



Using mixtures of common ancestors for estimating the probabilities of discrete events in biological sequences

Eleazar Eskin¹, William N. Grundy¹ and Yoram Singer²

¹Department of Computer Science, Columbia University, New York, NY, 10027 and

²School of CSE, Hebrew University, Jerusalem, Israel

ABSTRACT

Accurately estimating probabilities from observations is important for probabilistic-based approaches to problems in computational biology. In this paper we present a biologically-motivated method for estimating probability distributions over discrete alphabets from observations using a mixture model of common ancestors. The method is an extension of substitution matrix-based probability estimation methods. In contrast to previous such methods, our method has a simple Bayesian interpretation and has the advantage over Dirichlet mixtures that it is both effective and simple to compute for large alphabets. The method is applied to estimate amino acid probabilities based on observed counts in an alignment and is shown to perform comparably to previous methods. The method is also applied to estimate probability distributions over protein families and improves protein classification accuracy.

Contact: eeskin@cs.columbia.edu

INTRODUCTION

Many successful approaches to modeling sequences for computational biology problems involve statistical models. Many of these problems can be viewed as multi-class classification problems which involve classifying a sequence into one of a set of possible classes. The protein homology problem, for example, can be viewed as the problem of classifying an unknown protein sequence into a protein family. There are two approaches to these kinds of problems. Generative approaches employ a set of generative models, each which is trained over one class of data. An unknown sequence is evaluated by each generative model to determine which class (or protein family) is most likely to have “generated” the protein sequence. These models include approaches such as hidden Markov models (Krogh *et al.*, 1994; Eddy, 1995; Baldi *et al.*, 1994; Karplus *et al.*, 1998), probabilistic suffix trees (Bejerano & Yona, 1999; Apostolico & Bejerano, 2000) and profiles (Gribskov *et al.*, 1990). In contrast, discriminative models are trained over data containing multiple classes

and directly classify an unknown sequence into a class. Discriminative models have been successfully applied to protein homology using support vector machines (Jaakkola *et al.*, 2000) and sparse Markov transducers (Eskin *et al.*, 2000).

Many statistical models, both generative and discriminative, employ a mechanism for estimating a probability distribution from observations. This mechanism estimates a probability distribution over a given alphabet from a set of observed counts of that alphabet. For example, generative protein homology approaches, such as hidden Markov models or profiles, use this mechanism for estimating the probability of observing an amino acid in a certain portion of the model given a set of observations of amino acids at that position. Analogously, some discriminative statistical models such as sparse Markov transducers, use this mechanism for estimating the probability over protein families in a certain portion of the model given a set of observations of protein families.

For estimating over amino acids, many approaches have been presented (see below). Although these approaches were designed for estimating amino acid probabilities, they can be applied to other alphabets such as protein families for use in discriminative models. However, one problem is that the size of the alphabet for protein families is significantly larger ($\approx 2,500$) than the size of the alphabet amino acids (≈ 20). Because of the large alphabet, it is difficult to apply some of the best-performing approaches developed for amino acids. In this paper, we present an efficient robust method to compute these probability distributions. The method performs well on large alphabets. The method uses a mixture of common ancestors and has a straightforward Bayesian interpretation.

Estimation of probabilities of amino acids from observed counts is a very well studied problem, and many methods have been proposed (Durbin *et al.*, 1998). The simplest method to estimate probabilities from counts is the maximum likelihood estimate. Although this method performs well when there is an abundance of data, it is problematic when there is very little data relative to the

alphabet size. For this reason, we want to incorporate prior information into the estimation. Intuitively, in instances where we have few observed counts, we want to rely on the prior information more than in instances where we have more observed counts.

The simplest way to incorporate prior information into the probability estimation is to use the pseudo-count method. In this method, a vector of “virtual” counts is added to the observed counts of amino acids. Although this method is more robust than the simple maximum likelihood estimate when there is a small amount of data, a single pseudo-count vector cannot encode the relations between symbols in the alphabet, such as groupings of amino acids or related protein families. These groupings are important because amino acids tend to appear in groups that share similar biochemical properties such as the hydrophobic group. The presence of one amino acid in the group increases the likelihood of seeing other amino acids from the same group.

A method that addresses these problems is mixtures of Dirichlet distributions (Brown *et al.*, 1995; Sjolander *et al.*, 1996). The expectation and the maximum a posteriori value of a (single-component) random variable from the Dirichlet distribution can be viewed as a smoothing process with a pseudo-count. A mixture of Dirichlet distributions can encode the grouping information by having a component in the mixture for each group. However, in general it is difficult to compute the optimal Dirichlet components (the pseudo-count vectors) from the data. If there are 10 components in the mixture and there are 20 amino acids, then there are about 200 parameters that need to be estimated from the data. The parameters are set by using the EM algorithm to minimize an error function over training data. Since the EM algorithm is subject to local maxima, and there are a lot of parameters, it is very difficult to obtain with confidence the best set of components from the data. In the case of amino acids, the estimation is possible and some very good components have been discovered (Karplus, 1995). However, the computation is much more difficult for large alphabets such as in the case of estimating probability distributions over protein families. In the latest version of the Pfam database, there are close to 2,500 protein families (Sonnhammer *et al.*, 1997). Even with a small number of components, the total number of parameters for Dirichlet mixtures will be very large and will be difficult to optimize.

Another set of methods are based on using substitution matrices (Henikoff & Henikoff, 1996b; Schwartz & Dayhoff, 1978). Substitution matrices have the advantages that they explicitly encode the relations between amino acids and can be easily computed even for large alphabets. A problem with substitution matrix-based methods is that each amino acid has a fixed substitution probability with respect to each other amino acid. Heuristic approaches

to address these problems use the substitution matrix to set a pseudo-count vector (Tatusov *et al.*, 1994; Claverie, 1994; Henikoff & Henikoff, 1996b). Although these methods perform well in practice, they have little theoretical justification.

Another approach to the problem of a fixed substitution matrix is presented by Gribskov and Veretnik (Gribskov & Veretnik, 1996). The approach estimates amino acid probabilities by making an assumption that they were derived from a common ancestor and uses a substitution matrix to obtain the probability estimates. Gribskov and Veretnik first computed which of a set of substitution matrices fit the observed counts best using the measure of cross entropy. They typically used a set of 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048 PAM matrices. Because they choose one of a set of possible substitution matrices, their model is more flexible.

In this paper we present a mixture model of common ancestors. Our approach builds upon and generalizes the method presented in (Gribskov & Veretnik, 1996) to allow for efficient exact computation of a richer, model. In contrast to previous work on substitution-based methods, our method has a simple Bayesian interpretation and has three advantages over the Gribskov and Veretnik model (Gribskov & Veretnik, 1996). The first is that our model is richer since we consider infinitely many possible matrices instead of 12 choices. Second, we employ a set of priors that can be used to incorporate additional biological information into the model. Finally, we derive an analytical solution for the weight of each ancestor in our mixture as well as the possible mutation rates. This analytical solution can be computed efficiently in time that depends only on the size of the alphabet.

The mixture of common ancestors method is an exact method for computing the probability distribution over a discrete alphabet given a set of observed counts that takes into account as a prior the relations between elements of the alphabet. The method makes an assumption that the observations all were derived from a common ancestor through mutations. Given this assumption, if the common ancestor is known, as well as the mutation rate, we can easily compute the probability distribution. Since this information is not known, we implicitly estimate this information using a mixture technique. The observed counts are used to induce a distribution over the possible ancestors and their rate of mutations.

We present our method in the context of amino acids and later apply the method to compute probability distributions over another larger alphabet (protein families). We present an efficient method for computing the probability estimate. Due to space limitations, we omit the derivations of the efficient method. The derivations are presented in a longer version of the paper available at <http://www.cs.columbia.edu/compbio/mca/>. Using amino acid counts

obtained from an aligned database, we perform experiments comparing our method versus comparison methods. We show the ability of the method to handle large alphabets in experiments using this method as a basis for protein family classification.

PRELIMINARIES

Throughout the paper we use the following notation. We denote by Σ the set of possible observations. For instance, in the case of nucleotides $\Sigma = \{A, C, T, G\}$. A sequence of length t of observations is denoted by x^1, x^2, \dots, x^t where $x^s \in \Sigma$ for $1 \leq s \leq t$. The number of occurrences of $i \in \Sigma$ in x^1, x^2, \dots, x^t is denoted by n_i^t , that is, $n_i^t = |\{s | x^s = i, 1 \leq s \leq t\}|$. Analogously, the number of observations in x^1, x^2, \dots, x^t which are *different* from i is denoted by $\tilde{n}_i^t = t - n_i^t$. For convenience we define for all $i \in \Sigma$, $n_i^0 = \tilde{n}_i^0 = 0$.

A major statistical tool in this paper is a mixture model. A mixture model is a combination of simpler models called base or ground models. Each base model induces a probability estimate over events in Σ . Denote the prediction of the j th model on x^s by $P_j(x^s)$. Each base model is associated with a weight denoted w_j . These weights reflect the importance (or reliability) of each model in the mixture. That is, the higher w_j is the more we “trust” the j model. The mixture weights are updated after each observation as follows,

$$w_j^{t+1} = w_j^t P_j(x^t). \quad (1)$$

In this formulation the mixture weights are *unnormalized*. The prediction given by the mixture model is the weighted sum of its base model constituents, that is,

$$P(x^t) = \frac{\sum_j w_j^t P_j(x^t)}{\sum_j w_j^t}, \quad (2)$$

where j ranges over all base models in the mixture.

Equ. (2) is a direct application of Bayes rule assuming that the models are statistically independent. While this assumption is not true in practice, one can still use Eqs. (1) and (2). The formal properties of the Bayes inference algorithm in agnostic settings, i.e., settings in which the independence assumption does not hold, have been studied extensively in information theory, statistics, and learning theory (see for instance instance (Haussler & Oppel, 1998, 1997; Shtar'kov, 1987; Willems *et al.*, 1995; Freund, 1996) and the references therein). An attractive property of mixture models that we exploit in this paper is the simple adaptation to new examples via Equ. (1). In order to compute the prediction of the model, we must be able to easily compute Equ. (2).

THE COMMON ANCESTOR MODEL

The common ancestor model has strong biological motivations and can be best described in the context of protein homology. In the protein homology problem, the goal is to determine which proteins are derived from a common ancestor. The common ancestor model makes the assumption that at some point in the past, each protein sequence in a family was derived from a common ancestor sequence. That is, at each amino acid position in the sequence, each observed amino acid occurs due to a mutation (or set of mutations) from a common amino acid ancestor.

Determining the common ancestor is important in estimating the probability distributions over amino acids. Since amino acids have different chemical properties, different common ancestors mutate to other amino acids with different probabilities. We represent these probabilities in a mutation matrix. A mutation matrix encodes the probabilities that a common ancestor mutates to another amino acid *given that a mutation occurred*. Each row of the mutation matrix corresponds to a unique common ancestor. Each column of the matrix corresponds to the resulting (mutated) amino acid. We denote an element in the mutation matrix $M_{i,j}$. This element corresponds to the probability of the i th amino acid mutating to the j th amino acid if a mutation occurs. Note that for all i $M_{i,i} = 0$ and $\sum_k M_{i,k} = 1$. In a mutation matrix, the diagonal elements are all zeros because we assume that a mutation occurred. We can address the case of multiple mutations by defining mutations and non-mutations in terms of observations. If we observe the common ancestor, then we define that a mutation did not occur even if the common ancestor mutated to another amino acid and mutated back. Similarly, if we observe a different amino acid, we define this as a mutation regardless of the actual number of mutations that occurred from the original common ancestor. Thus the mutation matrix itself gives the probability of observing an amino acid given its common ancestor. This is a reasonable definition, because the mutation matrices are estimated using observed counts. Using the observed counts we will attempt to determine the common ancestor which determines which row of the matrix to use as the probability distribution.

A second issue we need to address in the probability estimation procedure is how likely a mutation is to occur at a given position in an alignment. The estimates depend on the evolutionary distance of the common ancestor. If there is a very short evolutionary distance between the common ancestor and the observed sequence, then there will be very few mutations. However, if the evolutionary distance is very large, there will be a significantly higher number of mutations. The evolutionary distance defines the probability of mutation. We denote in our model the probability of mutation by α . Likewise, the probability

that a mutation did not occur is $1 - \alpha$.

Assuming that we know the probability of mutation as well as the common ancestor, we can obtain a probability distribution over amino acids. We denote this probability by $P_{\alpha,c}$. For a common ancestor c and a mutation probability α , the probability of observing amino acid i is:

$$\forall i \neq c : P_{\alpha,c}(i) = \alpha M_{c,i} \quad (3)$$

$$i = c : P_{\alpha,c}(i) = 1 - \alpha \quad (4)$$

A mutation matrix can be obtained from a standard substitution matrix. Let $S_{i,j}$ be the i, j element of a standard substitution matrix S , such as the ones from the BLOSUM or PAM families of substitution matrices in a form where each element is a conditional probability (Henikoff & Henikoff, 1996b; Schwartz & Dayhoff, 1978). In our framework, each row in the substitution matrix is a common ancestor with a fixed mutation rate, denoted α_i . With this insight we can now compute the corresponding mutation matrix $M_{i,j}$ as follows. Our framework, described above, implies that $S_{i,i} = 1 - \alpha_i$. Furthermore, for each off-diagonal element in the matrix we have $S_{i,j} = \alpha_i M_{i,j}$. We thus get that M can be obtained from S according to the following transformation,

$$M_{i,j} = \begin{cases} \frac{S_{i,j}}{1-S_{i,i}} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

As we discuss in the next section, the fixed mutation rate, implied by the $S_{i,i}$, is instead replaced with a mixture of potential mutation rates where the mixture weights are automatically estimated by our mixture learning algorithm. Thus, our framework automatically adjusts the weights to the set of probable mutation rates based on the data.

MIXTURES OF COMMON ANCESTORS

The probability distribution over amino acids depends on two factors: the rate of mutation, α , and the common ancestor c . Apriori, we do not know either the common ancestor or the mutation rate. Intuitively, we want to estimate these values from observed data. The framework that we employ is a Bayesian mixture approach.

Informally, we can estimate the mutation by the spread of the amino acids. That is, if the observed amino acid counts are concentrated on a few amino acid, then we can intuitively say that the mutation rate is likely to be low, while if the observed amino acid counts are spread among numerous amino acids, the mutation rate is likely to be high. Similarly, we can attempt to estimate which amino acid is the common ancestor by the counts of the observed amino acids.

Formally, we examine the observed counts of each amino acid in the sequence x^1, \dots, x^t . We use the notation presented earlier and denote the counts at time t of a given amino acid i by n_i^t . If the common ancestor is c we can expect the mutation rate to be low if n_c^t is high relative to $\tilde{n}_c^t = \sum_{i \neq c} n_i^t$.

For each common ancestor c , we maintain several possible mutation rates. We denote the set of its possible mutation rates for ancestor c by $R(c)$. For a given mutation rate $\alpha \in R(c)$, we can use Eqs. (3) and (4) to form the estimates for the possible observations. We build a mixture model as discussed earlier: We associate a weight with each possible mutation rate, $w_{\alpha,c}^t$ and the prediction of the mixture is the weighted sum of the predictions of each model:

$$P_c(x^t) = \frac{\sum_{\alpha \in R(c)} w_{\alpha,c}^t P_{\alpha,c}(x^t)}{\sum_{\alpha \in R(c)} w_{\alpha,c}^t} \quad (6)$$

The weights are updated based on the performance of each model on the observed data. The models that better predict the observed data will have a higher relative weight in the mixture. For each possible ancestor c and mutation rate $\alpha \in R(c)$, we use the weight update described previously, that is,

$$w_{\alpha,c}^{t+1} = w_{\alpha,c}^t P_{\alpha,c}(x^t) \quad (7)$$

In the longer version of the paper we derive an efficient scheme for computing the mixture over all possible mutation rates for the predictions given by Equ. (6). For a common ancestor c , the mixture's predictions, which we get by averaging over all mutation rates, is,

$$x^{t+1} \neq c : P_c(x^{t+1}) = M_{c,i} \frac{\sum_{j \neq c} n_j^t + 1}{\sum_k n_k^t + 2} = M_{c,i} \frac{\tilde{n}_c^t + 1}{t + 2} \quad (8)$$

$$x^{t+1} = c : P_c(x^{t+1}) = \frac{n_c^t + 1}{\sum_k n_k^t + 2} = \frac{n_c^t + 1}{t + 2} \quad (9)$$

Note that due to our definition of n_i^t the equation above is given in terms of the predictions for x^{t+1} , and not x^t . Also note that our definition of n_i^0 and \tilde{n}_i^0 implies that $P_c(x^1) = 1/2 M_{c,x^1}$ when $x^1 \neq c$ and $P_c(x^1) = 1/2$ otherwise. That is, the prior probability, before we see any observation, is such that the probability of mutation happening is equal to the total probability of any of the possible mutation. This prior probability distribution can be modified to reflect prior knowledge, as we describe below.

Eqs. (8) and (9) conform to our intuitions for amino acids. Indeed, we can see that if we have observed many mutations in the observed amino acid counts, the probability of observing an amino acid other than the common ancestor is high. Likewise, if we observed

few mutations in the observed amino acid counts, the probability of observing the common ancestor is high. Thus the spread of the amino acids helps estimate the mutation rate.

We can also incorporate some additional prior biological information into the probability estimate. Notice that the rate of mutation is estimated by the relative number of observed mutations versus observed non-mutations. We can introduce a set of “virtual” counts for each common ancestor. These counts behave in a very similar way to pseudo-counts. For ancestor c we define m_c and \tilde{m}_c to be the virtual counts of non-mutations and mutations, respectively. These counts can be used to bias the predictions given by Eqs. (8) and (9) by relying on previous information, incorporated via the virtual counts, when there are few observed counts. The way these virtual counts are incorporated into the prediction is a simple extension of the derivation of Eqs. (8) and (9) and is also described in the longer version. The estimated distribution over Σ after incorporating the virtual counts is given according to the following:

$$x^{t+1} \neq c : P_c(x^{t+1}) = M_{c,i} \frac{\tilde{m}_c + \tilde{n}_c^t + 1}{m_c + \tilde{m}_c + t + 2} \quad (10)$$

$$x^{t+1} = c : P_c(x^{t+1}) = \frac{m_c + n_c^t + 1}{m_c + \tilde{m}_c + t + 2} . \quad (11)$$

Because we do not know a priori the common ancestor, we again apply the mixture technique and evaluate a weighted sum of models each with a different common ancestor. For each common ancestor, c we have a weight at time t , denoted w_c^t . The weight w_c^t reflects the total weight achieved by averaging over all possible mutation rates $\alpha \in R(c)$. The initial weight is set to be w_c^1 and using the predictions given by Eqs. (10) and (11), we update the weights as follows,

$$w_c^{t+1} = w_c^t P_c(x^t) . \quad (12)$$

The prediction of the mixture at time t , $P(x^t)$, is defined to be:

$$P(x^t) = \frac{\sum_{c \in \Sigma} w_c^t P_c(x^t)}{\sum_{c \in \Sigma} w_c^t} . \quad (13)$$

Since there are only 20 possible candidates for a common ancestor for amino acids ($|\Sigma| = 20$), we can use Eqs. (10) and (11) to compute the prediction of each common ancestor (while taking into account the mixture over the different mutations rates) and obtain the final prediction of the mixture efficiently. We can compute the mixture weights incrementally (online) using Equ. (12).

INCORPORATING DEFAULT MODELS

One advantage of the mixture framework is that we can explicitly incorporate other phenomena into the mixture.

That is, the actual probability distribution over amino acids can be dependent on more than just the evolutionary distance and common ancestor. For instance, the probability distribution can depend on the neighboring amino acids or some other structural constraints that are not taken into account in the model presented. Furthermore, in situations when there are vast amounts of data, we would like to “fall back” to a prediction model that is based solely on the empirical counts, namely, the maximum likelihood model.

In order to incorporate a background model B , we add a single component to the mixture along with a smoothing parameter. Specifically, we add a single component Dirichlet distribution with parameter λ whose prediction on x^{t+1} is,

$$P_B(x^{t+1}) = \frac{n_{x^{t+1}}^t + \lambda}{t + |\Sigma|\lambda} . \quad (14)$$

As the number of observations increase, this component will converge to the maximum likelihood estimate. We denote the initial weight of the background predictor by w_B^1 . Similar to the weight-update of the ancestor models, the weight-update of the background model is $w_B^{t+1} = w_B^t P_B^t(x^t)$.

Adding the background model as a component to the mixture, we get that the prediction of the mixture (Equ. (13)) becomes,

$$P(x^t) = \frac{w_B^t P_B(x^t) + \sum_c w_c^t P_c(x^t)}{w_B^t + \sum_c w_c^t} . \quad (15)$$

We set the prior weight of this component w_B^1 to be relatively low compared to the priors for the common ancestors, since we would like the background model to have a significant impact on the overall prediction of the mixture only after we have observed non-negligible amounts of data.

USING BIOLOGICAL INFORMATION TO SET THE PRIORS

In the common ancestor method, we have several variables that define the prior probabilities. The priors are the initial counts for each common ancestor m_c and \tilde{m}_c . We also have the initial weights for each common ancestor w_c^1 . Finally, we have the initial weight of the pseudo-count predictor w_B^1 and the initial count λ .

We can use biological information to set reasonable values of many of the priors. We assume we are given a substitution matrix. Typically if the substitution matrix was obtained from a certain family of substitution matrices, there are members of the family which have empirically been shown to perform best. For example, in the BLOSUM family, one of the most commonly used matrices is the BLOSUM50 matrix. We want to set m_c and \tilde{m}_c for

each component so that the default predictions will be the predictions of the BLOSUM50 matrix. This can be done by fixing the ratio between the the number of virtual mutations versus conservations. The magnitude of $m_c + \tilde{m}_c$ defines how much weight the priors are given in the prediction. This can be set based on how much difference in evolutionary distances are observed in the data set.

We set the values of w_i^1 equal to the background probability of amino acid i in the data set. The initial weight of the background model w_B^1 should be set low relative to the weights of the components w_i^1 to allow the method to perform well when there is little data. A reasonable value for λ is $\frac{1}{|\Sigma|}$ or 0.05 in the case of amino acids.

ESTIMATING THE PROBABILITIES OF AMINO ACIDS

To evaluate the method presented in this paper, we measure the performance of the mixtures of common ancestors on estimation of amino acid probabilities and compare this to the performance of the Dirichlet mixtures over the same data set.

A framework for evaluating methods for probability estimation for amino acids is presented in (Karplus, 1995). In this section we follow Karplus' notation and experimental setting for evaluating and comparing our mixture model of ancestors with previously studied models. Karplus presents an information theoretic framework for measuring the effectiveness of a method. Different methods for probability estimation are compared by independently estimating probabilities of columns of multiple alignments from the BLOCKS database (Henikoff & Henikoff, 1991, 1996a).

Using the notation from (Karplus, 1995), we denote the total count for each amino acid i in a column t by $F_t(i)$. If we use sequence weights such as those presented in (Henikoff & Henikoff, 1994), $F_t(i)$ are not necessarily integers. The total count for each column is denoted $|F_t| = \sum_i F_t(i)$. Using an entire column, we can estimate the probabilities for each amino acid using the maximum likelihood estimate $\frac{F_t(i)}{|F_t|}$.

In practice, we usually only have access to a small set of observed counts. The goal of the probability estimation method is to obtain estimates over the entire column using only the small set of observed counts. For a small set of observed counts s , we use the method to estimate the probabilities over amino acids $\hat{P}_s(i)$. Intuitively, the better the method, the closer the estimate $\hat{P}_s(i)$ will be to the maximum likelihood estimate over the entire column $\frac{F_t(i)}{|F_t|}$.

A way to measure how close a method estimates this probability is by calculating the encoding cost or conditional entropy. The encoding cost for the sample s

in a column t for a given method with estimate $\hat{P}_s(i)$ is given by:

$$H_s(t) = - \sum_i \frac{F_t(i)}{|F_t|} \log_2 \hat{P}_s(i) . \quad (16)$$

The more accurate the estimate, the lower the encoding cost. The minimum encoding cost is when the method's estimate equals the maximum likelihood estimate $\hat{P}_s(i) = \frac{F_t(i)}{|F_t|}$. We can measure the performance of a method by computing how much higher the encoding cost is above this minimum. This amount was referred to by Karplus as the *excess entropy*. The excess entropy is called the relative entropy or the Kullback-Leibler in information theory (Cover & Thomas, 1991). Since the focus of this section is comparison to the different probabilistic estimators discussed in (Karplus, 1995), we use the notation employed by Karplus.

Karplus examines the expected value of the encoding cost when a sample of size k is chosen. We compute the entropy over all possible samples of size k . This measures the performance of the method after k observations. This gives the following entropy of a column for a given sample size:

$$H_k(t) = \sum_{\text{sample } s, |s|=k} P(s|t) H_s(t) , \quad (17)$$

where $P(s|t)$ is the probability of selecting sample s from column t . By averaging over the entire database we obtain the encoding cost for a method with a given sample size:

$$\begin{aligned} H_k &= \frac{\sum_{\text{column } t} |F_t| H_k(t)}{\sum_{\text{column } t} |F_t|} \\ &= \frac{\sum_{\text{column } t} |F_t| \sum_{\text{sample } s, |s|=k} P(s|t) H_s(t)}{\sum_{\text{column } t} |F_t|} \end{aligned} \quad (18)$$

Karplus presents an efficient method for computing the entropy given in equation (18). By comparing encoding costs between two methods, we can compare the performance of two methods.

We perform our experiments over the same data set that was used in (Karplus, 1995). The data consists of sets of observed counts of amino acids taken from the BLOCKS database (Henikoff & Henikoff, 1991, 1996a). The counts are weighted using a position-specific weighting scheme described in (Henikoff & Henikoff, 1994) with slight variations presented in (Sjolander *et al.*, 1996). The data set was split into disjoint training and test subsets.

For each experiment, we compute equation (18) for a different size of the sample k where $0 \leq k \leq 5$. For each sample size, we compute the excess entropy of the methods over all possible samples drawn from each column. We compare the performance of the mixture of common ancestors (*CA-Mixture*) versus several previously

Table 1. Excess entropy for different probability estimation methods under different sample sizes over BLOCKS database. The encoding costs were computed over the BLOCKS database for CA-Mixture, Zero-1, Zero-0.0481, and Pseudo and were consistent with previously published results. The encoding costs for Dirichlet-S and Dirichlet-K reported are published results (Karplus, 1995). Note that the Dirichlet mixtures were obtained with a parameter search in order to minimize entropy over the dataset, while the mixture of common ancestors results are without parameter optimization.

Sample Size	CA-Mixture	Zero-1	Zero-0.0481	Pseudo	Dirichlet-S	Dirichlet-K
0	0.00633	0.12527	0.12527	0.00610	0.06123	0.00883
1	0.02241	1.07961	0.20482	0.13925	0.05336	0.00115
2	0.05557	1.17080	0.18636	0.13720	0.02402	0.00757
3	0.09525	1.16489	0.16843	0.13097	0.01970	0.01471
4	0.10887	1.13144	0.15311	0.12350	0.02083	0.02740
5	0.11337	1.09164	0.14203	0.11804	0.02455	0.03943

presented methods over the same data set. As baselines for comparison we examine two *zero-offset* methods. These methods are pseudo-count predictors with a fixed initial count for each symbol. We compute the results for a zero-offset method with initial count 1 (*zero-1*) and with a initial count optimized for the dataset .0481 (*zero-0.0481*). We also compute the results of a pseudo-count predictor (*pseudo*) with initial counts optimized for the data set and two Dirichlet mixtures. The first Dirichlet mixtures components were obtained from (Sjolander *et al.*, 1996) (*Dirichlet-S*) and the second set of components (*Dirichlet-K*) were obtained from (Karplus, 1995). All of the methods were optimized with a parameter search for the dataset as described in (Karplus, 1995). This gives the mixture of common ancestors a significant disadvantage because we did not perform a parameter search for the dataset.

The results of these experiments are given in Table 1. As we can see, the mixture of common ancestors performs better than all of the comparison methods except for the Dirichlet mixtures, which were optimized for the data. However, the method presented in this paper required very little optimization in order to achieve these results. Also note that the mixture of common ancestors out performs *Dirichlet-S* for small samples. A more complete set of experimental results as well as the mutation matrix and parameter definition files are available at <http://www.cs.columbia.edu/compbio/mca/>.

EXPERIMENTS WITH PROTEIN FAMILIES

As we have shown, the mixture of common ancestor method presented is comparable in performance to other probability estimation methods over amino acids. However, the main advantage to the mixture of common ancestors is that it can handle large alphabets. We apply the mixture of common ancestors presented in this paper to the estimation of probabilities over protein families.

In the current version of the Pfam database, there are approximately 2,500 protein families (Sonnhammer

et al., 1997). A Dirichlet mixture over this alphabet would contain many parameters making it difficult to optimize the parameters over the data. If we assume that with a 2,500 symbol alphabet, there are 100 components, this corresponds to 250,000 parameters.

The mixture of common ancestors, however, does not require this kind of training because the necessary parameters can easily be computed directly from the data. Assuming that we have many sets of observed counts of Σ we can derive a matrix and prior weights from the data. To derive a mutation matrix we use the method presented in (Henikoff & Henikoff, 1996b) to derive a substitution matrix and then convert it to a mutation matrix using the method presented earlier. We can set the prior component weights w_i^1 to the background probabilities computed over the data.

We apply the mixture of common ancestors to estimating distributions over protein families. We are interested in estimating probability distributions from short subsequences of amino acids that are contained in the protein family and using these distributions for protein classification (Bejerano & Yona, 1999; Apostolico & Bejerano, 2000). Probability distributions over protein families conditional on these short subsequences are used for protein family classification using sparse Markov transducers (SMTs) (Eskin *et al.*, 2000). Because SMTs are beyond the scope of this paper, we present a very simple model for protein family classification using subsequences and show how the method in this paper significantly improves the accuracy of protein classification.

In this classification model, we view a protein sequence as the set of subsequences of amino acids that it contains. These subsequences are obtained by a sliding window. For these experiments, we use subsequences of fixed size six. The size of the subsequences was chosen arbitrarily for the experiments. For each subsequence, we compute a probability distribution over protein families. This probability corresponds to how likely the subsequence is to have originated from that family. These probabilities are

computed from a database of classified protein sequences. A protein is classified into a family by computing the probability distribution over protein families for each subsequence. For each family, we compute a score for a protein by computing the normalized sum of logs of the probabilities of the subsequences in the protein originating from that family. This corresponds to assuming that the subsequences are conditionally independent. The protein is classified by determining which family has the highest score.

The key to classifying proteins under this model is the estimation of the probability distribution over protein families for each subsequence. We estimate the probabilities from the observed counts of the subsequence in the database. We compare the mixture of common ancestors to a simple pseudo-count. As we discussed earlier, because of the large alphabet size, it is difficult to use Dirichlet mixtures on this data.

To create the mixture of common ancestors, we first create a data set from the Pfam database version 6.1. In this database, there are a total of 2727 protein families. Here we present results of experiments over the first 100 families. Complete results are presented on the website. Our data set is created by obtaining all sequences of length six present in these families. For each sequence we obtain counts of how many proteins in each family contain that sequence. This gives us sets of counts of over protein families. There are a total of 1,656,277 distinct sequences of length six amino acids giving a total of 1,656,277 sets of counts. Using this data we create a mutation matrix and set the prior weights as described above. The mutation matrix is of size 100×100 .

We split each family in the Pfam database into a training and testing portion by a ratio of 4 : 1. There are a total of 13,889 protein sequences in the training set and 3,418 sequences in the testing set. Over the training portion, we compute the probabilities over protein families associated with each sequence of six amino acids obtained by a sliding window over the proteins. Using these estimates we classify a protein from the test set as described above. For comparison, we estimate the probabilities using both the mixture of common ancestors and the pseudo-count method with a virtual count of $\frac{1}{100}$.

To evaluate the protein classification we compute the ROC₅₀ score (Gribskov & Robinson, 1996). The ROC₅₀ score is the normalized area under a curve that plots true positives versus false positives up to 50 false positives. Table 2 compares the ROC₅₀ scores for the first ten protein families in the Pfam database using the two methods. The results for the complete Pfam database as well as the protein family mutation matrix and parameter definition files are available at <http://www.cs.columbia.edu/compbio/mca/>.

Table 2. ROC₅₀ scores showing results of protein classification over the Pfam database using mixture of common ancestors and pseudo-count predictors.

Family	Mixture of Common Ancestors	Pseudo-counts
14-3-3	1.000	1.000
2-Hacid_DH	1.000	0.969
2-oxoacid_dh	0.973	0.952
3A	0.956	0.900
3Beta_HSD	.889	.889
3HCDH	1.00	1.000
3HCDH_LN	0.977	0.9333
3_5_exonuclease	0.833	0.700
4A_glucanotrans	1.000	1.000
5-FTHF_cyc-lig	.765	0.667

DISCUSSION

We have presented the mixture of common ancestors and applied it to estimating amino acid probabilities and protein family probabilities from observed counts. The method is effective for large alphabets. The mixture of common ancestors leverages well-understood techniques for estimating substitution matrices from data (Henikoff & Henikoff, 1996b; Schwartz & Dayhoff, 1978). Using a variant of these matrices, mutation matrices, the method explicitly encodes the relationships between the symbols of the alphabet.

The mixture of common ancestors has several advantages over previous methods. Unlike previous substitution matrix-based methods, the mixture of common ancestors has a Bayesian interpretation. The method also performs well with both few and many observations. Unlike Dirichlet mixtures, the mixture of common ancestors parameters can be easily computed from training data. This gives the advantage of being able to handle large alphabets such as protein families. In addition, the method is strongly biologically motivated.

Future work involves investigating what kinds of mutation matrices and which prior weights lead to optimal performance under different alphabets. One direction of future work is to compute mutation matrices directly from the data optimized for this model as opposed to computing them from known substitution matrices.

ACKNOWLEDGEMENTS

WNG is funded by an Award in Bioinformatics from the PhRMA Foundation, and by National Science Foundation grant DBI-0078523. YS would like to thank the Bronfman family for their support. Part of his research was funded by Israeli Ministry of Science, grant number 489/00-1. Thanks to C. Leslie for useful comments.

REFERENCES

- Apostolico, A. & Bejerano, G. (2000). Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space. In *Proceedings of RECOMB2000*.
- Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 1059–1063.
- Bejerano, G. & Yona, G. (1999). Modeling protein families using probabilistic suffix trees. In *Proceedings of RECOMB99*. ACM, pp. 15–24.
- Brown, M., Hughey, R., Krogh, A., Mian, I., Sjolander, K. & Haussler, D. (1995). Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Rawlings, C., (ed.) *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 47–55.
- Claverie, J. M. (1994). Some useful statistical properties of position-weight matrices. *Computers and Chemistry*, **18**, 287–294.
- Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.
- Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge UP.
- Eddy, S. R. (1995). Multiple alignment using hidden Markov models. In Rawlings, C., (ed.) *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 114–120.
- Eskin, E., Grundy, W. N. & Singer, Y. (2000). Protein family classification using sparse markov transducers. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA.
- Freund, Y. (1996). Predicting a binary sequence almost as well as the optimal biased coin. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*.
- Gribkov, M., Lüthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods in Enzymology*, **183**, 146–159.
- Gribkov, M. & Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers and Chemistry*, **20**, 25–33.
- Gribkov, M. & Veretnik, S. (1996). Identification of sequence patterns with profile analysis. *Methods in Enzymology*, **266**, 198–212.
- Haussler, D. & Opper, M. (1997). Mutual information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics*, **25**.
- Haussler, D. & Opper, M. (1998). *The Mathematics of Information Coding, Extraction and Distribution*, chapter Worst case prediction over sequences under log loss. Springer Verlag.
- Henikoff, J. G. & Henikoff, S. (1996a). Blocks database and its applications. *Methods in Enzymology*, **266**.
- Henikoff, J. G. & Henikoff, S. (1996b). Using substitution probabilities to improve position-specific scoring matrices. *Computer Applications in the Biosciences*, **12**, 135–143.
- Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, **19**, 6565–6572.
- Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *Journal of Molecular Biology*, **243**, 574–578.
- Jaakkola, T., Diekhans, M. & Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*.
- Karplus, K. (1995). Regularizers for estimating distributions of amino acids from small samples. *Technical Report UCSC-CRL-95-11*.
- Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
- Krogh, A., Brown, M., Mian, I., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, **235**, 1501–1531.
- Schwartz, R. M. & Dayhoff, M. O. (1978). *Atlas of Protein Sequence and Structure*, chapter Matrices for detecting distant relationships. National Biomedical Research Foundation, Silver Spring, MD, pp. 353–358.
- Shtar'kov, Y. M. (1987). Universal sequential coding of single messages. *Problems of information Transmission (translated from Russian)*, **23**, 175–186.
- Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. & Haussler, D. (1996). Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences*, **12**, 327–345.
- Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
- Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 12091–12095.
- Willems, F., Shtarkov, Y. & Tjalkens, T. (1995). The context tree weighting method: basic properties. *IEEE Transactions on Information Theory*, **41**, 653–664.