Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families[‡]

Michael Brown
Computer Science
University of California
Santa Cruz, CA 95064
mpbrown@cse.ucsc.edu

Richard Hughey
Computer Engineering
University of California
Santa Cruz, CA 95064
rph@cse.ucsc.edu

Anders Krogh
Electronics Institute, Build. 349
Technical University of Denmark
2800 Lyngby, Denmark
krogh@nordig.ei.dth.dk

I. Saira Mian
Sinsheimer Laboratories
University of California
Santa Cruz, CA 95064
saira@fangio.ucsc.edu

Kimmen Sjölander
Computer Science
University of California
Santa Cruz, CA 95064
kimmen@cse.ucsc.edu

David Haussler
Computer Science
University of California
Santa Cruz, CA 95064
haussler@cse.ucsc.edu

Abstract

A Bayesian method for estimating the amino acid distributions in the states of a hidden Markov model (HMM) for a protein family or the columns of a multiple alignment of that family is introduced. This method uses Dirichlet mixture densities as priors over amino acid distributions. These mixture densities are determined from examination of previously constructed HMMs or multiple alignments. It is shown that this Bayesian method can improve the quality of HMMs produced from small training sets. Specific experiments on the EF-hand motif are reported, for which these priors are shown to produce HMMs with higher likelihood on unseen data, and fewer false positives and false negatives in a database search task.

Introduction

Hidden Markov models (HMMs) are a class of statistical models, related to profiles (Waterman and Perlwitz, 1986; Barton and Sternberg, 1990; Gribskov *et al.*, 1990; Bowie *et al.*, 1991; Lüthy *et al.*, 1991), that can be successfully applied to the problems of modeling protein and nucleic acid families (Churchill, 1989; White *et al.*, 1994; Stultz *et al.*, 1993; Krogh *et al.*, 1994; Hughey, 1993; Baldi *et al.*, 1992; Baldi and Chauvin, 1994; K. Asai and S. Hayamizu and K. Onizuka, 1993). HMMs can be extremely effective for database searching and, without the aid of three-dimensional structural information, can in some cases generate alignments equal in quality to those produced by methods incorporating such high-level information.

One disadvantage of HMM methods is that they require many training sequences from the protein family or domain of interest. When training sets are small, calculating the optimal model for a given protein family is difficult because there are insufficient data to properly estimate the parameters. As only a small number of sequences is available for most protein families and domains, to date the method has only been applied to large, well studied families such as the EF-hand family of proteins that posses a Ca metal ion binding motif. Experimenting with the globin family, we found that 200 randomly chosen family members were required to obtain quality models. The majority of protein families represented in the databases contain far fewer members.

One natural solution is to introduce additional prior information into the construction of the HMM. In this paper, we present methods for incorporating prior knowledge of typical amino acid distributions over positions in multiple alignments to the problem of HMM training. In fact, our HMMs themselves include a linear chain of match states that capture amino acid distributions for each position in the multiple alignment of a protein family. Thus, in a bootstrapping procedure, we can use distributions from our previously built HMMs to generate prior information for the *next* model. Additionally, databases can be searched with the model built from a small training data set to find new members of the family, increasing the size of the training set.

In this paper, we introduce Dirichlet mixture densities (Antoniak, 1974) as a means of representing prior information about typical amino acid distributions. A related use of mixture priors, in this case Gaussian mix-

[‡] This work was supported in part by NSF grants CDA-9115268 and IRI-9123692, ONR grant N00014-91-J-1162, NIH grant GM17129, a grant from the Danish Natural Science Research Council, and funds granted by the UCSC Division of Natural Sciences.

ture priors used in the context of neural net training, was given in (Nowlan and Hinton, 1992). The Dirichlet mixtures cluster amino acid distributions into prototypical classes of distributions. Using Bayes' rule, Dirichlet mixture densities can be combined with observed frequencies of amino acids to obtain posterior estimates of amino acid distributions. In a detailed set of experiments on building HMM models of the EF-hand motif, we show that such posterior estimates lead to superior HMMs. In particular, we show that HMMs for the EF-hand motif trained using appropriate priors produce models that have higher likelihood with respect to independent (non-training) sets of EF-hand motifs. Furthermore, we show that these models produce fewer false positive and false negative sequences when searching a database.

Our present work has several conceptual similarities with profile methods, particularly in regard to seeking meaningful or prototypical amino acid distributions for use in database search and multiple alignment (Waterman and Perlwitz, 1986; Barton and Sternberg, 1990; Gribskov *et al.*, 1990; Bowie *et al.*, 1991; Lüthy *et al.*, 1991). In particular, Lüthy, McLachlan and Eisenberg (1991) have analyzed multiple alignments using secondary structure information to construct a set of distributions describing the columnar statistics of secondary protein structures. The result of their work is a set of nine probability distributions, which we will call the *LME distributions*, describing the distribution of amino acids in nine different structural environments in a protein.[1] These LME distributions have been shown to increase the accuracy of profiles in both database search and multiple alignment by enabling them to take advantage of prior knowledge of secondary structure.

There are two difficulties in applying the LME distributions to HMMs. First, there is no measure of how much variance is associated with each of the distributions. This is important because Bayes rule demands that in computing the posterior, the observed frequency counts be modified less strongly when the prior distribution has a very high variance. Second, the LME distributions are directly associated with secondary structure, whereas we assume no secondary structure information is available.

Instead of beginning with secondary structure, our approach is to use unlabeled training sequences to discover, through clustering, those classes of distributions of amino acids that are intrinsic to the data. We do this with statistical methods that directly estimate the most likely Dirichlet mixture density from observed counts of amino acids. In several cases, the amino acid distributions we find are easily identified as typifying some commonly found distribution (e.g., a large non-polar), but we do not set out *a priori* to find distributions representing these structures.

For a review of the essentials of the HMM methodology we use, including architecture, parameter estimation, multiple alignments, and database searches, see (Krogh *et al.*, 1994).

Modeling amino acid distributions with Dirichlet mixtures

Examining the columns in a large multiple alignment of a homologous set of protein sequences, we see a variety of distributions of amino acids. In the extreme case, when an amino acid is highly conserved at a certain position in the protein family, such as the proximal histidine that coordinates the heme iron in hemoglobin, the distribution of amino acids in the corresponding column of the multiple alignment is sharply peaked on that one amino acid, whereas in other cases the distribution is spread over many possible amino acids.

There are many different commonly occurring distributions. Some of these reflect preference for hydrophobic amino acids, some for small amino acids, and some for more complex combinations of physicochemical features. Using a purely statistical method, we have attempted to discover and model the major types of amino acid distributions found in columns of multiple alignments. Our principle intent was to use this information to produce better multiple alignments, but the results may also be of independent biological interest.

Our primary data is a set of N count vectors. Each count vector in this set represents data from a specific column in a specific multiple alignment. Many multiple alignments of different protein families are included, so N is typically in the thousands. Let us suppose that we fix a numbering of the amino acids from 1 to 20. Then, each count vector has the form $\vec{n} = (n_1, \ldots, n_{20})$, where n_i is the number of times the i^{th} amino acids occurs in the column represented by this count vector. We make the simplifying assumption that the amino acids in a particular column are generated independently at random according to an underlying probability distribution $\vec{p} = (p_1, \ldots, p_{20})$ over the 20 amino acids. Each column, however, is assumed to have its own unique probability distribution. Our goal is to model the kinds of distributions $\vec{p}$ that are most likely generating the actual observed count vectors.

A trivial approach would be to estimate a probability distribution $\vec{p}$ separately for each count vector or column. Under our independence assumption, a single count vector $\vec{n}$ is interpreted as data from a multinomial distribution with unknown parameters $\vec{p}$. We can estimate the p_i parameters from this data using the usual maximum likelihood method, *i.e.* by finding the p_i's that maximize

$$\mathrm{Prob}(n_1, \ldots, n_{20} | p_1, \ldots, p_{20}).$$

As is well known, this leads to the obvious estimate $\hat{p}_i = n_i/n$, where $n = \sum_{i=1}^{20} n_i$. These $\hat{p}_i$ values are just

[1] In more recent work, they have used 18 different distributions (Bowie *et al.*, 1991).

a summary of the raw data, and for small n provide only poor estimates for the actual underlying probability distributions.

To solve this problem, we propose a two-stage stochastic model for the data. We assume that, for each count vector $\vec{n}$, first a distribution $\vec{p}$ is chosen independently from an unknown probability density ρ over all such distributions, then the count vector $\vec{n}$ is generated according to the multinomial distribution with parameters $\vec{p}$. Our goal is now to bypass the estimation of the individual $\vec{p}$ parameter vectors and instead use the data from the count vectors to directly estimate the underlying density ρ.

To make this feasible, we have assumed a simple parametric form for the density ρ, initially choosing a *Dirichlet density* with unknown parameters $\alpha_1, \ldots, \alpha_{20}$ (Berger, 1985; Santner and Duffy, 1989). The value of ρ at a particular point $\vec{p}$ is given by:

$$\rho(\vec{p}) = \frac{\prod_{i=1}^{20} p_i^{\alpha_i - 1}}{Z}, \tag{1}$$

where Z is the normalizing constant such that ρ integrates to unity. Letting $\alpha = \sum_{i=1}^{20} \alpha_i$, it is easy to see that the Dirichlet density with parameters $\alpha_1, \ldots, \alpha_{20}$ is peaked around the amino acid distribution where $p_i = \alpha_i / \alpha$. The larger α is, the more peaked is the density. Thus, modeling ρ by a simple Dirichlet density assumes that all amino acid distributions are deviations from one central underlying amino acid distribution.

Because this latter assumption seems dubious, in further experiments we have used a more complex form for the density ρ. In particular, we assume that ρ has the form

$$\rho = q_1 \rho_1 + \ldots + q_k \rho_k, \tag{2}$$

where each ρ_j is a Dirichlet density and the numbers $q_1, \ldots, q_k$ are positive and sum to one. A density of this form is called a *mixture density* (or, in this specific case, a *Dirichlet mixture density*), and the q_j values are called *mixture coefficients*. Each of the densities ρ_j is called a *component* of the mixture. By using a Dirichlet mixture, we hope to discover several underlying "prototypical" amino acid distributions: a collection of amino acid distributions such that each observed column count from a multiple alignment is very likely obtained from a minor variant of one of the prototypes. The process is similar to clustering amino acid distributions into types. However, instead of having just 20 parameters $\alpha_1, \ldots, \alpha_{20}$ to estimate, as in the case of a single Dirichlet density, we now have $21 \times k$ parameters to estimate: twenty α_i values for each of the k components of the mixture and twenty mixture coefficients. This is feasible if k small and the number of count vectors available is large.

We have used the maximum likelihood method to estimate the parameters of ρ from the set of count vectors. Thus, we searched for the parameters of ρ that would maximize the probability of occurence of the observed count vectors. In the simplest case, we have simply fixed the number of components k to a particular value and then estimated the $21 \times k$ parameters. In other experiments, we tried to estimate k as well. Unfortunately, even for fixed k, there does not appear to be an efficient method of estimating these parameters that is guaranteed to always find the maximum likelihood estimate. However, the standard expectation-maximization (EM) algorithm for mixture density estimation works well in practice.[2]

The final result of this statistical estimation is a set of k mixture coefficients, $q_1, \ldots, q_k$, and a set of k Dirichlet parameter vectors, $\vec{\alpha}_1, \ldots, \vec{\alpha}_k$, where $\vec{\alpha}_j$ is the vector $\alpha_1^{(j)}, \ldots, \alpha_{20}^{(j)}$ of parameters of the j^{th} Dirichlet component in the mixture. These parameters are possibly interesting in themselves, in terms of what they reveal about protein structure (as discussed in the next section), however their main use will be in improving multiple alignments and other models derived from multiple alignments, such as profiles and HMMs.

Consider the production of a multiple alignment for a protein family. From one column in a rough, initial alignment, a count vector $\vec{n}$ is obtained. One immediate question to consider is whether or not the count vector is similar to one of the distributions on amino acids that commonly occurs in protein families. If this is the case, then this can be considered evidence for the accuracy of the alignment (otherwise, it may be considered evidence against that particular alignment). Furthermore, assuming a correspondence, one may ask what structural role is usually played by positions that have this kind of distribution and use this information to discover the common structure of the proteins family. Finally, if only a relatively small number of proteins make up the alignment (less than 30), then one does not expect the counts $\vec{n}$ to yield good estimates of the actual probabilities $\vec{p}$ that each amino acid will appear in that position in other proteins from the family not yet included in the alignment. Thus, it is difficult to use this alignment to search a database for other proteins in the family, by profile, HMM, or alternative methods. By combining these counts with prior information from the Dirichlet densities, better estimates of the p_i parameters can be obtained. In this sense the Dirichlet mixture prior provides an alternative to the use of the Dayhoff matrix (Dayhoff *et al.*, 1978), and other means of "smoothing" probability estimates based on a few occurrences of amino acids.

Once we have estimated the parameters of a Dirichlet mixture, these issues can all be addressed in a purely statistical manner. We do this by treating the Dirichlet mixture density ρ as a *prior probability den-*

[2] An introduction to this method of mixture density estimation is given in the book by Duda and Hart (1973). We have modified their procedure to estimate a mixture of Dirichlet rather than Gaussian densities. The mathematical details of this will be described in a separate paper (Brown *et al.*, 1993).

sity over the possible actual distributions $\vec{p}$ in the new protein family being modeled. Then, given the actual counts $\vec{n}$ for a particular column of a multiple alignment, we can use a Bayesian method to determine the type of distribution that may have generated these counts, (i.e., which of the k components of the Dirichlet mixture may have produced the underlying probability distribution $\vec{p}$ for this position) and to produce estimates $\hat{p}_1, \ldots, \hat{p}_{20}$ of the actual p_i values. The latter estimates will differ from the maximum likelihood estimates, and should be much better when n is small.

It is straightforward to derive the formulas for these Bayes estimates, assuming a Dirichlet mixture prior (Brown *et al.*, 1993). In the first case, for each j between 1 and k, we want to calculate $\mathrm{Prob}(j|\vec{n})$, the posterior probability that the underlying probability distribution $\vec{p}$ that produced the observed counts $\vec{n}$ was chosen from the j^{th} component of the Dirichlet mixture. Hence, instead of identifying one single component of the mixture that accounts for the observed data, we determine how likely each individual component is to have produced the data. Using Bayes rule,

$$\mathrm{Prob}(j|\vec{n}) = \frac{q_j \mathrm{Prob}(\vec{n}|\rho_j)}{\sum_{l=1}^{k} q_l \mathrm{Prob}(\vec{n}|\rho_l)}. \qquad (3)$$

And, if $n = \sum_{i=1}^{20} n_i$ and $\alpha^{(j)} = \sum_{i=1}^{20} \alpha_i^{(j)}$,

$$\mathrm{Prob}(\vec{n}|\rho_j) = \frac{\Gamma(n+1)\Gamma(\alpha^{(j)})}{\Gamma(n+\alpha^{(j)})} \prod_{i=1}^{20} \frac{\Gamma(n_i + \alpha_i^{(j)})}{\Gamma(n_i+1)\Gamma(\alpha_i^{(j)})},$$

where $\Gamma(x)$ is the gamma function. Hence this gives an explicit formula for the first kind of estimate.

For the second kind of estimate, the estimate of the p_i parameters from the counts n_i, again using Bayes rule[3],

$$\hat{p}_i = \sum_j \mathrm{Prob}(j|\vec{n}) \frac{n_i + \alpha_i^{(j)}}{n + \alpha^{(j)}} \qquad (4)$$

We propose this method as a new way of interpreting count data from multiple alignments. In particular, we suggest that a comprehensive Dirichlet mixture density be constructed that covers most of the amino acid distributions that have been found in existing multiple alignments. Then, when new multiple alignments are constructed, we suggest that the statistics from each column be used to classify that column based on the posterior probabilities of the components of the Dirichlet mixture, using Equation 3, and that the underlying probabilities of the 20 amino acids for that column be estimated using Equation 4. In the following section we describe the experiments we have done using this method.

Sequences	Columns in Alignment	Protein Family
400	147	Globins
193	254	Kinases
88	401	Elongation

Figure 2: Protein families included in the HMM data set.

Results

Obtaining Priors

As described above, our approach focuses on an automated construction of priors based on multiple alignments. Here we describe the construction of several Dirichlet mixture priors and demonstrate the effectiveness of these priors in building accurate models for the EF-hand motif.

We used two sources of multiple alignments for our raw count data: alignments from the HSSP database (Sander and Schneider, 1991) (Figure 1), and multiple alignments we generated using HMMs to model the kinase, globin and elongation factor families (Haussler *et al.*, 1993; Krogh *et al.*, 1994) (Figure 2). The total number of columns from the HSSP alignments was 5670; the number of columns from the HMM alignments totaled 802.

The HSSP database contains multiple alignments of proteins obtained by taking a single protein whose three dimensional structure is known, and aligning to it other proteins that are deemed homologous above a certain threshold to this protein but whose structure is not known. In (Sander and Schneider, 1991), a representative set of HSSP multiple alignments is suggested that includes a variety of different protein types. We used all the multiple alignments in this representative set with 30 or more sequences to obtain our HSSP count data. These proteins are listed in Figure 1.

Sequences used to create HMM alignments were obtained from various sources. The training data we used to create our kinase alignment came from the March 1992 release of the protein kinase catalytic domain database maintained by S. K. Hanks and A. M. Quinn (1991). This set is biased towards sequences from vertebrates and higher eucaryotes but includes some from lower eucaryotes. There are only two kinases encoded by viral genomes. Training data for the globin alignment consisted of all globins from the SWISS-PROT database, release 22 (Barioch and Boeckmann, 1991). Elongation factor sequences were drawn from the SWISS-PROT database, releases 22 and 23. Multiple alignments for these sequences were produced by HMMs we built for these families, as described in (Krogh *et al.*, 1994; Hughey, 1993). Summary information for these data sets is given in Figure 2.

Using the maximum likelihood procedure described in the previous section, we estimated the parameters of both a one component Dirichlet mixture density and

[3] This is actually the mean posterior estimate of the parameters $\vec{p}$.

Sequences	HSSP identifier	Protein Family
948	1HDS	HEMOGLOBIN
475	1FDL	IG*G1 FAB FRAGMENT
372	2FBJ	IG*A FAB FRAGMENT
287	2PKA	KALLIKREIN A
251	7ADH	ISONICOTINIMIDYLATED LIVER ALCOHOL DEHYDROGENASE
242	1TRC	CALMODULIN
191	2TGP	TRYPSINOGEN COMPLEX WITH PANCREATIC TRYPSIN INHIBITOR
178	1TGS	TRYPSINOGEN COMPLEX WITH PORCINE PANCREATIC SECRETORY
130	3HLA	HUMAN CLASS I HISTOCOMPATIBILITY ANTIGEN A2.1
126	2RUS	RUBISCO
109	3CYT	CYTOCHROME $C
107	4INS	INSULIN
102	5P2P	PHOSPHOLIPASE A=2
89	1R08	RHINOVIRUS 14
89	2RR1	RHINOVIRUS 14
89	2RS3	RHINOVIRUS 14
82	3SGB	PROTEINASE B FROM STREPTOMYCES GRISEUS
81	1CDT	CARDIOTOXIN V=4===/II$==
77	2MEV	MENGO ENCEPHALOMYOCARDITIS VIRUS COAT PROTEIN
71	1NXB	NEUROTOXIN $B
65	2LTN	PEA LECTIN
63	1GD1	HOLO-*D-*GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
63	1WSY	TRYPTOPHAN SYNTHASE
60	1FC2	IMMUNOGLOBULIN FC AND FRAGMENT B OF PROTEIN A COMPLEX
59	1FC1	FC1 FC FRAGMENT
54	1ETU	ELONGATION FACTOR TU
53	8RSA	RIBONUCLEASE *A
49	5HVP	HIV$-1 PROTEASE COMPLEX WITH ACETYL-PEPSTATIN
46	4LYZ	LYSOZYME
46	9API	MODIFIED ALPHA=1=-*ANTITRYPSIN
41	2CD4	CD4$
39	1GCR	GAMMA-/II$ CRYSTALLIN
38	2SBT	SUBTILISIN NOVO
38	2SOD	CU,ZN SUPEROXIDE DISMUTASE
36	1CSE	SUBTILISIN CARLSBERG
35	9WGA	WHEAT GERM AGGLUTININ
33	3ICB	CALCIUM-BINDING PROTEIN
31	1CMS	CHYMOSIN B
31	5LDH	LACTATE DEHYDROGENASE H=4= AND S-$LAC-/NAD$==+== COMPLEX
30	1MHU	CD-7 METALLOTHIONEIN-2
30	2MRT	CD-7 METALLOTHIONEIN-2

Figure 1: Protein families included in the HSSP data set.

a nine component Dirichlet mixture density from the 5670 count vectors obtained from the HSSP multiple alignments. We call these Dirichlet mixtures *HSSP1* and *HSSP9*, respectively. As mentioned in the previous section, the EM method we use is not guaranteed to always find the optimal setting of the parameters. However, multiple runs of the program with different initial parameter settings, yielded virtually identical priors, indicating that these solutions are very stable. In addition, we conducted an experiment to find a prior with a larger number of components. For this experiment, we started with 100 components using random initial values for the Dirichlet parameters. After eliminating those components found to not represent any of the data, we obtained a mixture prior having 62 components, which we call *HSSP62*.

Similar experiments were done for the HMM alignments, obtaining Dirichlet mixture priors with one component, nine components, and 33 components (*HMM1, HMM9, HMM33*). The results for the single-component and nine-component priors were also shown to be stable with respect to the initial starting point of the estimation procedure.

We studied the priors we obtained from the HMM alignments and the HSSP alignments and found several components common to both sets. Similarity between components was determined by Kullback-Leibler distance and by examination of the physico-chemical attributes of the distributions. The α parameters of the HMM1 and HMM9 priors are given in Figure 3. The distributions and physico-chemical attributes of the components of these priors are summarized in Figures 4 and 5.

In general, the physico-chemical attributes of the components of the HMM9 prior are consistent with biological intuition. When we order the components with respect to their mixture coefficients (i.e., their probabilities), the first component, HMM9.1, contains mostly small residues. The second component, HMM9.2, is large, charged and polar. HMM9.3 is polar, and has mostly negatively charged residues, except for Alanine (A), which is small, neutral, and can be found in virtually every environment. HMM9.4 has a weak tendency towards hydrophobic residues and contains three large non-polar residues, Isoleucine (I), Leucine (L), and Valine (V) with high probability. However, it also contains a single charged residue, Lysine (K) with high probability, but it is worth noting that Lysine possesses a long hydrophobic carbon chain in addition to the positively charged nitrogen atom. HMM9.5 is strongly hydrophobic and contains uncharged, nonpolar amino acids. HMM9.6 is charged, hydrophilic and polar, and HMM9.7 is negatively charged and aromatic. HMM9.8 is strongly hydropho-

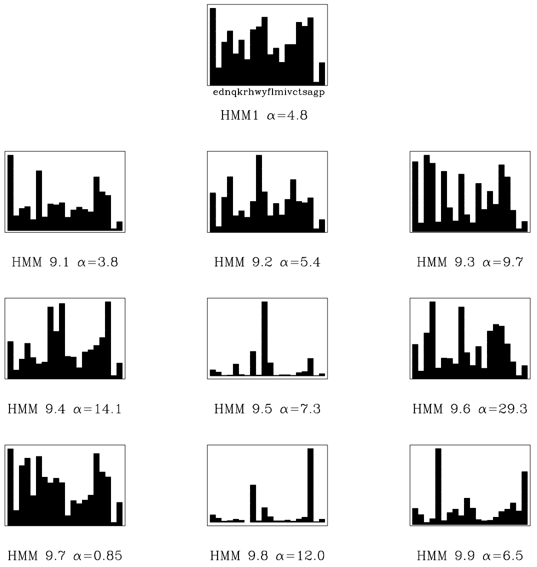

Figure 3: α parameters for the HMM1 and HMM9 priors. These bar charts show the 20 α_i parameters of the Dirichlet density of HMM1 and of each of the nine components of HMM9. The ordering of the residues used is alphabetic in the single-letter representation of each amino acid. Each bar chart is scaled by the largest α_i. The parameter α is the sum of the α_i, which gives some idea of the real magnitude of the parameters.

bic, non-polar and uncharged. HMM9.9 greatly emphasizes large residues aswell as aromatic, hydrophobic and uncharged residues.

In addition to the priors we obtained via maximum likelihood estimation, we tested the effectiveness of some additional priors: the standard uniform prior called *Add One*, [4] priors obtained directly from the nine-component LME distributions (Lüthy *et al.*, 1991) and a 29-component *EF-hand custom* prior in which each component is derived from a column in our EF-hand multiple alignment. The prior derived from the nine-component LME distributions was obtained by forming Dirichlet mixture components for each of the nine LME amino acid distributions with the same means and a fixed variance.[5] The 29-component EF-hand custom prior was designed specifically as a kind of control for the EF-hand motif experiments reported in the next section. Each component of this mixture is a Dirichlet distribution that is strongly peaked on the particular amino acid distribution at one of the positions in a multiple alignment of 885 EF-hand motifs. We use it as a control in our EF-hand experiments to indicate what kind of performance we might expect if we used the best possible prior for obtaining HMMs and multiple alignments of EF-hand sequences. Of course this particular prior will be useless for other kinds of proteins.

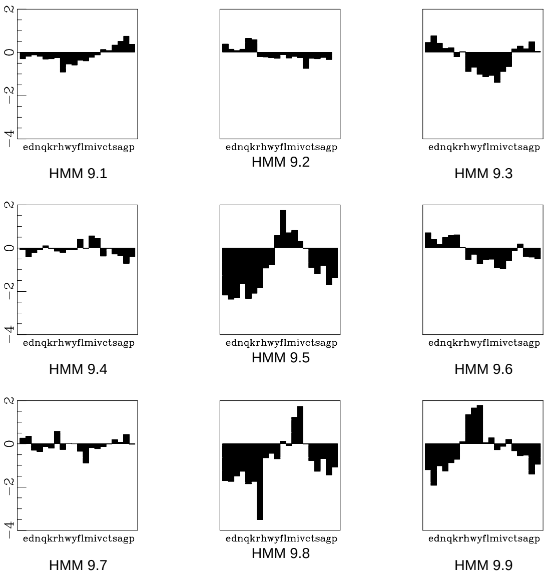

Figure 4: Log ratios for the distributions represented by the HMM1 and HMM9 priors. The i^{th} bar in the graph for HMM9.j shows the logarithm of the ratio p_i/q_i where p_i is the probability of the i^{th} amino acid in the mean of the j^{th} component of the HMM9 prior. The variable, q_i, is the probability of the i^{th} amino acid in the overall mean of HMM1. These values represent the difference in mean amino acid distribution of the j^{th} component from the background distribution. Positive numbers indicate higher values than the background distribution; negative numbers represent lower values than the background.

Using Priors to Build HMMs

We conducted a series of experiments on building HMMs for the EF-hand motif. EF-hands are an approximately 29-residue structure present in cytosolic calcium-modulated proteins (Nakayama *et al.*, 1992; Persechini *et al.*, 1989; Moncrief *et al.*, 1990). These proteins bind the second messenger calcium (Ca^{2+}) and in their active form function as enzymes or regulate other enzymes and structural proteins. The EF-hand motif consists of an α-helix, a loop binding a Ca^{2+} ion and a second helix. We chose EF-hands to

[4] This prior is often used to avoid zero probabilities on distributions in states of HMMs. Posterior estimates for this prior are obtained by simply adding one to all observed frequency counts and then normalizing so that the parameters sum to one.

[5] The variance was set so that the sum of the α_i parameters was 10. This appeared to work best for our experiments with EF-hand sequences, but further experimentation would be required to find the optimal value.

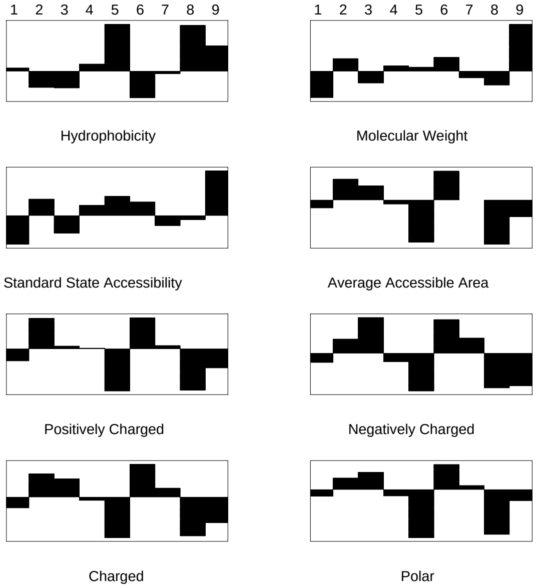

Figure 5: Physico-chemical characteristics of components of the HMM9 prior. Here each bar chart shows the relative scores of one characteristic for all nine components of HMM9. A positive score indicates that the distribution represented by the component puts more weight on residues with that characteristic than does the background distribution (represented by HMM1). A negative score indicates that less weight is put on residues with this characteristic. Each characteristic is defined by a numerical value for all residues, then these are averaged with respect to the distribution, and finally the background average is subtracted. Definitions of the numerical scores are taken from (Fasman, 1989) (Hydrophobicity, Standard-state accessibility, Average accessible area), (Hunter, 1987) (Molecular Weight), and (King and Sternberg, 1990) (Polar, Charged, Positively and Negatively Charged).

demonstrate the ability of mixture-priors to compensate for limited sample sizes because the motif's small size allowed many experiments to be performed relatively rapidly. Furthermore, a large number of EF-hand motif sequences are available.

For these experiments we used the June 1992 database of EF-hand sequences maintained by Kretsinger and co-workers (Nakayama *et al.*, 1992). Sequences in this database are proteins containing two or more copies of the EF-hand motif. We extracted the EF-hand structures from each of the 242 sequences in the database, obtaining 885 EF-hand motifs having an average length of 29. Training sets were constructed by randomly extracting subsets of size 5, 10, 20, 40, 60, 80, and 100.

For each training set size and each prior, several HMMs were built. We evaluated each HMM on a separate test set containing EF-hand sequences not in the training set, yielding an average negative log likelihood (*NLL*) score over all test sequences for each model (Krogh *et al.*, 1994). Lower scores represent more accurate models. For every combination of training sample size and prior used, we took the average test-set NLL-score across all models, and the standard deviation of the test-set NLL-scores. The results for the *Add One*, HMM1, HMM9, and EF-hand custom priors are shown in Figure 6. The results of tests using priors derived from the HSSP alignments are shown in Figure 7.

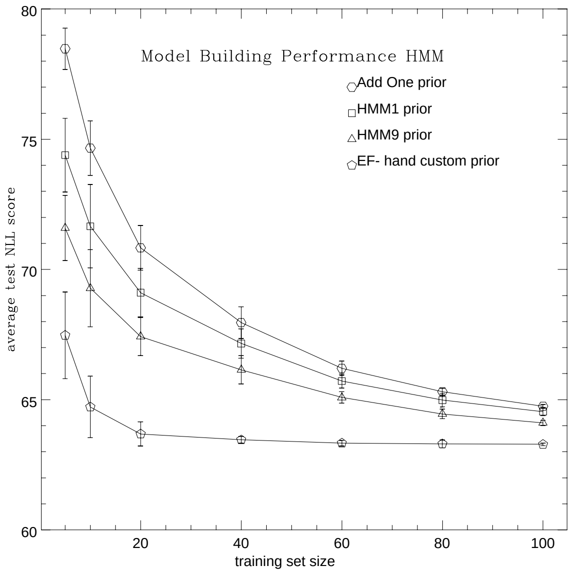

Figure 6: Average NLL scores on test data for HMMs built using different combinations of training set sizes and priors estimated from the HMM data. Bars indicate one standard deviation above and below the mean. For sample sizes 5, 10, and 20 we did 15 repetitions with independently chosen training sets. For other sample sizes we performed five repetitions.

From these Figures, we see that Add One and HSSP1 perform the worst, followed by HMM1, HSSP9, HMM9 and EF-hand custom prior. HSSP62 and HMM33, not shown, both perform about the same as HMM9, which was close in performance to HSSP9. We conducted tests using the nine LME distributions on sample size of 10. While these are not shown, the results were at the midpoint between the performance of HMM1 and HMM9. Further tests on the nine LME and the 18 LME distributions are in progress.

In our previous work, the NLL score has always been almost perfectly correlated with superior multiple alignments and database search. To further demonstrate the latter point, we tested some of the HMMs built from various priors on their ability to discriminate sequences containing the EF-hand domain from

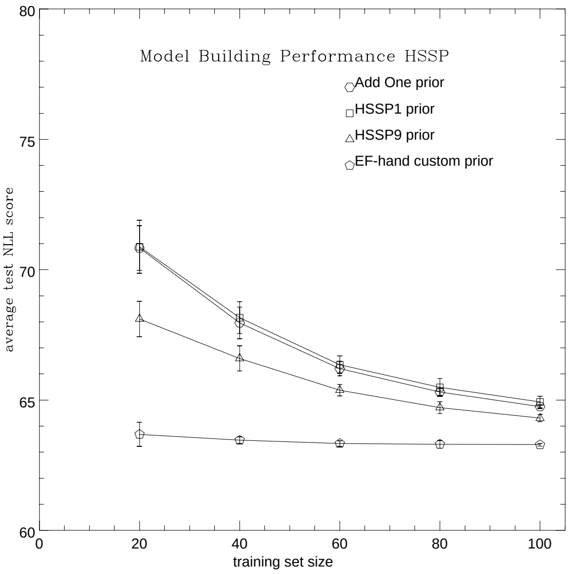

Figure 7: Average NLL scores on test data for HMMs built using different combinations of training set sizes and priors estimated from the HSSP data. Bars indicate one standard deviation. For sample size 20 we did 15 repetitions with independently chosen training sets. For other sample sizes we performed five repetitions.

those not containing the domain. To do this we choose models built from training samples of sizes 5, 10, and 20, and using the Add one, HMM1, HMM9 and EF-hand custom priors. For each sample size and prior, we built an HMM as above and then used it to search the SWISS-PROT database for sequences that contain the EF-hand motif, using the method described in (Krogh *et al.*, 1994). The results are given in Figure 8.

The results show again that HMM9 performs better than HMM1, which performs better than Add one. Unfortunately, only one test was done for each combination of sample size and prior, so the results are not as statistically clear as those for NLL-score.

Finally, we note that while the HSSP alignments contain EF-hand-specific proteins, the HMM alignments do not. Interestingly, results of experiments show that the HMM-derived priors perform better. This confirms that these priors do indeed capture some universal aspect of amino acid distributions that are meaningful across different protein families.

Conclusions

The use of Dirichlet mixture priors has been shown to increase the accuracy of HMMs for protein families where only a small number of sequences are available. In particular, the ability of models trained using prior information to discriminate members of protein families from non-members is enhanced. Thus, database search using these models can potentially yield previously unknown members of the family, enlarging the training set. From this new set, an even better model can be obtained, enabling the iterative refinement of the HMM in a bootstrapping fashion.

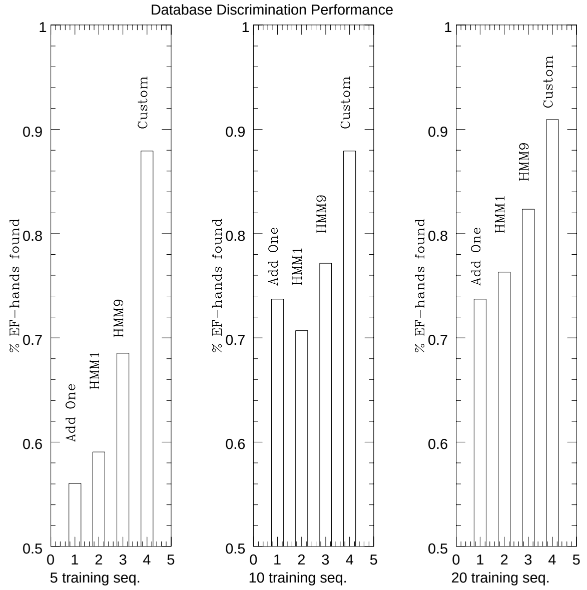

Figure 8: Discrimination results for models trained with training set sizes of 5, 10, and 20 sequences and priors. The percentage of EF-hand sequences that are found through a database search is reported for models trained with different training set sizes and priors. The cutoff is set so that there are no false positive classifications.

As experiments on the EF-hand domain using custom priors demonstrate, if one has a library of Dirichlet priors spanning a variety of amino acid distributions, such that virtually all possible distributions are represented, even extremely small training sets can in principle yield final models that are close to optimal. Ideally, such a library would be continually updated as new models and alignments are produced. We plan to do this as we continue to build HMMs for protein families.

Acknowledgments

We wish to thank James Bowie and his colleagues for providing us with an additional 18 distributions spanning a large variety of secondary structures. Experiments using these distributions are currently in progress.

The protein project has also included: Brad Gulko, Harry Noller, Rebecca Underwood, and Anne Urban. We gratefully acknowledge Tony Fink for his comments on our amino acid distributions. Jon Becher and MasPar Corporation generously provided time on several

MasPar computers.

References

Antoniak, C. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* 2:1152–1174.

Baldi, P. and Chauvin, Y. 1994. Smooth on-line learning algorithms for hidden Markov models. *Neural Computation* 6(2):305–316.

Baldi, P.; Chauvin, Y.; Hunkapiller, T.; and McClure, M. A. 1992. Adaptive algorithms for modeling and analysis of biological primary sequence information. Technical report, Net-ID, Inc., 8 Cathy Place, Menlo Park, CA 94305.

Barioch, A. and Boeckmann, B. 1991. *Nucleic Acids Research* 19:2247–2249.

Barton, G. J. and Sternberg, M. J. 1990. Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *Journal of Molecular Biology* 212(2):389–402.

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Bowie, J. U.; Lüthy, R.; and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170.

Brown, M. P.; Hughey, R.; Krogh, A.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1993. Dirichlet mixture priors for HMMs. In preparation.

Churchill, G. A. 1989. Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* 51:79–94.

Dayhoff, M. O.; Schwartz, R. M.; and Orcutt, B. C. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, D. C. chapter 22, 345–358.

Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* Wiley, New York.

Fasman, G. 1989. *Prediction of protein structure and the principles of protein conformation.* Plenum Press, New York.

Gribskov, M.; Lüthy, R.; and Eisenberg, D. 1990. Profile analysis. *Methods in Enzymology* 183:146–159.

Hanks, S. K. and Quinn, A. M. 1991. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods in Enzymology* 200:38–62.

Haussler, D.; Krogh, A.; Mian, I. S.; and Sjölander, K. 1993. Protein modeling using hidden Markov models: Analysis of globins. In *Proceedings of the Hawaii International Conference on System Sciences*, volume 1, Los Alamitos, CA. IEEE Computer Society Press. 792–802.

Hughey, Richard 1993. Massively parallel biosequence analysis. Technical Report UCSC-CRL-93-14, University of California, Santa Cruz, CA.

Hunter, L. 1987. *Representing Amino Acids with Bitstrings.* Benjamin/Cummings Pub. Co., Menlo Park, California.

K. Asai and S. Hayamizu and K. Onizuka, 1993. HMM with protein structure grammar. In *Proceedings of the Hawaii International Conference on System Sciences*, Los Alamitos, CA. IEEE Computer Society Press. 783–791.

King, R. D. and Sternberg, M. J. 1990. Machine learning approach for the prediction of protein secondary structure. *Journal of Molecular Biology* 216:441–457.

Krogh, A.; Brown, M.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* 235:1501–1531.

Lüthy, R.; McLachlan, A. D.; and Eisenberg, D. 1991. Secondary structure-based profiles: Use of structure-conserving scoring table in searching protein sequence databases for structural similarities. *PROTEINS: Structure, Function, and Genetics* 10:229–239.

Moncrief, N. D.; Kretsinger, R. H.; and Goodman, M. 1990. Evolution of EF-hand calcium-modulated proteins. I. relationships based on amino acid sequences. *Journal of Molecular Evolution* 30:522–562.

Nakayama, S.; Moncrief, N. D.; and Kretsinger, R. H. 1992. Evolution of EF-hand calcium-modulated proteins. ii. domains of several subfamilies have diverse evolutionary histories. *Journal of Molecular Evolution* 34:416–448.

Nowlan, S. J. and Hinton, G. E. 1992. Soft weight-sharing. In Moody, ; Hanson, ; and Lippmann, , editors 1992, *Advances in Neural Information Processing Systems 4*, San Mateo, CA. Morgan Kauffmann Publishers.

Persechini, A.; Moncrief, N. D.; and Kretsinger, R. H. 1989. The EF-hand family of calcium-modulated proteins. *Trends in Neurosciences* 12(11):462–467.

Sander, C. and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1):56–68.

Santner, T. J. and Duffy, D. E. 1989. *The Statistical Analysis of Discrete Data.* Springer Verlag, New York.

Stultz, C. M.; White, J. V.; and Smith, T. F. 1993. Structural analysis based on state-space modeling. *Protein Science* 2:305–315.

Taylor, W. R. 1986. The classification of amino acid conservation. *Journal of Theoretical Biology* 119:205–218.

Waterman, M. S. and Perlwitz, M. D. 1986. Line geometries for sequence comparisons. *Bull. Math. Biol.* 46:567–577.

White, James V.; Stultz, Collin M.; and Smith, Temple F. 1994. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Mathematical Biosciences* 119:35–75.