Pattern Discovery in Sequences under a Markov Assumption

Darya Chudova
Information and Computer Science
University of California, Irvine
CA 92697-3425, USA
dchudova@ics.uci.edu

Padhraic Smyth
Information and Computer Science
University of California, Irvine
CA 92697-3425, USA
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. We investigate the fundamental aspects of this data mining problem that can make discovery "easy" or "hard." We present a general framework for characterizing learning in this context by deriving the Bayes error rate for this problem 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

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 applications include computational biology (DNA, RNA, and protein sequences), telecommunication networks (alarm message sequences), and user modeling (sequences of Web-page requests). An important data mining problem in this context is the unsupervised discovery of recurrent patterns in such sequences, i.e., the detection and discovery of relatively short, relatively rare, possibly noisy, repeated substrings in the data. What makes the problem difficult is that relatively little is known a priori about what these patterns may look like and there are typically a combinatorially-large number of possible patterns that could be present.

SIGKDD '02 Edmonton, Alberta, Canada

One of the simplest models for patterns is the "fixed-length noisy pattern", namely substrings of fixed length L where each position has a dominant symbol, but other "noise" symbols can be substituted as part of a noise process. As an example consider a 4-letter alphabet `A, B, C, D` and a pattern `ADDABB` of length $L = 6$, where the designated symbol in each of the 6 locations may be substituted with one of the other 3 symbols with some noise probability ϵ. Consider the case where multiple noisy copies of this pattern are embedded in a background process of uniformly distributed and independent occurrences of the 4 symbols `A, B, C, D`, e.g.,

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

The actual occurrences of the patterns in this sequence are enclosed in brackets. Note that the second occurrence of the pattern is a slightly noisy version of the basic pattern, with the symbol `C` substituted for `B` in the second-to-last position. 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.

Note that there are 3 different levels of problems here, from easiest to hardest:

1. **Pattern Detection** in a sequence, given a known model for patterns and a detection algorithm.

2. **Parameter Estimation** of the parameters for a pattern model, given a known functional form for the pattern (such as a Markov model), and given known locations of the patterns. This is in effect "supervised learning."

3. **Pattern Discovery** from a sequence, or set of sequences, where the locations and parameters of the pattern are unknown, but a known form for the patterns is assumed. This is the "unsupervised learning" problem and is the problem of primary interest in this paper.

In this paper we focus on characterizing the general nature of a class of sequential pattern discovery problems, both theoretically and empirically. 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. Well-known examples of this approach for multivariate classification problems include the Bayes error rate as a lower bound on the average error rate of any possible classifier for a given set of features (Chow 1957; Duda and Hart, 1973; McLachlan, 1992 (chapter 1 in particular); Ripley, 1996) and the risk minimization framework for upper-bounding test error rates (Vapnik, 1998).

Prior work on the Bayes error rate has led to fundamental and important insights into the nature of classification in multivariate feature spaces. In particular, the Bayes error rate provides a theoretical target (it terms of the *lowest achievable average error rate*) for the performance of *any* classifier on a given problem. The Bayes error rate can only be computed exactly if we know the true conditional densities (or distributions) and class probabilities for the classification problem, e.g., if we assume the classes are multivariate Gaussian. For most practical problems of course the true distributions are not known, but nonetheless, these theoretical results provide fundamental insight into the nature of multivariate classification problems and quantify the role of problem dimensionality, class separation, and so forth (e.g., see Chapter 3.8 in Duda and Hart (1973) for a quantification of how dimensionality affects the Bayes error rate under a Gaussian model).

In this paper we will apply the general Bayes error rate framework to sequential pattern discovery problems. As with prior work on the Bayes error rate, we will need to assume that the data are being generated by a specific type of model in order to compute the Bayes error. In particular, we will use a first-order Markov framework as our base model. If we make an analogy with the role of multivariate Gaussian models for classification, the first-order Markov model can be viewed as much less restrictive in scope (for sequential data) than Gaussians are for the multivariate case. Thus, we will use a Markov model as a useful baseline reference framework for characterizing the pattern discovery problem, focusing on a particular class of Markov patterns that can be modeled by a hidden Markov model with certain constraints.

It is important to understand the nature of learning with this baseline model before we can understand detection and discovery more complex pattern structures. For example, if learning is hard even in the case where the correct functional form of the model is assumed known, it is reasonable to infer that real-world problems (where we may not be able to assume knowledge of the correct form of the pattern) will be even harder.

Among sequential pattern discovery applications, motif-discovery in computational biology is the single application in this general class of problems that has received the most attention in prior work. It is certainly an important motivator for the work we will describe in this paper. Nonetheless our focus will be on understanding a general class of pattern discovery problems in sequences, with a view towards understanding the fundamental nature of this rather challenging unsupervised learning problem.

The primary novel contributions of this paper are as follows:

- We provide an approximate expression for the Bayes error rate for pattern discovery under a Markov assumption, and experimentally demonstrate that the resulting expression closely matches the true Bayes error rate. To our knowledge there has been no previous work on the Bayes error rate for sequential pattern discovery and detection problems.
- We illustrate how different factors such as alphabet size, pattern length, pattern frequency, and pattern autocorrelation can directly affect the Bayes error rate, and in turn, increase or decrease the difficulty of the pattern discovery problem.
- We empirically investigate several well-known algorithms for pattern discovery in a Markov context and demonstrate how far away they are from the Bayes error rate in practice; a significant finding is that these algorithms can be relatively far away from the optimal (Bayes) performance, unless very large training data sets are available.
- We apply these ideas to motif-finding problems in computational biology and demonstrate how the theoretical framework of the Bayes error rate can shed light on the reliability of automated motif-discovery algorithms for real data.

2. A MARKOV MODEL FOR SEQUENTIAL PATTERNS

We use the following notation throughout this paper:

- $|A|$ denotes the size of the observable alphabet (the number of unique symbols that can occur in a sequence).
- L denotes the length of the pattern, for the case of fixed-length patterns.
- The *consensus* pattern is a string of L symbols, where the ith symbol is the most likely symbol to appear in position i in the pattern, $1 \leq i \leq L$.
- ε denotes the 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.
- n_s denotes the expected number of substitutions in each pattern and can be computed from L and ε.
- 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 model patterns in the form of "fixed-length plus noise", i.e., a consensus pattern of fixed length L, with substitutions allowed with probability ε. Although this assumption on pattern structure is quite simple it has nonetheless proven to be very useful as a model for motif discovery in computational biology (see Liu et al. (1995), Pevzner and Sze (2000), and Buhler and Tompa (2001) for benchmark problems and detailed discussions). Motifs can be thought of as relatively short highly-conserved regions in a DNA sequence. 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

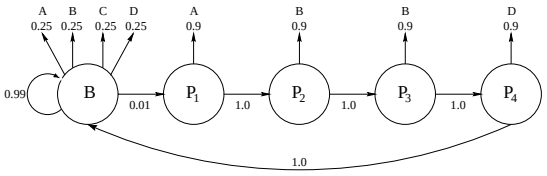

Figure 1: Example of the HMM state transitions for a pattern whose most likely instantiation is $ABBD$ in a uniform background sequence.

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.

To model the embedding of these patterns in a background sequence, we will use hidden Markov models (HMMs), where there is a hidden state for each position in the pattern (L states for patterns of length L), and a background state to model the background. We can think of the HMM as a *generative* process (a method for simulating sequences) for generating patterns of length L, with a particular consensus pattern, a noise level ε, and frequency F. Specifically, we assume a Markov process with $L+1$ states, consisting of a background state B and L pattern states, P_1 to P_L. In the simplest case of fixed-length patterns we assume a strictly "linear" state transition matrix, where the background state B can only transition to either itself or to the first pattern state P_1, where each pattern state P_i can only transition to state P_{i+1}, $1 \leq i \leq L-1$, and where P_L can only transition back to the background state B. This can be viewed as being similar to the product multinomial model for block motifs suggested by Liu et al. (1995). Under this assumption, two occurrences of a pattern can only differ due to substitution errors. We will later relax this assumption to allow for insertion and deletion states.

Given the model above we can generate a Markov state sequence on the $L+1$ states. The state-sequence generated in this manner is not directly observed, hence, the *hidden* nature of the Markov model. We will let h_j denote the value of the hidden state variable at position j, where $h_j \in \{B, P_1, \ldots, P_L\}$. The generative model produces an observed symbol for each hidden state value h_j in the sequence. The symbols are produced according to a state-specific multinomial distribution. For the background state B the multinomial distribution on symbols corresponds to the frequency with which each symbol appears in the background (we use a uniform distribution by default). For the pattern states P_i, $1 \leq i \leq L$, the multinomial probability is $1-(L-1)\varepsilon$ for the consensus symbol in position i and ε for the $L-1$ non-consensus symbols. Thus each pattern state has an output multinomial distribution that is typically "tuned" to a specific symbol for that position.

Figure 1 shows a model for generating patterns of length 4 for a consensus pattern `ABBD`, using a 5-state HMM. The background state is characterized by a high entropy distribution for the background frequencies of symbols in the sequences. Emissions in each of the L pattern states have low entropy with the probability mass concentrated on the consensus symbol in the corresponding position of the pattern. The transition probability from the background to the pattern governs the frequency with which patterns are observed. More generally, patterns could be assumed to be generated by a hidden stochastic finite state machine, i.e., pattern states could have arbitrary transition probabilities with exit transitions to the background state allowing for the generation of variable length patterns. Alternatively, one could include specialized insertion and deletion states, as is usually done in the computational biology for multiple sequence alignment (see, for example, Baldi et al. (1994), Eddy (1995)).

Consider the pattern discovery problem under the HMM assumption above, where it is assumed that the length L of the pattern is known (i.e., the number of states in the HMM) . In this case the pattern discovery problem is reduced to that of learning the parameters of the "correct" HMM for the problem. Even though there exist well-known techniques for fitting such models (such as the expectation-maximization (EM) procedure), we will see that real-world problems are complex enough such that learning the right model is often very difficult. In what follows we characterize in mathematical terms how easy or hard pattern discovery tasks are in the HMM context, by deriving the Bayes error rate for problems of this nature.

3. THE BAYES ERROR RATE

3.1 Motivation

The difficulty of learning a particular pattern can be characterized along multiple dimensions. For example it should be affected by the size of the observable alphabet $|A|$, the length of the pattern L, the variability within the pattern as characterized by the substitution probability 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.

Let $\mathbf{O}$ denote the sequence of observed symbols, and let o_j be the jth elements of the 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 (class 1) or the pattern (class 2), where $h_j = B$ denotes the background and $h_j = P_{1\ldots L} = P_1 \vee P_2 \vee \ldots \vee P_L$ denotes the disjunction of pattern states.

The Bayes error rate is a well-known concept in classification: it is the minimum error rate for a given problem, achieved by using the true model to make the optimal Bayes decision for the class variable given an observed feature vector. If the state sequence were memoryless (independent draws from a distribution on state values, rather than being Markov) then we would have a standard non-sequential classification problem and the Bayes error rate would be defined as:

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

where h is the state value and the sum is over the $|A|$ different values that the symbol o can take in the alphabet. This definition makes intuitive sense: for each value o the optimal Bayes decision is to choose the most likely (maximum

probability) value of B or $P_{1\ldots L}$. The probability of making an error is the probability that the "other" (minimum probability) event actually occurred, averaged over the different possible values of o. For categorical data, the Bayes error will be a function of how "close" the pattern multinomial distribution on symbols is to the background multinomial distribution.

When we re-introduce the Markov dependence, to make an optimal decision on the class at location j, we must now consider not only on the observation o_j but the *whole sequence* of observed values $\mathbf{O}$. If we adopt a directly analogous definition of the Bayes error rate to that defined above, this requires that we sum over the infinite number of possible observation sequences $\mathbf{O}$. This introduces a number of technical issues (such as steady-state probabilities) that are not of direct interest to us in this paper. For this reason we use a more "operational" definition of the Bayes error rate that is in practice asymptotically equivalent to its theoretical cousin, but is both easier to understand and can be directly estimated empirically:

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

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). Intuitively, the classification mistakes occur whenever a background symbol looks more similar to a pattern state than the background state, given the context, or vice-versa. The Bayes error rate in principle indicates how difficult it is (even in the "perfect model knowledge" case) to isolate the occurrences of the pattern from the sea of background, and thus, characterizes the difficulty of the original unsupervised learning (pattern discovery) problem.

3.2 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^*.

There has been little prior work on deriving closed-form expressions for the P_e^* under a Markov assumption. The only related work that we are aware of is by Chu (1970) and Lee (1974), who provide bounds on the Bayes error in a Markov context using *sub-optimal* decision rules. Deriving an expression for the Bayes error rate as a function of the hidden Markov model parameters appears difficult if not impossible for the general case. However, for the special case of a linear transition structure in the HMM (as assumed above), we can use certain simplifying assumptions to derive an approximate expression for P_e^*.

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 L subsequent symbols $o_i, \ldots, o_{i+L-1}$. The problem is simpler because of the reduction of context from the whole observable sequence to the adjacent L characters for each possible position. This simplification is 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 implied by the IID assumption are often very close to the ones obtained from an HMM decoding, and the IID Bayes error rate can serve as a tight upper bound for the true error rate $P_e^{IID} \geq P_e^*$. The only exceptions are patterns with periodic internal structure (consider, for example, pattern $ABABABAB$). In such cases the optimal decision depends on symbols outside the L-context—we study this effect in more detail in section 4.

It is possible to evaluate P_e^{IID} exactly in closed form for any given pattern. However the derivations and corresponding expressions are rather complex and we refer the reader to Chudova and Smyth (2002) for details. We can further simplify the analysis, and obtain close approximations to P_e^{IID}, by ignoring 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 the entire substring is generated either by a run of pure background, or a sequence of L pattern states. We call this model *IID-pure* and the associated error rate $\widetilde{P}_e^{IID}$. As we will see later, the assumption of "pure" state sequences leads to a close approximation of P_e^{IID} as long as the marginal probability of pattern states $F \times L$ is small. We omit the derivation details here due to space limitations (see Chudova and Smyth (2002) for details), and only state that under the IID-pure assumption the Bayes error rate is given by

$$\begin{aligned} P_e^* \quad &\leq P_e^{IID} \approx \widetilde{P}_e^{IID} = \sum_{l=0}^{L} \binom{L}{l} \Big\{ (|A|-1)^l \qquad (3) \\ \times \quad &\min\left[(1-\varepsilon)^{(L-l)} \left(\frac{\varepsilon}{|A|-1}\right)^l F, \quad \left(\frac{1}{|A|}\right)^L (1-F) \right] \Big\} \end{aligned}$$

We look at various interpretation and limiting cases of this expression later in this section. In general, the Bayes error can vary between zero and the probability of the minority class, which for our purposes will be $F \times L$. In practice, it is useful to bring problems with different pattern frequencies and pattern lengths 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*. Naturally, 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.

Experiments have shown that equation 3 approximates closely the true P_e^* across a variety of problems of inter-

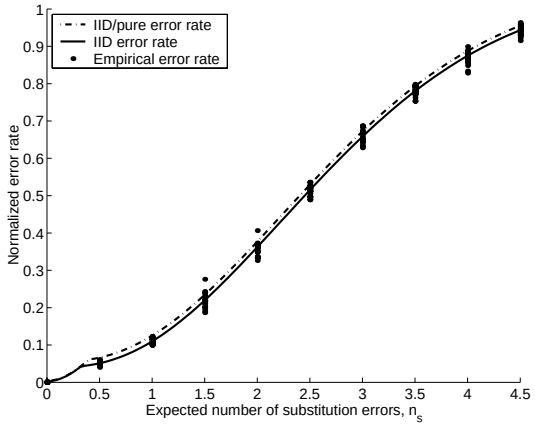

Figure 2: Analytical and empirical estimates of the normalized probability of error as the symbol substitution probability varies, $L = 10$, $F = 0.01$.

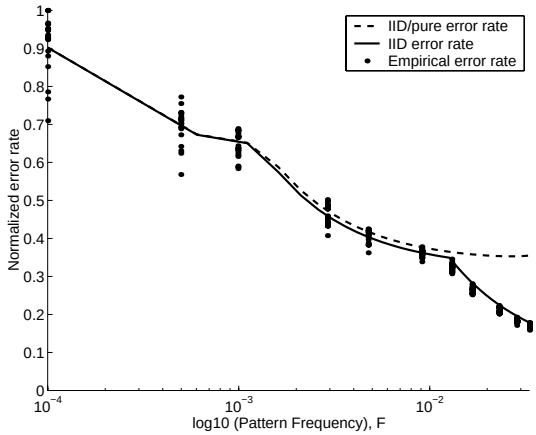

Figure 3: Analytical and empirical estimates of the normalized probability of error as the pattern frequency varies, $L = 10$, $n_s = 1$.

est. 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 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. Note that the empirical estimates become very noisy as the probability of patterns approaches zero, and an accurate empirical evaluation of the Bayes error becomes very expensive computationally due to the lengths of sequences required to get accurate results. Also, the solid line that corresponds to the analytical approximation correctly captures the non-linearity of this dependence. The "switching" that is seen on these plots occurs whenever substrings with one more substitution relative to the consensus pattern become recognized as the patterns. We also see that the IID-pure approximation (dotted line) starts to deviate from the empirical results only when the marginal pattern

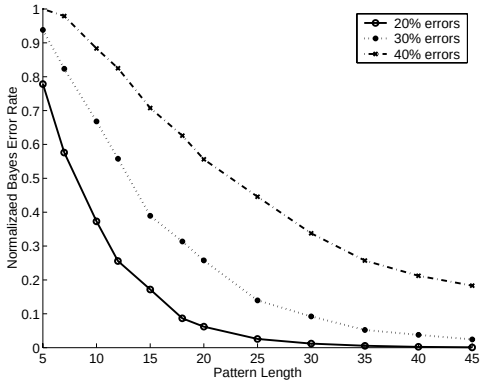

Figure 4: The normalized Bayes error rate for three sets of problems; each set is characterized by exactly the same expected percentage of substitution errors.

probability $F \times L$ increases above 0.1, which is consistent with the theory since this is where we would expect this approximation to break down.

Other heuristic measures have previously been used in computational biology to characterize the difficulty of a particular motif-discovery problem. For example, by Sze, Gelfand and Pevzner (2002) use the expected percent of errors in the pattern. However, there are cases when such a metric fails to differentiate between easy and hard problems. For example, consider a set of problems defined over the same alphabet and having the same substitution probability and pattern frequency. Suppose that problems within the set have different pattern lengths; then the expected percentage of errors for the problems in the set is constant. We illustrate in Figure 4 that for such sets (each of the three lines corresponds to a different set) the normalized P_e^* can vary between 0.2 and 1 while the metric of Sze, Gelfand and Pevzver (2002) remains the same. Equation 3 provides a principled way of combining different parameters of the problems into a single summary characteristic of the problem's difficulty level.

3.3 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 the parameters of the problem. In this section, we will use a realistic example from the biological domain to illustrate qualitatively the difficulty of the pattern discovery problem. Suppose we are trying to discover patterns (motifs) of length $L = 5$ in a DNA sequence (hence, the alphabet size is $|A| = 4$) and the patterns appear with probability $F \times L = 0.025$. 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 (the symbol noise in the pattern) goes to zero? The normalized Bayes error approaches

$$P_e^N \rightarrow \min\left[1, \frac{1-F}{F}\left(\frac{1}{|A|}\right)^L\right]$$

Plugging in the specific values for our hypothetical problem, even the optimal detection algorithm will incorrectly misclassify on the order of 20% of all pattern symbols even if no 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 recognized as background if

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

In our hypothetical problem this corresponds to a substitution probability of $\varepsilon = 0.28$. Equivalently, if the average number of substitutions is greater than $n_s = 1.39$, the optimal procedure will miss *all* of the patterns, 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 misses all of the patterns and classifies them as background? All occurrences of a pattern will be classified as the background if

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

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:

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

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

- 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 23% of these errors will be made due to false negatives, and 77% due to false positives, i.e., the optimal detection procedure is going to overestimate substantially the number of patterns even when the true model is known. This dominance of false positives in the Bayes error expression is typical for these problems, and provides a theoretical explanation for the observed empirical fact that pattern detection algorithms for biological sequences tend to systematically suffer from high false positive rates (see Robison et al., 1998).

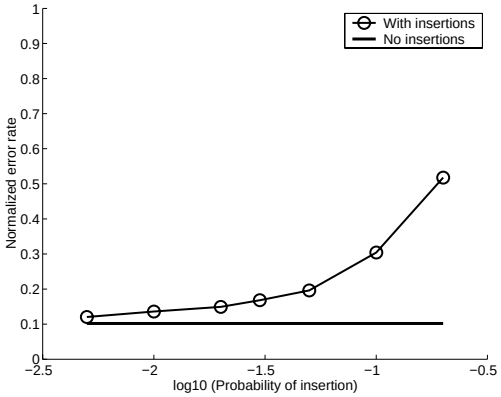

Figure 5: Normalized probability of error as a function of the probability of insertion, $L = 10$, $n_s = 1$, $F = 0.01$.

To extend the model to handle arbitrary insertions and deletions between consecutive pattern positions, one can introduce an additional $(L-1)$ insertion states and $(L-1)$ 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 (e.g., Baldi et al (1994)). In the experiment in Figure 5 we fixed the parameters of the model and empirically evaluated the Bayes error rate of problems with the probability of insertions ranging 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.

The analysis above can be modified to handle risk functions other than the $0-1$ loss that we are implicitly using in counting classification errors. In addition we can represent the presence of multiple consensus patterns (potentially of different length, frequency, etc.) embedded into a common background, and extend the analysis of the Bayes error to handle multiple patterns. 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 satisfies

$$P_e^{M_{12}} = P_e^{M_1} + P_e^{M_2} - (S_1F_2 + S_2F_1)P(O_L|B)$$

where 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 $P(O_L|B)$ is the probability of a random L-mer being produced by the background. See Chudova and Smyth (2002) for more general cases and exact derivations.

4. 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

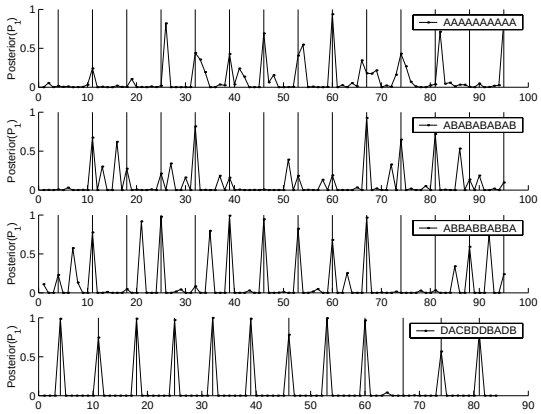

Figure 6: Posterior probability of the first pattern state according to the estimated HMM model, with $L = 10$, $n_s = 1$, $F = 0.01$. The X-axis represents position in the sequence.

higher Bayes error rate (sometimes by an order of magnitude). 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 one whenever the $(L-i+1)$-th prefix of the pattern coincides with the $(L-i+1)$-th suffix. Autocorrelations have been studied extensively in the literature in the past 20 years (see Regnier 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 varied from "random" patterns (zero autocorrelation vector) to patterns with period 1 (the autocorrelation vector has 1's in all positions). All other parameters except for the symbols comprising the pattern were the same in each problem: pattern length, frequency and substitution probability. From the table we see that with 1 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 number 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 Van 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. Figure 6 illustrates 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 (note that long runs of the background have been cut-out from this plot). Here the HMM algorithm is learning the parameters and locations of the patterns. The difference between random and structured patterns is quite dramatic. The algorithm can discover the boundaries of the random pattern much better than for the non-random ones (and the corresponding Bayes error rates are quite different). Fortunately, patterns with significant autocorrelation structure are relatively rare both in motif discovery and in randomly generated patterns. Nonetheless it is useful to be aware of the potential problems associated with discovering such patterns.

5. 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 well-known motif-discovery algorithms on a set of simulation-based motif-finding problems (similar to the "challenge problems" of Pevzner and Sze (2000)). The Bayes error rate provides us with a performance bound for *any* algorithm on these problems and our analysis demonstrates the contribution to learning error that is introduced by two separate effects: (a) not knowing the positions of the patterns in the sequence, and (b) not knowing the parameters of the true data-generating model, even when the positions are known.

Two of the most-widely used algorithms for motif discovery are the Motif sampler algorithm proposed by Liu et al. (1995) and the MEME algorithm by Bailey and Elkan (1995). Both use an underlying 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 the consecutive letters beyond the context of fixed length 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 algorithm that is quite similar to the MEME algorithm. We will refer to these as IID Gibbs and IID EM, 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 Linear HMM algorithm. IID EM and the Linear HMM both use the EM algorithm for pattern discovery, but differ in the nature of their underlying model. The IID Gibbs uses the same IID model for the problem as IID EM, but uses Gibbs sampling to locate patterns and learn their parameters rather than EM.

In the experiments below we ran each of the IID EM, IID Gibbs and Linear HMM algorithms on the same sets of

Table 1: Empirical estimates of the normalized Bayes error rate for 4 types of pattern structure: random, period= 3 ($BCCBCCBCCB$), period= 2 ($BCBCBCBCBC$), period= 1 ($BBBBBBBBBB$).

E[# errors]	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

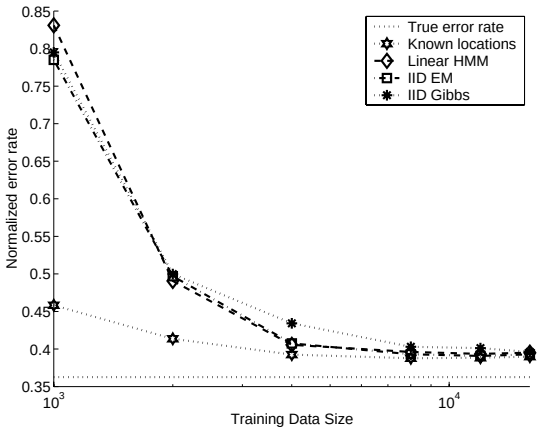

Figure 7: Normalized probability of error for the learned models as a function of training set size with $L = 10$, $F = 0.01$, $n_s = 2$.

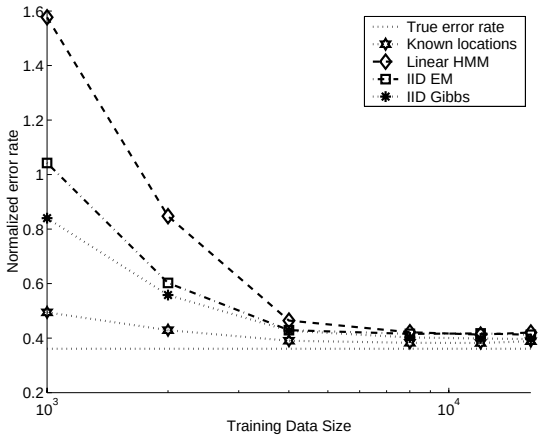

Figure 8: Normalized probability of error for the learned models as a function of training set size with $L = 10$, $F = 0.01$, $n_s = 2$, using a weak prior on pattern frequency.

simulated problems. We used the same shifting heuristic for restarting the EM algorithm between runs for both IID EM and HMM to avoid partial solutions (see Liu et al. (1995), Bailey and Elkan (1995) and Chudova and Smyth (2002) for details).

We performed experiments with data simulated from known true HMM models in the class of challenge problems from Pevzner and Sze (2000). The performance of the models was measured by the normalized error rate on relatively long sequences of unseen data to provide stable estimates of the error rate. Note that this error rate can exceed a value of 1 for 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.

We studied the effect of increasing the size of the training data for problems with different Bayes error rates. The length of the training sequence was varied from $1K$ to $16K$ symbols. For a given training set size, the estimated model can differ from the original true model due to (1) noise in the actual training patterns, and (2) ambiguity in locating the boundaries of the patterns within the training data. The results below allow us to explore the contribution of each of these two factors to the overall algorithm performance.

Figures 7 and 8 illustrate the results of these experiments. Each point on the plots was obtained by averaging over 20 different training sets of the same size for the same true model. On both plots, the lowest curve shows the normalized Bayes error rate of the problem (given by the true model)—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). The uppermost curves show the normalized error rate for three different algorithms: IID EM, Linear HMM, and the IID Gibbs. The plots in Figure 7 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 8 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 significantly outperforms the EM-based algorithms. In fact, the EM algorithms greatly overestimate the frequency of pattern occurrences when allowed to deviate from the prior (we omit the plots here due to the space limitations). We attribute this effect 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 Gibbs sampler makes the changes "on-line", for a single occurrence of the pattern at each step. If an *incorrect* pattern frequency (mismatched to the data) is specified in the prior, the IID Gibbs algorithm seems to be better able to handle this, outperforming the EM-based algorithms (plots not shown here). So, whenever there is an uncertainty about the pattern frequency, the IID Gibbs appears to be more reliable in dis-

covering the patterns.

The lower two curves in Figure 7 are worth discussing further. The lowest one is the Bayes error rate (estimated empirically here) and is the best any discovery algorithm could possibly do on this problem (e.g., using the true model, with an infinite amount of training data). 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 one were told where the patterns are, one would 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 curve is systematically higher than 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 parameters are, a "pattern location error."

This allows us to decompose the total error of any algorithm into three additive components: (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) that the algorithm being used can efficiently discover the patterns given that the other two constraints are met.

The curves for the unsupervised algorithms are quite close to each other across different values of the training set size, indicating that the IID model is quite adequate for non-structured patterns, and that the EM and Gibbs learning algorithms are both equally successful in finding the pattern locations on these simulated problems. From a practical viewpoint it is interesting to note that the algorithms require significantly different computation times for learning. While the IID Gibbs can fit a model in seconds, the HMM can take several minutes due to the more complex forward-backward calculations.

6. 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 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 would expect the error rate to approximate the Bayes error of the true model.

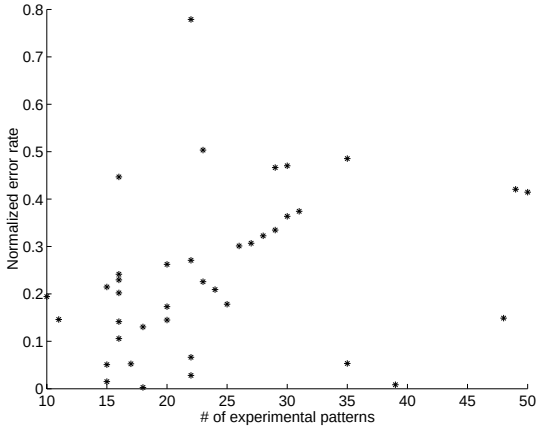

Figure 9: The normalized error rate of learning the binding sites of 42 different *E. coli* DNA-binding proteins.

Robison et al. (1998) reported experimentally-confirmed instances of binding sites for 55 different *E. coli* DNA-binding protein families. 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 9 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 pattern ambiguity rather than the small alphabet size.

Sze, Gelfand and Pevzner (2002) have evaluated the performance of the Motif Sampler, MEME and CONSENSUS (see Stormo (1989)) algorithms on 3 of these problems. The performance metrics they use 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 or protein structure) with sequence information.

7. CONCLUSIONS

Pattern discovery in categorical sequences is an important and challenging data mining problem. In this paper we

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

Family	Normalized Bayes error rate	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

demonstrated the use of the Bayes error rate framework for analyzing problems of this nature in a Markov context. In particular the Bayes error provides insight on how different factors influence the learnability of certain types of patterns. For the particular problem of motif-discovery, we discovered that for both simulated and real data sets the Bayes error rate can be quite high. Directions for future work include multivariate extensions and the design and application of new algorithms for pattern finding.

Acknowledgements

This work was supported by the National Science Foundation under grants IIS-9703120 and IIS-0083489, and by grants from NASA and the Jet Propulsion Laboratory, the National Institute of Standards and Technology, Lawrence Livermore National Laboratory, the UCI Cancer Center, Microsoft Research, and an IBM Faculty Partnership award.

References

Bailey, T. and Elkan C. (1995) Unsupervised learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Machine Learning Journal*, 21, pp. 51–83.

Baldi, P., Chauvin, Y., McClure, M. and Hunkapiller, T. (1994) Hidden Markov Models of Biological Primary Sequence Information. *Proceedings of the National Academy of Science*, 91, pp. 1059–1063.

Buhler, J., and Tompa, M. (2001), *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01)*, Montreal, Canada, 2001, pp.69–76

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

Chu, J. T. (1974) Error Bounds for a Contextual Recognition Procedure. *IEEE Trans. Elect. Comput.*, Vol. C-20, No. 10.

Chudova, D. and Smyth P. (2002) Pattern discovery in sequences under a Markov assumption, Technical Report ICS-TR-02-08, University of California, Irvine.

Duda, R. O., and Hart, P. E. (1973) *Pattern Recognition and Scene Analysis*, New York, NY: John Wiley.

Eddy, S.R. (1995) Multiple Alignment Using Hidden Markov Models. *Intelligent Systems for Molecular Biology*, 3, pp. 114–120.

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

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

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

Liu X, Brutlag D L, Liu J S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput.* 127–138.

Liu, J. S., Neuwald, A.F., and Lawrence, C.E. (1995) Bayesian models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *Journal of the American Statistical Association*, 90, pp. 1156–1170.

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

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

Pevzner, P. A., and Sze, S.-H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'2000)*, pp. 269–278.

Regnier, M. and Szpankowski, W. (1998) On the approximate pattern occurrences in a text. In *Compression and Complexity of Sequences 1997*, pp. 253–264.

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

Robison, K., McGuire, A.M., Church, G.M. (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.

Stormo, G.D., Hartzell, G.W. *Proc. Natl. Acad. Sci.*, 86:1183-1187, 1989

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

Vapnik, V. (1998) *Statistical Learning Theory*, New York, NY: John Wiley.