algorithm also appears from experiments to be better able to handle this situation, outperforming the EM-based algorithms (results not shown here). Whenever there is an uncertainty about the pattern frequency F, the IID-Gibbs algorithm appears to be more reliable in estimating the models.

6.4 A Component-Wise Breakdown of the Error Rates

The lower two curves in Figures 8 and 9 are worth discussing further. The lowest one is the Bayes error rate (estimated empirically here) and is the best any discovery algorithm can possibly do on this problem. The next curve is the performance of the HMM algorithm where it is told the locations of the patterns (also estimated empirically using a standard supervised HMM training algorithm). The difference between these two lower curves is in effect the contribution to the error rate simply from parameter estimation of the multinomials, i.e., because of small sample sizes, even if a learning algorithm is told where the patterns are, it will still get noisy estimates of the pattern parameters and this contributes to additional error. We can call this contribution to the error the "parameter estimation error."

The three "real" algorithms must of course also discover the parameter locations, and naturally their error curves are systematically higher than that for the "known location" curve. In fact, the distance between a "real algorithm curve" and the "known location" curve can be considered a contribution to error that is coming from not knowing where the patterns are, a "pattern location error."

This allows us to decompose the total error of any algorithm into three additive components:

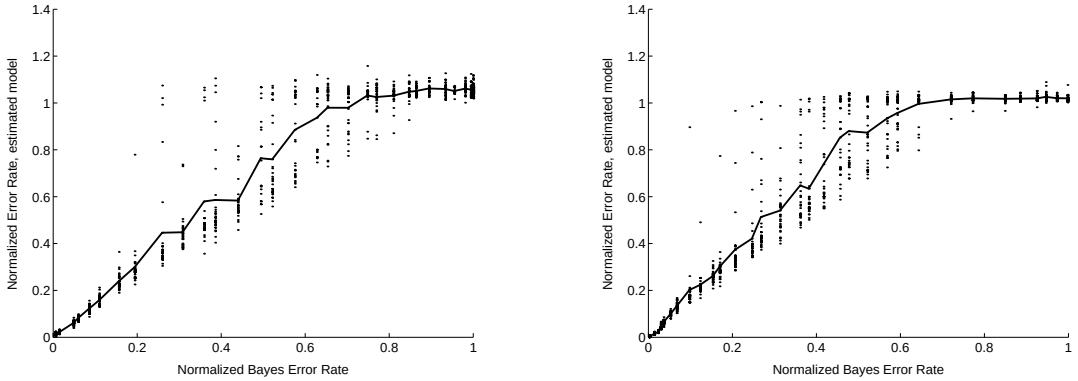

Figure 11: Performance of the IID-Gibbs algorithm on problems with different Bayes error rates. The bold line shows the mean value across all problems with a given Bayes error rate. Here $F = 0.01$, $L = 10$, and (a) $n_A = 6$, (b) $n_A = 10$.

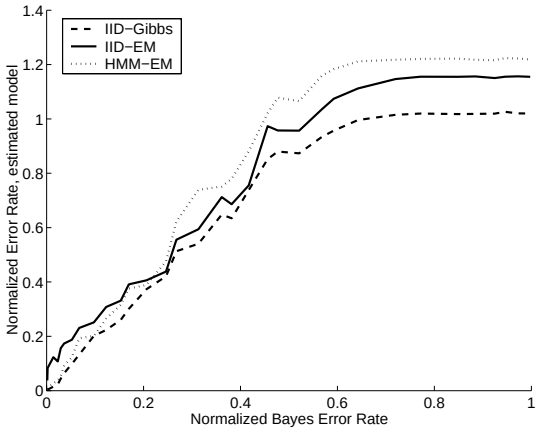

Figure 12: Mean performance of the HMM-EM, the IID-EM and the IID-Gibbs algorithms on problems with different error rates. $f = 0.01$, $L = 10$, $n_A = 10$

(1) the basic Bayes error, (2) additional error due to noise in the parameter estimates (the difference between the Bayes error and "known location" curves) and, (3) further additional error from not knowing where the patterns are located (the difference between the "known location" curves and the real algorithm curves). This tells us, for example, that we can only expect a pattern discovery algorithm to have high accuracy in terms of pattern detection if *all three* error terms are close to zero, i.e., (1) the pattern itself is easily detectable even when known (the Bayes error is close to 0), (2) the parameter estimation noise is low (typically implying that we have a large amount of data or prior knowledge), and (3) the algorithm being used can efficiently discover the patterns given that the other two constraints are met.

6.5 Algorithm Performance as a Function of the Bayes Error Rate

We investigated the behavior of each of the three algorithms on a set of simulated challenge problems with varying Bayes error rates. We fixed the pattern frequency ($F = 0.01$) and the pattern length ($L = 10$) to maintain the same total number of pattern symbols across different problems, and varied the probability of substitutions and the alphabet size to generate a large set of problems

with systematically varying Bayes error rates. We also fixed the training set size to 2000 symbols and then generated multiple instances of the training set from every model. This combination of pattern frequency and training set size was deliberately chosen to present a range of relatively difficult problems for the learning algorithms. All three algorithms used the same set of default priors described earlier on pattern frequency and emission probabilities in the background and pattern states.

Figure 11 shows the performance of the IID-Gibbs on these problems (as measured by the estimated out of sample probability of error) for problems with alphabet size equal to 6 (left hand side) and 10 (right hand side). Naturally, we do not see any points below the diagonal, since it represents the lowest achievable error rate for any classifier.

One would expect the problems with larger alphabet size to be slightly more difficult from the learning viewpoint since they involve a larger parameter space. This can be seen from the two plots in Figure 11: the mean out of sample performance (bold line) on the problems with the same Bayes error rate is slightly worse for the problems with alphabet size 10 (on the right). Another important feature on these plots is extremely high variance in performance for problems with significant Bayes error rate. This variance appears to be largely due to the random variation of the training sets rather than any "local maxima" problems with the learning algorithms—for example, allowing the IID-Gibbs more random restarts or more runs from the same initial conditions does not lead to a lower mean performance curve or smaller variance.

On these simulated problems, the IID-Gibbs algorithm showed better average performance than both the IID-EM and HMM-EM algorithms. Figure 12 shows the mean curves of all three algorithms for the set of problems with alphabet size equal to 10. The error rate for each of the three unsupervised algorithms (the y-axis) appears to be an approximately linear function of the Bayes error (the x-axis) in the range of Bayes error from 0 to 0.5; specifically the error rate obtained by the algorithms is roughly twice the Bayes error (note that this relationship only holds for this particular problem set—for example, if the training set sizes were increased we would expect the algorithms to perform closer to the Bayes error). The strong dependence of actual algorithm performance on the true Bayes error rate provides empirical support for the use of the Bayes error as a fundamental characteristic of the hardness of a problem for a given pattern discovery algorithm. For problems with Bayes error above 0.5 the three discovery algorithms cannot recover the patterns at all and asymptote at a normalized error rate of approximately 1.

The scatter plots in Figure 13 describe on a problem-by-problem basis the error rates of the IID-Gibbs algorithm versus the HMM-EM algorithm (plot on the left), and IID-Gibbs algorithm versus the IID-EM algorithm (plot on the right), for the same set of data summarized by mean curves in Figure 12. If the algorithms performed equally well, the points would be centered around the diagonal. The plots are slightly skewed to the right indicating again the general superiority of the IID-Gibbs.

In Figure 13 we used a very strong prior on the pattern frequency, essentially assuming that the pattern frequency is known. If this assumption is relaxed, and a weaker prior is used, we know from the previous results that the IID-EM and the HMM-EM algorithms can greatly overestimate the pattern frequency, and hence build more "fuzzy" models of the pattern, leading to higher error probabilities.

This behavior is supported by the scatter plots in Figure 14 which show the degradation in performance of the IID-EM and the IID-Gibbs algorithms when using a weak prior on the pattern frequency. In these two scatter plots we compare the performance of the two algorithms with known pattern frequency (strong prior) against the performance of the same algorithms when the frequency is not known (weak prior). The IID-Gibbs tolerates the change quite well (its performance is hardly effected by the change in the prior) but the IID-EM algorithm is quite sensitive to the

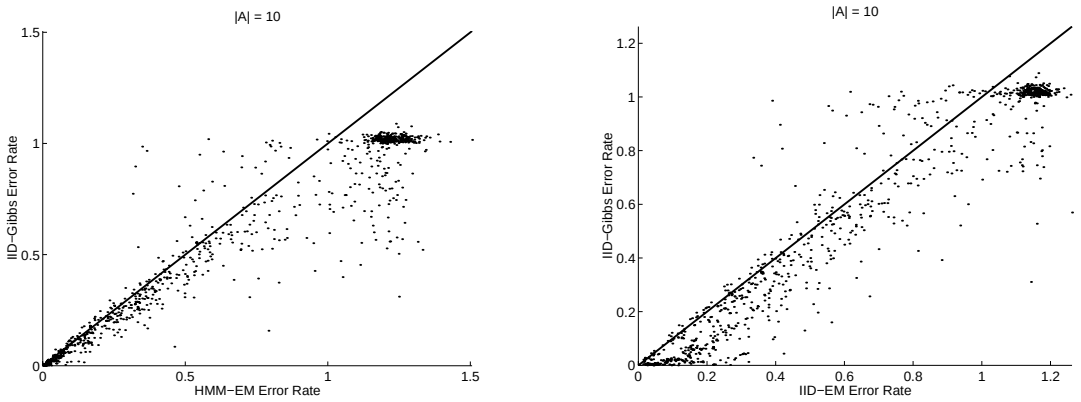

Figure 13: Performance comparison of the HMM-EM and IID-EM algorithms versus the IID-Gibbs algorithm, on problems with different Bayes error rates. Plot (a) compares the HMM-EM algorithm with the IID-Gibbs, plot (b) compares the IID-EM algorithm with the IID-Gibbs algorithm on the set of problems with the following parameters: $F = 0.01$, $L = 10$, $n_A = 10$

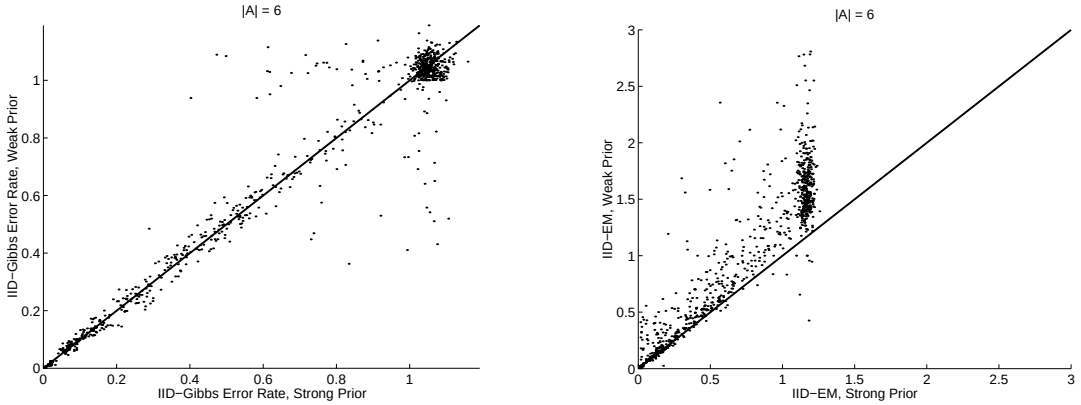

Figure 14: Degradation in performance of the IID-Gibbs and the IID-EM algorithms when the pattern frequency is unknown (weak prior on the pattern frequency). $F = 0.01$, $L = 10$, $n_A = 6$

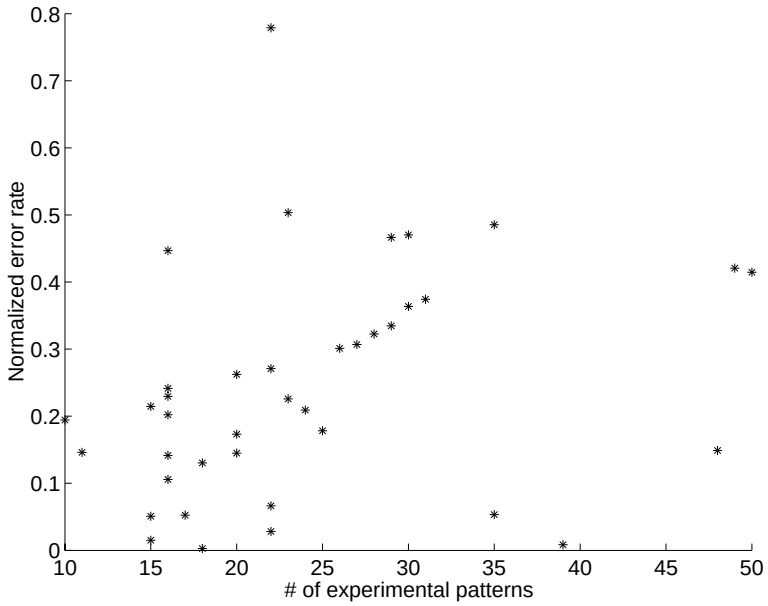

Figure 15: The normalized error rate of learning the binding sites of 42 different *E. coli* DNA-binding proteins.

equivalent sample size. The high values (significantly above 1) of the probability of error also indicate overestimation of the pattern frequency F by the algorithms. The same effect is seen with the HMM-EM algorithm.

7 Finding Real Motifs

Finally we consider an application of the Bayes error rate framework to actual motif-finding problems in DNA sequences. Evaluation of the Bayes error rate requires knowledge of the true data generating model, which is not possible for the real-world problems. However, if the locations of the instances of a given pattern are known from experimental data, then one can consider the supervised version of the problem and construct the corresponding model from these instances (which is what we do here). Given a reasonable number of instances of the pattern, we expect the error rate calculated in this manner to approximate the Bayes error of the true data generating process.

We used data from the DPInteract database which contains experimentally-confirmed instances of binding sites for 55 different *E. coli* DNA-binding protein families (Robison et al., 1998). We constructed the MAP estimates of the parameters of the HMM models from these instances. We used a Dirichlet prior matching the overall letter frequencies with an estimated sample size of 1 to regularize the emission probabilities. Figure 15 shows the empirically evaluated Bayes error rate as a function of the number of given instances that had 10 or more pattern instances per problem (there were 42 such problems, with a total of approximately 800 experimentally verified binding sequences). The normalized Bayes error varies from near 0 to 0.8, independent of the number of training patterns, indicating the presence of significant variability in the difficulty of these discovery problems. Note that the relatively high values of Bayes error are due to high degree of pattern

Table 2: Error metrics for the Motif Sampler, MEME, and CONSENSUS algorithms on known motif problems with different Bayes error rates.

Family	P^*_{Ne}	Motif Sampler	MEME	CONSENSUS
pur	0.05	0.11	0.06	0.06
argR	0.14	0.31	0.52	0.29
crp	0.26	0.53	0.62	0.65

ambiguity rather than the small alphabet size.

Performance of the Motif Sampler, MEME and CONSENSUS (see Stormo and Hartzell (1989)) algorithms on 3 of these problems has been evaluated in Sze et al. (2002). The performance metrics used in this paper are different from classification error, but are correlated. In Table 2 we see that these reported errors increase (non-linearly) with the estimated Bayes error rate, demonstrating how Bayes error directly influences the "learnability" of motif-discovery.

The relatively high Bayes error rates on real motif problems suggests that motif discovery from sequence information *alone* is a hard problem. Note in particular that the Bayes error rate cannot be reduced either by seeking better learning algorithms or by using larger data sets. The Bayes error can, however, in principle be reduced if we provide additional "features" to the motif classifier. This suggests that a profitable future direction for motif-discovery in computational biology is to combine additional information outside of the sequence (such as gene expression measurements, chemical properties or 3-dimensional protein structure) with primary sequence information. See Chen et al. (2001) and Steffen et al. (2002) and references therein for the work in that direction.

Current algorithms have been also used to make predictions about previously unknown binding sites in collections of DNA sequences, see, for example, Sze et al. (2002). We estimated the Bayes error rate based on the published hypothetical matches for each of the two patterns. The normalized error rates in these case were 0.03 for the IRON-FACTOR family, and 0.02 for the PYRO-PURINES family. We use this example to illustrate that analyzing the capabilities of the motif finding algorithms on datasets with *known* motifs is a much simpler task than making predictions about an *unknown* pattern using the same algorithm. While it is straightforward to note that the algorithm is capable of finding a rather complex *known* pattern in a given dataset (i.e., one with a high associated Bayes error rate), confident predictions can only be made about rather simple and well conserved patterns. Current algorithms appear to be able to successfully solve only the simplest of these truly unsupervised problems.

8 Discussion and Conclusions

The analytical expressions for the Bayes error derived earlier in the paper allowed us to investigate the effect of different factors such as pattern frequency and alphabet size on the Bayes error. We then experimentally investigated how the Bayes error is related to actual error rates of well-known pattern discovery algorithms on a set of simulated problems. The size of the training data was found to have a significant effect on algorithm performance (not surprisingly). In the challenge problems investigated the algorithms achieved near-optimal performance when trained on sequences of length 16,000 but missed all the patterns (in effect) when trained on sequences of length 1,000.

While the relative performance of the three algorithms was in generally comparable on the problems investigated, there were some systematic differences. In terms of sensitivity to the priors

on pattern frequency, the IID-Gibbs algorithm was found to be systematically more robust than the EM-based algorithms. On problems where the Bayes error rate was varied, the IID-Gibbs algorithm again had a slight edge in terms of overall mean accuracy over the full range of Bayes error.

We also saw that the actual error rate can be thought of as having three components: (1) fundamental ambiguity in pattern detection as captured by the Bayes error itself, (2) errors arising from sub-optimal detection of pattern locations, and (3) estimation error arising from sample-size effects in parameter estimates for the models. For relatively small (but realistic) training data sizes, each of these three error components can be quite large, with the result that the three algorithms we investigated (IID-Gibbs, IID-EM, and HMM-EM) often yielded performance that was quite far away from the optimal possible performance.

The "three-way decomposition" of the error tells us how much improvement one could possibly get from *any* learning algorithm on this problem. No learning algorithm can overcome the parameter noise problem—this is a data limitation, not an algorithmic one. Thus, although the Bayes error rate is the lowest achievable error rate asymptotically with an infinite amount of training data, the "known location" curve is provably the lowest possible error rate that is achievable given a finite sample size. In other words, the difference between the three actual learning algorithms and the "known location" curve (in Figures 8 and 9) is the largest improvement one can achieve with *any algorithm* on this problem. For small sample sizes the gap is large, but for larger sample sizes the curves tell us that there is little room for improvement over the existing approaches with that much training data. In conclusion, this suggests that to get better results in sequential pattern discovery it may be more effective to try to increase the amount of training data, to improve the quality of prior knowledge (e.g., in the form of priors or model structure), or to lower the Bayes error rate of the problem by adding additional information beyond the sequence, rather than trying to improve the performance of discovery algorithms.

Acknowledgements

The work described in this paper was supported by the National Science Foundation under Grant IRI-9703120.

9 Appendix

Appendix 1: Derivation of the Bayes Error Rate under the IID Assumption

Let us now consider the Bayes error rate of the IID problem. The expression for the P_e^{IID} depends on the dominant letters in each of the pattern positions and involves enumeration of some strings of length L.

In the IID problem we classify the observed strings of length L as patterns and non-patterns as above, but the hidden state sequence may contain both the background and the pattern states - consider, for example, $H = (B, B, B, P_1, P_2, P_3)$ or $H = (P_4, P_5, P_6, B, B, B)$ or $H = (P_6, B, B, B, P_1, P_2)$. We denote the state sequence that contains the first j pattern states by $\mathbf{P_j} = (B, \ldots, B, P_1, \ldots, P_j)$, the one that contains the last j pattern states by $\mathbf{P_{-j}} = (P_{L-j+1}, \ldots, P_L, B, \ldots, B)$, and finally the one that contains both the k last pattern states and j first pattern states by $\mathbf{P_{kj}} = (P_{L-k+1}, \ldots, P_L, B, \ldots, B, P_1, \ldots, P_j)$. Hence, the H variable takes values from the set $dom(H) = \{\mathbf{P}, \mathbf{B}, \mathbf{P_j}, \mathbf{P_{-j}}, \mathbf{P_{kj}} : \ 1 \le k \le (L-1), \ 1 \le j \le (L-1), \ (k+j) \le (L-1)\}$.

The optimal procedure classifies each L-mer as the pattern if the corresponding posterior probability of the pure pattern sequence $p(H = \mathbf{P}|O)$ is greater than the sum of posterior probabilities of all other possible hidden sequences $p(H \in \{dom(H) \setminus \mathbf{P}\}|O) = p(H \neq \mathbf{P}|O)$. We can write the Bayes error rate as

$$P_e^{IID} = \sum_{O=(o_1,\ldots,o_L)} \min [\ p(O|H = \mathbf{P})\ p(H = \mathbf{P}), \quad p(O|H \neq \mathbf{P})\ p(H \neq \mathbf{P})\ ]$$

As before, we can break the overall sum into pieces corresponding to exactly l substitutions, and denote by $O(l)$ any L-mer with exactly l substitutions relative to the consensus pattern, and by C_l the set of such L-mers.

$$P_e^{IID} = \sum_{l=0}^{L} \sum_{O(l) \in C_l} \min [\ p(O(l)|H = \mathbf{P})\ p(H = \mathbf{P}), \quad p(O(l)|H \neq \mathbf{P})\ p(H \neq \mathbf{P})\ ] \tag{4}$$

However, different L-mers from C_l are produced by the non-pure state sequences with different probabilities, so we can no longer collapse the sum over all elements of C_l. This will lead later to enumeration of some L-mers, but first we expand the second term corresponding to non-pattern state sequences. We do this by writing out the corresponding disjunction of H values

$$\begin{aligned} p(O|H \neq \mathbf{P}) \qquad & p(H \neq \mathbf{P}) = \\ & p(O|H = \mathbf{B})\ p(H = \mathbf{B}) + \\ & \sum_{j=1}^{(L-1)} [\ p(O|H = \mathbf{P_j})\ p(H = \mathbf{P_j}) + p(O|H = \mathbf{P_{-j}})\ p(H = \mathbf{P_{-j}})\ ] + \\ & \sum_{1 \le k,j \le (L-1),\ k+j \le (L-1)} [\ p(O|H = \mathbf{P_{kj}})\ p(H = \mathbf{P_{kj}})\ ] \end{aligned} \tag{5}$$

We now compute the prior probabilities of the values that the hidden state sequence H can take. Note that if the pattern frequency is F, then the prior probability of seeing a background symbol is $P(B) = (1 - F\ L)$, and the probability of transition from background to pattern is $p(P_1|B) = \frac{F}{(1.0 - L*F)}$. From this we can compute the prior probability of values that the state sequence H can take:

$$p(H = \mathbf{B}) = p(B)\ (1 - p(P_1|B))^{(L-1)} \tag{6}$$

$$p(H = \mathbf{P}) = F \tag{7}$$

$$p(H = \mathbf{P_j}) = p(B)\ (1 - p(P_1|B))^{(L-j-1)}\ p(P_1|B) \tag{8}$$

$$p(H = \mathbf{P_{-j}}) = F\ (1 - p(P_1|B))^{(L-j-1)} \tag{9}$$

$$p(H = \mathbf{P_{kj}}) = F\ (1 - p(P_1|B))^{(L-(k+j)-1)}\ p(P_1|B) \tag{10}$$

Now we come back to the problem of calculating the likelihood of the L-mers from C_l under different values of the generating hidden state sequence $P(O(l)|H)$. As we mentioned earlier, this requires enumeration of all elements of C_l. The likelihood of the L-mer $O(l)$ under $\mathbf{P_j}$, $\mathbf{P_{-j}}$, and $\mathbf{P_{kj}}$ assumptions would be different depending on the specific composition of $O(l)$. To eliminate the need to enumerate all of the L-mers, we can break down the sum over l in (4) into two pieces: one corresponding to $0 \leq l \leq R^*$, and the other to $R^* < l \leq L$. We can pick R^* as the smallest integer such that the sequences with R^* substitutions have higher joint probability of the sequence and pure background, than the sequence and pure pattern states $P(O(R^*)|H = \mathbf{B}\ p(H = \mathbf{B})) > P(O(R^*)|H = \mathbf{P})\ p(H = \mathbf{P})$. Then for L-mers with more than R^* substitutions the minimum in 4 is reached at the first argument corresponding to the pattern states $H = \mathbf{P}$, and we can write

$$\begin{aligned} P_e^{IID} = \quad & \textstyle\sum_{l=0}^{R^*} \quad \sum_{O(l)\in C_l} \min[\ p(O(l)|H = \mathbf{P})\ p(H = \mathbf{P}), \quad p(O(l)|H \neq \mathbf{P})\ p(H \neq \mathbf{P})\ ] + \\ & \textstyle\sum_{l=R^*+1}^{L} \quad \sum_{O(l)\in C_l} p(O(l)|H = \mathbf{P})\ p(H = \mathbf{P}) \end{aligned}$$

The probabilities $P(O(l)|H = \mathbf{B})$ and $P(O(l)|H = \mathbf{P})$ are the same for all patterns with a given number of substitutions relative to the consensus pattern, so we can collapse the second sum and enumerate only the L-mers that differ from the consensus pattern in less than R^* positions.

$$\begin{aligned} P_e^{IID} = \quad & \textstyle\sum_{l=0}^{R^*} \quad \sum_{O(l)\in C_l} \min[\ p(O(l)|H = \mathbf{P})\ p(H = \mathbf{P}), \quad p(O(l)|H \neq \mathbf{P})\ p(H \neq \mathbf{P})\ ] + \\ & \textstyle\sum_{l=R^*+1}^{L} \quad N_l\ p(O(l)|H = \mathbf{P})\ p(H = \mathbf{P}) \end{aligned} \tag{11}$$

We now enumerate all L-mers with exactly l substitutions by first fixing the positions in which the substitutions occur $i = (i_1, \ldots, i_l)$, and then fixing the values at substituted positions $v = (v_1, \ldots, v_l)$. We will denote by $O(i, v)$ the L-mer that has values v in positions i and matches the consensus pattern in all other positions. The contribution of substrings with l substitutions to the probability of error is given by

$$P_e^{IID}(l) = \sum_{i=(i_1,\ldots,i_l)} \sum_{v=(v_1,\ldots,v_l)} \min\left(p(O(i,v)|H = \mathbf{P})\ F, \quad p(O(i,v)|H \neq \mathbf{P})p(H \neq \mathbf{P})\right)$$

When the number of substitutions l, substituted positions i, and substituted symbols v are fixed, we have for individual likelihoods in the right-hand side of (5):

$$p(O(i,v)|H=\mathbf{B}) = \left(\frac{1}{n_A}\right)^{L} \tag{12}$$

$$p(O(i,v)|H)=\mathbf{P_j}) = \left(\frac{1}{n_A}\right)^{L-j}(1-\varepsilon)^{q(i,v,j)}\left(\frac{\varepsilon}{n_A-1}\right)^{j-q(i,v,j)} \tag{13}$$

$$p(O(i,v)|H=\mathbf{P_{-j}}) = \left(\frac{1}{n_A}\right)^{L-j}(1-\varepsilon)^{r(i,v,j)}\left(\frac{\varepsilon}{n_A-1}\right)^{j-r(i,v,j)} \tag{14}$$

$$p(O(i,v)|H=\mathbf{P_{kj}}) = \left(\frac{1}{n_A}\right)^{L-j-k}(1-\varepsilon)^{s(i,v,k,j)}\left(\frac{\varepsilon}{n_A-1}\right)^{j+k-s(i,v,k,j)} \tag{15}$$

Here $q(i,v,j)$ denotes the number of the coinciding symbols in the j suffix of the L-mer with given substitutions $O(i,v)$ and the j-prefix of the consensus pattern. Similarly, $r(i,v,j)$ denotes the number of the coinciding symbols in the j-th prefix of $O(i,v)$ and the j suffix of the consensus pattern. Finally, $s(i,v,k,j)$ denotes the number of the coinciding symbols in the k-th prefix of $O(i,v)$ and the k suffix of the consensus pattern plus the number of the coinciding symbols in the j-th suffix of $O(i,v)$ and the j-th prefix of the consensus pattern.

Thus, we have obtained all the expressions necessary to evaluate the final equation for the Bayes error (11): the likelihood of the L-mers $p(O(l)|H)$ is given in equations (12) - (15)) and the prior probabilities of the state sequences are given in equations (6) through (10).

The evaluation of the Bayes error under the IID assumption can potentially be computationally expensive, as it requires the enumeration of all the L-mers that can be recognized as the patterns. For a lot of practical problems this set is still manageable and the estimate of the Bayes error can be evaluated in a matter of minutes, however for a large alphabet size or very long and noisy patterns the evaluation of the Bayes error may become infeasible. Note, that in this case the empirical estimates would also require decoding of rather long sequences.

Appendix 2: Extensions to Multiple Patterns

In this section, we derive the expression for the Bayes error rate under the IID/pure assumption when multiple patterns are present in the model. We limit our analysis to the case when the pattern models are rather "distant" from each other (the corresponding consensus patterns are dissimilar and the noise level is such that the probability of mistaking an occurrence of one pattern for another pattern is negligibly small). We also limit the number of patterns to two (multiple mutually-dissimilar patterns can be handled in a similar fashion). Let's denote the model corresponding to pattern C_i with frequency F_i by M_i ($i = 1, 2$) and the model allowing both patterns with frequencies F_1 and F_2 by M_{12}. Here we assume that models M_1, M_2, and M_{12} share the same distribution in the background state. We can split the set of all L-mers into three non-intersecting sets:

- $S(C_i)$ is the set of all L-mers recognized as pattern C_i by M_i and M_{12}, $i = 1, 2$
- $S(B)$ is the set of all L-mers recognized as the background by M_{12}

Again, we assume that the patterns are distant enough to be most often confused with the background, not with one another. This assumption ensures that the three sets above are correctly defined and are non-overlapping. In this section, we denote the probability of the L-mer O being produced by the pattern sequence

We can then write down the $P_e^{M_{12}}$ of the combined model M_{12} as

$$\begin{aligned} P_e^{M_{12}} = & \sum_{O \in S(C_1)} [P(O|B)(1 - F_1 - F_2) + P(O|C_2)F_2] + \\ & \sum_{O \in S(C_2)} [P(O|B)(1 - F_1 - F_2) + P(O|C_1)F_1] + \\ & \sum_{O \in S(B)} [p(O|C_1)F_1 + p(O|C_2)F_2] \end{aligned} \tag{16}$$

Similarly, we can write for the single pattern models M_i, $i = 1, 2$

$$P_e^{M_1} = \sum_{O \in S(C_1)} [P(O|B)(1 - F_1)] + \sum_{O \in S(C_2)} [P(O|C_1)F_1] + \sum_{O \in S(B)} p(O|C_1)F_1 \tag{17}$$

$$P_e^{M_2} = \sum_{O \in S(C_1)} [P(O|C_2)F_2] + \sum_{O \in S(C_2)} [P(O|B)(1 - F_2)] + \sum_{O \in S(B)} p(O|C_2)F_2 \tag{18}$$

By subtracting equations 17 and 18 from equation 16 we obtain

$$P_e^{M_{12}} = P_e^{M_1} + P_e^{M_2} - (F_1|S(C_2)| + F_2|S(C_1)|)P(O|B)$$

Namely, the probability of error of the combined model is equal to the sum of the probabilities of error of the single pattern models minus a small term reflecting the fact that the two patterns look more similar to the background than to the each other.

Appendix 3: The EM Algorithm for the IID/pure model

We assume that each substring of L consecutive letters in the full sequence could have been generated either by the pure background state sequence or by L consecutive pattern states $(P_1, \ldots, P_L)$. We introduce binary hidden variables $\mathbf{C} = (c_1, \ldots, c_n)$ (one for each element of the training sequence), which take value 0 when the sequence $o_i, \ldots, o_{(i+L-1)}$ is produced by the pure background, and 1 when it is produced by the pattern states.

Let Θ denote the full set of model parameters: the pattern frequency F and the multinomial distributions for background and pattern states. We will use the matrix $E \in R^{(L+1)\times n_A}$ to denote the emission probabilities: the first row corresponds to the background distribution, and the other L rows to distributions within the pattern states. We derive below the likelihood of the data under a given set of parameters and the equations for the EM algorithm that estimates the unknown parameters Θ.

The EM algorithm proceeds in the following 2 steps:

- E-step: finding the expected value of the hidden variables given the observed data and the current parameter estimates. This can be implemented as follows:

$$p(c_i = 0|O, \Theta) \propto (1 - F) \prod_{j=1}^{L} E(1, o_{i+j-1}), \qquad 1 \leq i \leq (n - L + 1)$$

$$p(c_i = 0|O, \Theta) = 1, \qquad (n - L + 1) < i \leq n$$

$$p(c_i = 1|O, \Theta) \propto F \prod_{j=1}^{L} E(j + 1, o_{i+j-1}), \qquad 1 \leq i \leq (n - L + 1)$$

$$p(c_i = 1|O, \Theta) = 0, \qquad (n - L + 1) < i \leq n$$

  These two equations allow one to calculate current membership probabilities.

- M-step: find the parameters Θ that maximize the full joint likelihood of the observed and hidden variables given the membership probabilities. The following equations show the parameter updates:

$$\widehat{F} = \frac{\sum_{i=1}^{n} p(c_i = 1)}{(n - L + 1)}$$

$$\widehat{E}(k, m) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{L} p(c_i = j) I(o_i = m)}{\sum_{m=1}^{n_A} \sum_{i=1}^{n} \sum_{j=1}^{L} p(c_i = j) I(o_i = m)}$$

The M-step equations can be easily modified into maximum a posteriori (MAP) estimates of the parameter values by taking into account prior information on pattern frequency and emission probabilities.

The complete data likelihood can be evaluated using the following equation:

$$p(O|\Theta) = \sum_{i=1}^{n} \left[(1-F) \prod_{j=1}^{L} E(1, o_{i+j-1}) + F \prod_{j=1}^{L} E(j+1, o_{i+j-1}) \right]$$

References

Bailey, T. and C. Elkan (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning 21*(1/2), 51–80.

Baldi, P., Y. Chauvin, T. Hunkapillar, and M. McClure (1994). Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Science 91*, 1059–1063.

Buhler, J. and M. Tompa (2001). Finding motifs using random projections. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01)*, pp. 69–76.

Chen, S., A. Gunasekera, X. Zhang, T. Kunkel, R. Ebright, and H. Berman (2001). Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: Alteration of DNA binding specificity through alteration of DNA kinking. *J. Mol. Biol. 314*, 75–82.

Chow, C. K. (1962). A recognition method using neighbor dependence. *IEEE Trans. Elect. Comput. EC-11*, 683–690.

Chu, J. T. (1974). Error bounds for a contextual recognition procedure. *IEEE Trans. Elect. Comput. C-20*.

Duda, R., P. Hart, and D. Stork (2001). *Pattern Classification*. John Wiley and Sons, Inc., New York.

Eddy, S. (1995). Multiple alignment using hidden Markov models. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pp. 114–120.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, New York, NY.

Helden, J., B. Abdre, and J. Collado-Vides (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology 281*, 827–842.

Hu, Y.-H., S. Sandmeyer, and D. Kibler (1999). Detecting motifs from sequences. In *Proc. 16th International Conf. on Machine Learning*, pp. 181–190. Morgan Kaufmann, San Francisco, CA.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton (1993, October). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science 262*, 208–214.

Lee, E. (1974). Bounds and approximations for error probabilities in character recognition. In *Proceedings of the International Conference on Cybernetics and Society*, pp. 324–329.

Liu, J. S., A. Neuwald, and C. E. Lawrence (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association 90*(432), 1156–170.

Liu, X., D. Brutlag, and J. Liu (2001). Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of coexpressed genes. In *Pacific Symposium on Biocomputing*, pp. 127–138.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

Pevzner, P. A. (2000). *Computational Molecular Biology: an Algorithmic Approach*. Cambridge, Massachusetts: The MIT Press.

Pevzner, P. A. and S.-H. Sze (2000). Combinatorial approaches to finding subtle signals in DNA sequences. In *International Conference on Intelligent Systems for Molecular Biology*, pp. 269–278. AAAI Press.

Rabiner, L. R. (1989, February). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–285.

Raviv, J. (1967, October). Decision making in Markov chains applied to the problem of pattern recognition. *IEEE Transactions on Information Theory 3*(4), 536–551.

Régnier, M. and W. Szpankowski (1998). On the approximate pattern occurrences in a text. In *Compression and Complexity of SEQUENCES 1997*, pp. 253–264. IEEE Computer Society. Proceedings SEQUENCE'97, Positano, Italy.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, United Kingdom: Cambridge University Press.

Robison, K., A. McGuire, and G. Church (1998). A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *escherichia coli* k-12 genome. *Journal of Molecular Biology 284*, 241–254.

Steffen, N., S. Murphy, L. Tolleri, G. Hatfield, and R. Lathrop (2002). DNA sequence and structure: Direct and indirect recognition in protein-DNA binding. *Bioinformatics 1*, 1–9.

Stormo, G. and G. Hartzell (1989). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences 86*, 1183–1187.

Sze, S.-H., M. Gelfand, and P. A. Pevzner (2002). Finding weak motifs in DNA sequences. In *Pacific Symposium on Biocomputing*, pp. 235–246.