Bayesian Protein Family Classifier

Kunbin Qu, Lee Ann McCue, Charles E. Lawrence

Wadsworth Center for Laboratories and Research
Albany, NY 12201
quk, mccue, lawrence@wadsworth.org
(518)473-3382, FAX: (518)474-7992

Abstract

A Bayesian procedure for the simultaneous alignment and classification of sequences into subclasses is described. This Gibbs sampling algorithm iterates between an alignment step and a classification step. It employs Bayesian inference for the identification of the number of conserved columns, the number of motifs in each class, their size, and the size of the classes. Using Bayesian prediction, inter-class differences in all these variables are brought to bare on the classification. Application to a superfamily of cyclic nucleotide-binding proteins identifies both similarities and differences in the sequence characteristics of the five subclasses identified by the procedure: 1) cNMP-dependent kinases, 2) prokaryotic cAMP-dependent regulatory proteins, CRP-type, 3) prokaryotic regulatory proteins, FNR-type, 4) cAMP gated ion channel proteins of animals, and 5) cAMP gated ion channels of plants.

Introduction

Modern high-throughput sequencing technologies and genome sequencing projects have greatly accelerated the growth of the sequence databases. This expansion has led to the need for more sensitive and efficient methods of classifying proteins computationally, so that more specific inferences can be made about the structures and functions of specific classes of proteins.

Several databases of aligned collections of proteins (protein families) have been developed, such as Blocks (Henikoff et al., 1996), PROSITE (Bairoch et al., 1997), and Pfam (Sonnhammer et al., 1997). Identification of family membership using these databases can be very useful for prediction of putative biological function of unknown sequences. Several methods to create such databases have been developed: Hidden Markov Model (HMM; Krogh et al., 1994), neural nets (Wu, 1996), and Gibbs sampling (PROBE; Neuwald et al., 1997). Protein families often consist of many proteins that share some common characteristic(s), such as binding of a small molecule; alignments produced by the methods listed above provide valuable information about the family and its shared features. However, such families often include many subclasses based on more specific functions which the current database and methods do not differentiate. For example, the Ras family in Pfam contains several subfamilies, including the Ras, Rab, and Rho subfamilies (Casari et al., 1995). Therefore, protein family classification becomes important to separate the proteins within a family into subfamilies, thus providing researchers with a better understanding of the specific functional and structural characteristics of the subfamilies. Since there is no universal agreement on the definitions of the terms superfamily, family, and subfamily, we will call the starting group of sequences input into a classification procedure a collection and the output groups subclasses.

Phylogenetic tree reconstructions are the most popular method for classifying proteins into related groups, but these methods focus on evolutionary relatedness rather than on the sequence characteristics that distinguish subclasses (Golding & Felsenstein, 1990). Classification through principle components, a frequentist statistical method, has been described by Casari, et al. (1995). Classification is achieved by projections of high-dimension sequence space onto a small number of principle components. While this procedure has a number of useful features, it has two important limitations: 1) classification is limited by the amount of information captured in the principle components, accordingly there can be substantial loss of information, and 2) since the classification is based on the principle components of the entire aligned class, differences of aligned subclasses are not available to the classification procedure.

Recently, it has been shown that Bayesian statistics offers several advantages for the analysis of biopolymer data (Zhu et al., 1997; Liu & Lawrence, 1998; Lawrence tutorial). Specifically, Bayesian inference and model selection methods provide a sound means to relax or eliminate the need to specify parameter settings, and to make inference on all unknowns in a problem. Here, we describe a Bayesian procedure which classifies a collection of sequences into subclasses and multiply aligns the members of each subclass. These Bayesian methods yield inference about many important variables for each subclass, including the number and sizes of

subclasses, the number and sizes of motifs, and the number of conserved columns.

Methods

We begin our classification with any collection of sequences. Most often this collection will represent a diverse family or a superfamily of proteins which share some common sequence characteristics stemming from common functional or structural features. Often these collections will contain subclasses. The goal of this procedure is to classify the collection into subclasses and to identify the similarities and differences in the sequence characteristics of the subclasses. To achieve this goal, we employ a procedure which iteratively aligns the subclass sequences and re-assigns sequences to the post-classified subclasses. Alignment is carried out by using a recently developed Bayesian multiple sequence alignment tool, PROBE (Neuwald et al., 1997). Classification is achieved by using the predictive update version of the Gibbs sampler (Liu et al., 1995), which employs an algorithm similar in principle to the motif sampler described by Neuwald, Liu and Lawrence (Neuwald et al. 1995; Liu et al., 1995). The models used in these procedures are identical with those used by PROBE (Neuwald et al., 1997) which describe a multiple alignment by a product of multinomial models for the aligned positions while the remaining positions are modelled by a single multinomial model: the "background".

A collection of sequences is said to be classified when every sequence in the collection is assigned to one of $M_{\max}$ subclasses. Let $M_j = 1,2,\ldots M_{\max}$ be an assignment variable which indicates to which subclass sequence j has been assigned. Also let $P(M_j \mid p)$ be multinomially distributed with parameters $p = (p_1, p_2, \ldots p_{M_{\max}})$, and let $R_{j_{[1,t]}} = (R_{1,j}, R_{2,j}, \ldots R_{t,j})$ be the j-th sequence, with the length of t, in the sequence collection *R*. The alignment model used here is an extension of block based Gibbs sampling models (Lawrence et al., 1993; Neuwald et al., 1995; Liu et al., 1996; Neuwald et. al., 1997). Accordingly, the alignment of any sequence R_j can be described by a vector of the indices which specify the first position in each block, $A_j = (A_{1,j}, A_{2,j}, \ldots A_{k,j})$ where *k* is the number of blocks in the model, and the *k*-th block has a length of l_k.

The algorithm proceeds by iterating between an alignment step and a classification step. The alignment step begins with the sequences assigned to subclasses. For each subclass the sequences are aligned using the propagation algorithm (Liu and Lawrence, 1996; Neuwald et al., 1997). The MAP criterion is employed to determine the number of blocks and the number of conserved columns for the model that fits each subclass (Neuwald et al.,1997). The fragmentation procedure is employed to allocate conserved positions within and across motifs and to infer the width of each (Liu et al., 1995; Neuwald et al., 1997).

The classification step uses the predictive update version of the Gibbs sampler (Liu et al., 1995). As with all Gibbs sampling algorithms the process advances by iterating through the sequences one at a time. At any point in the iteration it begins with all the sequences assigned to subclasses and multiply-aligned within each subclass. Sequences are removed from the models one at a time, either in successive order or via sampling. In the current iteration assume that sequence j is withdrawn. Let $A_{[j]}^{(m)}$ be the alignment of all the sequences in subclass m with possible exception of sequence j which may have been removed from this subclass.

The Bayesian procedure which re-assigns sequence j to a subclass is based on joint probability of the sequence j and its alignment, i.e.

$$P(R_j, A_j \mid A_{[j]}, R_{[j]}, M_j = m) = P(R_j \mid A_j, A_{[j]}, R_{[j]}\, M_j = m) P(A_j) \qquad (1)$$

where

$$P(R_j \mid A_j, A_{[j]}, R_{[j]}, M_j = m)$$

is obtained by predictive inference (Liu et al., 1995). Historically the so-called entropic explosion in the number of alignments with increasing number of gaps has been treated by assigning gap penalties. Zhu, Liu and Lawrence (1997) have recently shown, using a Bayesian approach, that the more direct approach of discounting alignments with k gaps inversely proportional to the number of alignments with k gaps, is an alternative which shares some advantages over the historic approach. Accordingly, here we set $P(A_j) = \frac{1}{N_{A_j,m}}$, where $N_{A_j,m}$ is the number of alignments of sequence j to the model describing subclass m.

The probability that sequence j belongs subclass m is obtained from equation (1) as follows:

$$P(M_j = m \mid A_{[j]}, R_{[j]}, R_j) = \frac{P(R_j \mid A_{[j]}, R_{[j]}, M_j = m)}{\sum_M P(R_j \mid A_{[j]}, R_{[j]}, M_j = m)} \qquad (2.a)$$

where

$$P(R_j \mid A_{[j]}, R_{[j]}, M_j = m) = \sum_{all\ A_j} P(R_j, A_j \mid A_{[j]}, R_{[j]}, M_j = m) \qquad (2.b)$$

The sum in equation (2.b) is obtained recursively in the following manner. Given the alignment of *k* blocks in the subsequence $R_{[1,t]}=R_1,R_2,\ldots,R_t$. There are only two possibilities for extension to *t+1* residues: 1) residue R_{t+1}

belongs to the background model, or 2) R_{t+1} becomes the last residue in the *k*-th block. Because these two situations are mutually exclusive and exhaustive, the probability of finding *t+1* residue in sequence R_j with *k* blocks is the probability sum of the above cases:

$$P(R_{j[1,t+1]} \mid P, K = k) = P(R_{j[1,t]} \mid P, K = k) + P(R_{j[1,t-l_k+1]} \mid P, K = k-1) * P(A_{k,j} = t - l_k + 2 \mid P) \quad (3)$$

Using this basic recursion, we sum over all alignments in a manner similar to the sum forward step of the propagation recursion (Liu and Lawrence, 1996) to complete the summation in equation (2.b). Sequence j is now assigned to a subclass, say subclass m', by sampling in proportion to the probabilities given by equation (2.a). Re-assignment back to the subclass from which it was drawn is permitted.

Once sequence j has been sampled into subclass m', its data must be incorporated into the model describing subclass m'. This requires its alignment with the subclass, which is obtained by using a recursive back sampling procedure similar to the forward step of the recursion. With this sequence added to the alignment of subclass m', the process is repeated until the collection is exhausted.

After this re-classification, each of the subclasses is re-aligned using the propagation algorithm. Sometimes the initial collection will contain sequences that are too closely related. Sequence weighting (Henikoff et al., 1994) or purging (Neuwald et al., 1995) can be employed to address this problem. Here we employ a purging strategy. Since a subclass often is characterized by sequences that are more closely related than the collection from which it was derived, an adjustment in the purge level is made. The algorithm iterates between the alignment and classification until convergence. These two steps together simultaneously build the alignments of each subclass and the assignment of sequences to subclasses in proportion to the posterior probability of class membership.

The most important product of this process is the characterization of the sequence similarities and differences between the subclasses. These are obtained by comparing the posterior Dirichlet distributions of residue probabilities of the subclasses. Since Bayesian inferences are applied separately to each subclass, the subclasses often differ in the number of motifs, size and alignment of each motif, and the conserved columns in the alignment.

As shown in the results, the subclasses can contain motifs that are common to the entire collection and other motifs which are only present in a subset of classes. Motifs which are specific to subclasses describe sequence characteristics which distinguish them from the rest of the subclasses. These often describe distinctive structural and/or functional characteristics. While motifs common to the entire collection tend to describe common features, subclasses may still differ within these motifs as well. These differences are reflected in the posterior Dirichlet distributions' residue probabilities of the subclasses. We have found sequence logos to be a useful tool to display these differences (Schneider and Stephens, 1990).

Results

As a test case for Classifier, we examined putative cyclic nucleotide (cNMP) binding proteins. The cyclic nucleotides cAMP and cGMP are second messengers - changes in the intracellular concentration of these small molecules are caused by the activation of adenylate and guaninylate cyclase, respectively, which occurs in response to an extracellular signal (Alberts et al., 1994). Binding of cAMP or cGMP to target intracellular proteins then triggers a change in the activities of those target proteins. cAMP acts as a gene regulation signal in prokaryotes (Kolb et al., 1993). The cAMP-binding proteins are homodimers that bind DNA non-specifically, but undergo an allosteric change upon binding 2 molecules of cAMP (one to each monomer). The cAMP-protein complex is a highly specific DNA-binding protein that regulates transcription initiation. Cyclic nucleotides mediate their effects in eukaryotic cells primarily by activating cAMP-dependent protein kinases (Su et al., 1995). These kinases exist as heterotetramers composed of two catalytic and two regulatory subunits in the inactive state. When the regulatory subunits bind 4 molecules of cAMP (2 to each regulatory monomer), the catalytic subunits are released from the complex and are thereby activated to phosphorylate substrate proteins. A small group of cGMP-dependent protein kinases have also been identified in eukaryotes which act as homodimers, and for which the kinase activity and cGMP-binding activity are present on the same polypeptide. The only other proteins known to bind cyclic nucleotides in eukaryotes are a small group of gated ion channels in sensory neurons responsible for cell depolarization in response to external stimuli.

Existing computational methods do not divide the cNMP-binding protein superfamily into subclasses or identify sequence similarities and differences that distinguish the subclasses within the superfamily. The PROSITE database (Bairoch et al., 1997) contains two motifs describing the cNMP-binding domain, both of which are present in each of the proteins described above. Pfam (Sonnhammer et al., 1997) defines a superfamily of cNMP-binding proteins that was developed from a seed of 32 cNMP-binding sites and includes (in the full alignment) sequences from 41 different proteins with 67 cNMP-binding sites; this family has representatives from all of the types of proteins described above. Separately, Pfam defines a family of 26 prokaryotic regulatory proteins related to *Escherichia coli* CRP, a cNMP-binding regulatory protein; this Pfam pattern describes only the helix-turn-helix DNA-binding motif possessed by all members of this family of prokaryotic regulatory proteins.

We started our analysis with *Streptomyces griseus* P3, a protein of unknown function that is expressed only during sporulation of the bacterium. When P3 was compared to the PROSITE database, a potential cNMP-binding site was identified only by allowing one mismatch in each of the PROSITE cNMP-binding motifs. When compared to the Pfam database, P3 was again identified as having a potential cNMP-binding domain, but with a low bit value score (25.12). We identified a collection of 114 proteins using PROBE (Neuwald et al., 1997), given the *S. griseus* P3 sequence (gi|1196910) as the seed and using the non-redundant database (NCBI). The collection of proteins identified by PROBE as having motifs in common with *S. griseus* P3 contained representatives from all of the known types of cNMP-binding proteins. The PROBE model consisted of four motifs, three of which appeared to correspond to the cNMP-binding region, based on the PROSITE and Pfam motifs and comparisons to the proteins of known structure in the group, *E. coli* CRP (PDB accession 3GAP; Kolb et al., 1993) and bovine KAP0 (PDB accession 1RGS; Su st al., 1995). This group of 114 proteins became our cNMP-binding protein superfamily. Sequence logos for motifs 2 (part of the cNMP-binding domain) and 4 of the superfamily model are shown in Fig. 1.

Because the proteins in the superfamily appeared to

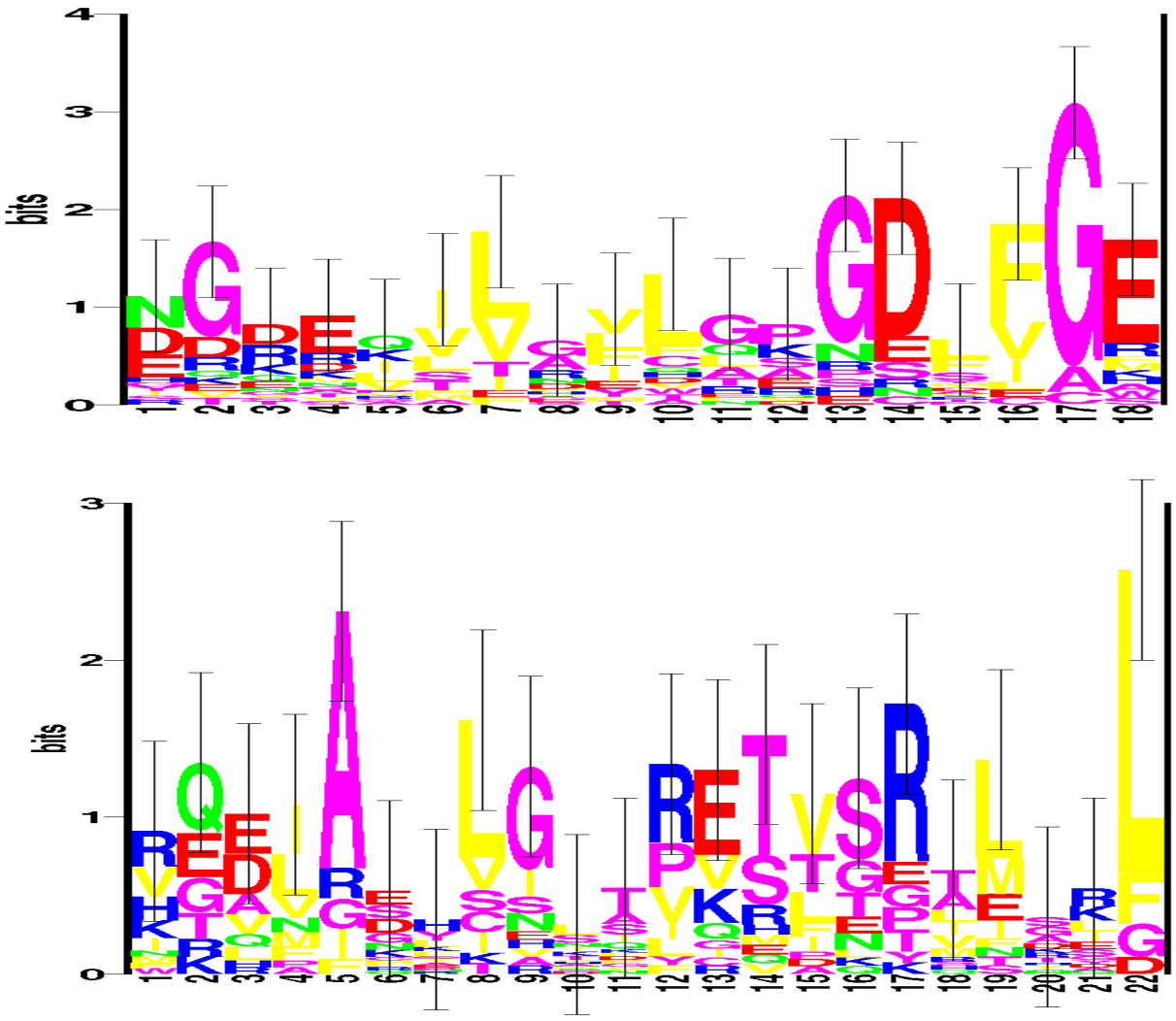

Figure 1. Sequence logos for motifs 2 (top) and 4 (bottom) of the cNMP-binding protein superfamily. There were 114 proteins in the original collection identified by PROBE. The sequence logos were produced using a BLAST cutoff of 150, which includes 24 sequences. Motif 2 represents a region common among the subclasses of this superfamily; three conserved glycine residues involved in β barrel formation are shown (positions 2, 13, and 17), as well as the glutamate that is involved in cNMP binding (position 18). Motif 4 extends from position 1 to position 33; the last 11 positions have been truncated to improve readability.

belong to approximately three separate subfamilies we randomly divided our collection into three subclasses. Therefore, the sequences of all three subfamilies were fairly evenly distributed across the subclasses. Application of the Classifier to these three groups, using a purge cutoff of 150, yielded one empty subclass and two occupied subclasses. The first of these consisted of primarily kinases and channel proteins, while the second subclass consisted of primarily kinases and the prokaryotic regulators; therefore, it appeared that the prokaryotic regulators were separated from the channel proteins, but that the kinases remained divided between these two subclasses. Each of the two subclasses were then randomly divided again, resulting in 4 subclasses. Classifier was applied again to these 4 groups, but with purge cutoff of 250; at convergence, four subclasses remained with four distinct models.

One of the subclasses contained only 5 proteins, all of which are ion channel proteins from plants. Another of the subclasses contained 27 proteins; most are known or predicted ion channel proteins from animals, though 5 are proteins of no known or predicted function. These unknowns included 3 hypothetical protein sequences from the genomes of the eukaryotes, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*, as well as 1 sequence from the chloroplast genome of the algae *Porphyra purpurea* and 1 sequence from the genome from the prokaryote *Alcaligenes eutrophus*. We have focused on the remaining two subclasses, which contained the eukaryotic kinases and the prokaryotic gene regulatory proteins, respectively, in part because each contains a protein sequence for which the structure has been solved.

Both *E. coli* CRP and bovine KAP0 form a flattened β barrel structure, which is a major part of the cAMP-binding site (Kolb et al., 1993; Su et al., 1995), and contain 5 highly conserved glycine residues that are important for the formation of the β barrel. These glycines are evident in motifs 1 (positions 11 and 23) and 2 (positions 2, 13, and 17) of the prokaryotic regulatory protein subclass (Fig. 2). Highly conserved glycine residues also appear in the eukaryotic kinase subclass motifs discussed below. In fact, these residues illustrate a motif common to all the subclasses of our superfamily.

The prokaryotic regulatory protein subclass contained 44 proteins, all of which are of prokaryotic origin and many are known to bind to DNA and regulate transcription. Of these, 5 have no known or predicted function; the remaining 39 proteins are described as members of the CRP/FNR family of regulatory proteins, for which only the CRP proteins are known to bind a cyclic nucleotide. The model for this CRP/FNR subclass contained 5 motifs. The first 3 motifs of this model correspond to the first 3 motifs of the superfamily model, and therefore to the cNMP-binding region. Motif 2 of this subclass, which corresponds to motif 2 of the superfamily model, had been noticeably modified, however. Refinement of the cNMP-binding region to better describe the prokaryotic regulatory protein subclass is reflected in the stronger conservation of the glycine-arginine-glutamate residues at positions 2-4 and the relaxed conservation at position 18 (Fig. 2, motif 2) relative to the corresponding positions in the superfamily model (Fig. 1, motif 2 positions 2-4 and 18). Additionally, motif 5 of this CRP/FNR subclass is unique to this group of proteins and encompasses the helix-turn-helix motif that is responsible for DNA binding (Fig. 2); this motif corresponds to motif 4 of the superfamily model (Fig. 1). Again, modification of this motif was apparent when positions 9, 13, 14, and 17 of motif 4 in Fig. 1 was compared to positions 13, 17, 18, and 21 of motif 5 in Fig. 2. The indicated positions in motif 5 of the subclass are important for the formation of the helix-turn-helix; the glycine (Fig. 2; motif 5 position 13) is part of the turn and the other positions are involved in DNA binding. In the subclass model, these positions exhibit clear conservation, whereas in the superfamily model, conservation at these positions is relaxed due to the inclusion of protein sequences from the kinases and channel proteins, which do not contain a helix-turn-helix. Therefore, this represents a motif unique to this subclass.

The 44 proteins in the prokaryotic regulatory protein subclass (CRP/FNR subclass) were further classified by randomly dividing the group into three subclasses and applying Classifier with a purge cutoff of 500. At convergence, three subclasses remained: a CRP subclass containing 14 proteins, a FNR subclass containing 28 proteins, and a subclass containing 2 proteins, both of which are hypothetical (i.e., translations of DNA sequence) and so are of unknown function. The CRP subclass contained *E. coli* CRP and related proteins that likely bind a cyclic nucleotide, and the FNR subclass contained *E. coli* FNR (which is also a transcription regulation protein) and related proteins that are believed to be regulated by a mechanism other than cNMP (Fischer, 1994). These subclasses have 2 motifs in common: one in the region of the 5 glycines necessary for β barrel formation, and one in the region of the helix-turn-helix. These motifs show no significant differences from the corresponding motifs of the CRP/FNR parent model (Fig. 2). The difference between these two subclasses resides in the sequence motif downstream of the conserved glycines. Both subclasses have significantly conserved motifs in this region; these motifs do not, however, resemble each other (not shown). Of particular interest is a conserved arginine, known to be important for interaction of *E. coli* CRP with cAMP (Kolb et al., 1993), that is present in the CRP subclass model, but not present in the FNR subclass model.

The kinase subclass contained 38 proteins, most of which are known kinases. Only one protein in this subclass is of unknown function, a hypothetical protein sequence from the genome of *C. elegans*. The cNMP-binding kinases bind 2 molecules of cyclic nucleotide per monomer (Su et al., 1995), which resulted in a kinase subclass model that contained two very similar cNMP-binding motifs. The final model for the kinase subclass

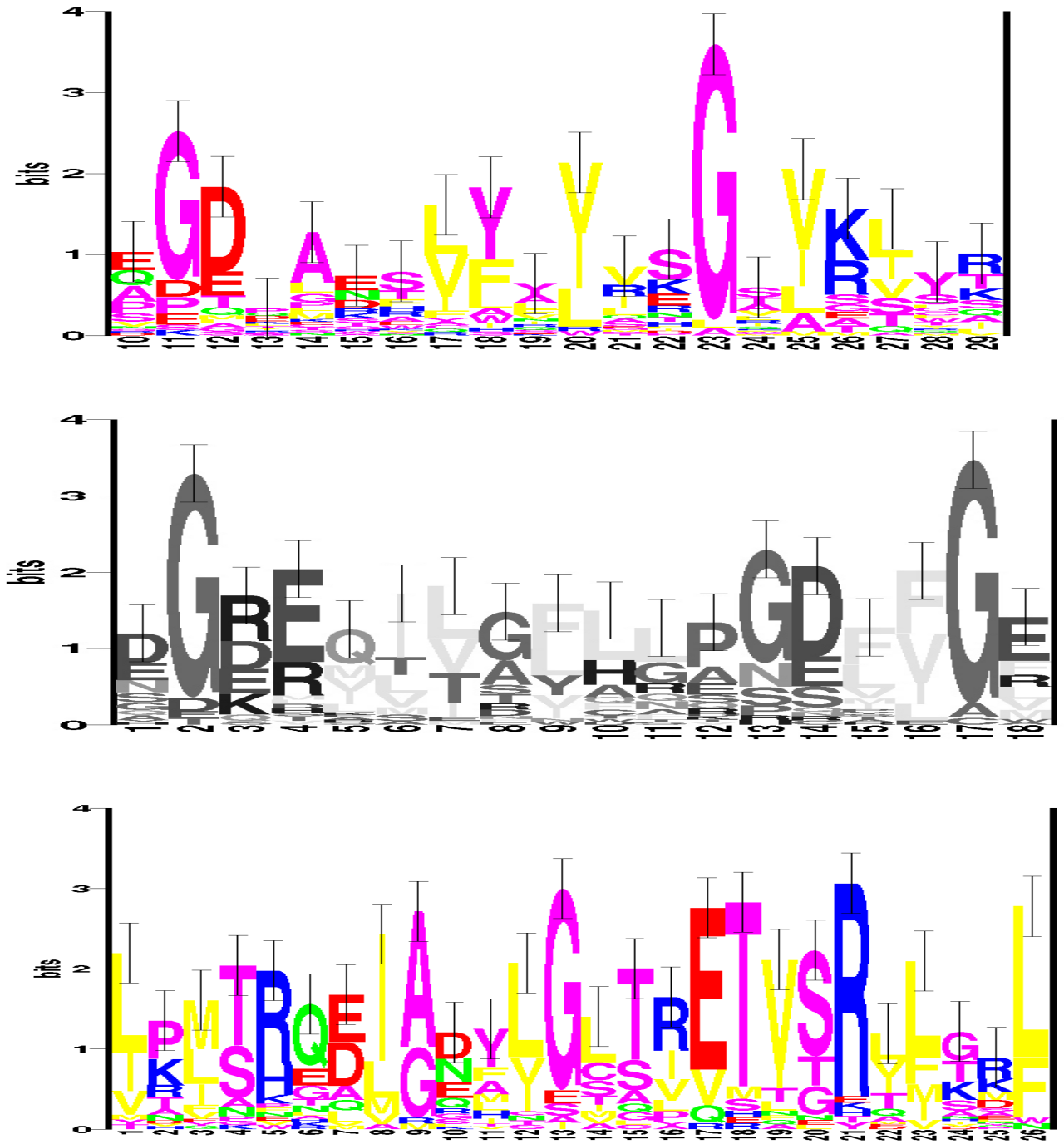

Figure 2. Sequence logos for motifs 1 (top), 2 (middle), and 5 (bottom) of the prokaryotic regulatory protein (CRP/FNR) subclass of cNMP-binding proteins. The BLAST cutoff was raised to 1000 to include enough sequences for the sequence logos, because the sequences are more closely related in a subclass. Motifs 1 and 2 show the five conserved glycines of the β barrel (see text). Motif 5 is a distinctive motif for this subclass, and represents a helix-turn-helix motif; the highly conserved glycine of the turn is at position 13, and the highly conserved glutamate and arginine at positions 17 and 21 are involved in DNA contacts (Kolb et al., 1993). The first 9 positions of motif 1 and the last 15 positions of motif 5 have been truncated to improve readability.

contained 7 motifs, of which motifs 4 and 6 corresponded to regions known to be important for cNMP binding, based on the structure of the bovine kinase, KAP0. These regions correspond, in part, to motif 2 of the superfamily model. The eukaryotic kinase cNMP-binding motifs are in Fig. 3, and show the strongly conserved glutamate and arginine residues, which distinguish the cNMP-binding pocket and have been shown by Su et al (1995) to be involved incNMP binding (Fig. 3, motif 4 positions 5 and 14, motif 6 positions 6 and 15). In addition, these two kinase motifs show two of the highly conserved glycine residues that are important for β barrel formation (Fig. 3, motif 4 positions 1 and 4, motif 6 positions 1 and 5), as well as a relatively strong conservation of hydrophobic residues immediately following the conserved glutamate (Fig. 3, motif 4 positions 6-9, motif 6 positions 7-10); these hydrophobic residues are not conserved in the cNMP-binding region of the prokaryotic regulators and so are not included in that model (Fig. 2, motif 2). Additionally, it could be expected that the two cNMP-binding regions of the kinases might differ, based on the evidence that when a molecule of cAMP binds to a regulatory subunit of a heterotetrameric kinase, allosteric changes occur in the protein structure that allow cooperative binding of the second molecule of cAMP (Su et al., 1995). Indeed, the two motifs are significantly different, $p \leq 2^{-11}$ (signed test). However, from these data we cannot determine if the differences reflect structural/functional constraints or phylogenetic ancestry.

Based on the sequence number found for the above four subclasses, estimation of the proportions of the collection that belong to each subclass are as follows: 1) 33% for

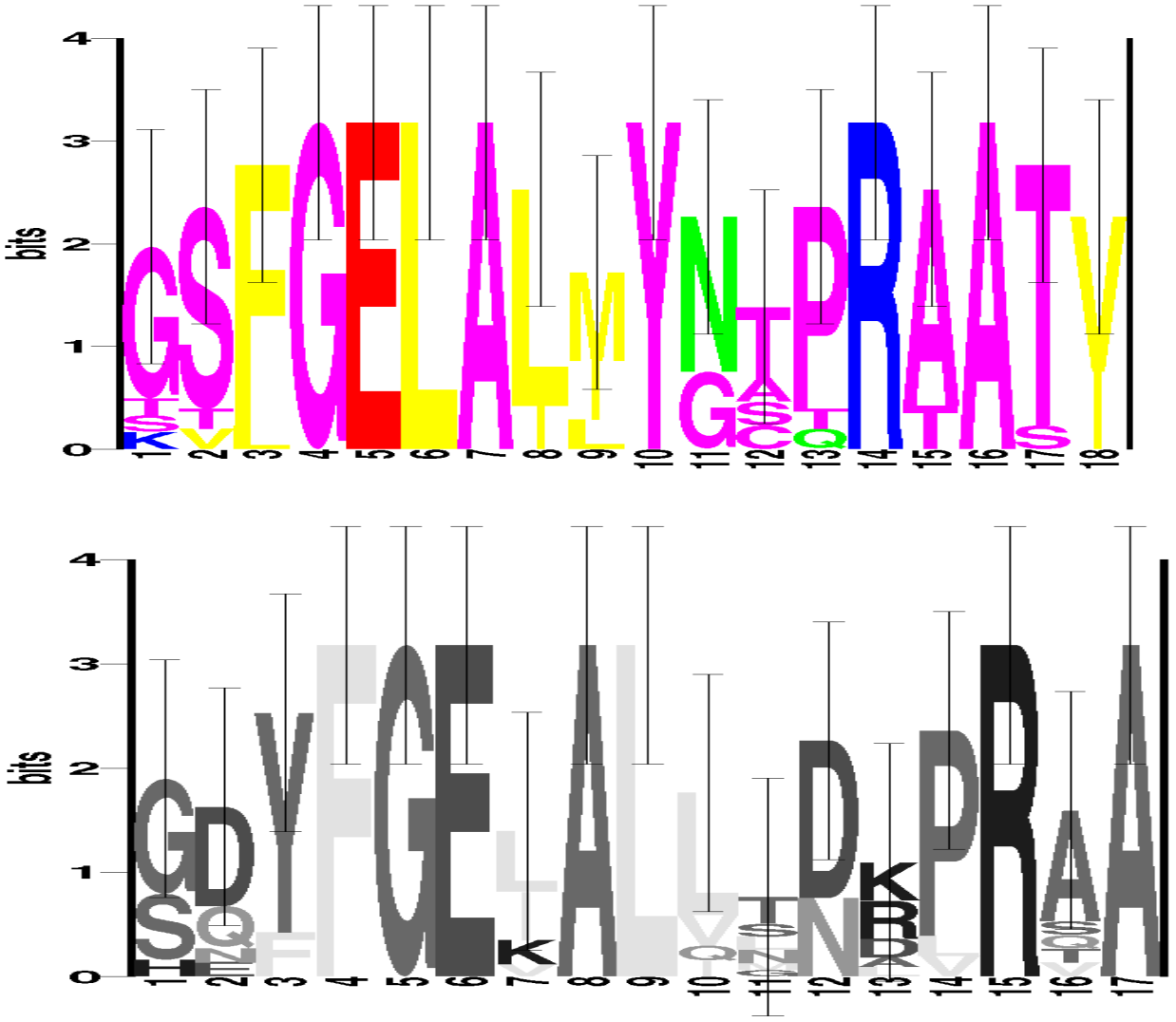

Figure 3. Sequence logos for motifs 4 (top) and 6 (bottom) of the kinase subclass of cNMP-binding proteins. The BLAST cutoff was raised to 1000 for the same reason as in Fig. 2. Positions 5 and 14 of motif 4 and positions 6 and 15 of motif 6 are the conserved glutamate and arginine residues involved in cNMP binding (Su et al., 1995).

kinase, 2) 38% for prokaryotic regulatory proteins (32% for CRP and 62% for FNR), 3) 24% for ion channel proteins from animals, and 4) 5% for ion channel proteins from plants.

Discussion

The classification procedure we have described utilizes iterations between a multiple alignment step and a classification step, and simultaneously updates all model parameters (including that of block numbers, block width, column position and amino acid composition for each column), as well as class membership. An important, distinct feature of this algorithm is that equation (2) incorporates all these inter-class differences in the variables for improved classification, via recent Bayesian prediction inference. This is a dynamic process based only on the sequence data, no manual interference is involved and no prior information about any of the model parameters is provided.

The example illustrates the application of this procedure to classifying the cNMP-binding protein superfamily into subclasses. The original collection contained proteins with a common motif that represented a common structural/functional feature, the β barrel. We have demonstrated that this procedure is able to identify differences between the subclasses of this superfamily; this was illustrated by differences in the common motifs between subclasses, as well as the identification of motifs unique to subclasses. Additionally a protein of unknown function, *S. griseus* P3, was classified with the FNR subclass of bacterial regulatory proteins, illustrating the use of this technique to classify unknown protein sequences that do not convincingly belong to a known superfamily.

The results also illustrate some limitations. From the biological data, we believe that the heterotetrameric and homodimeric kinases constitute two separate but related subfamilies. Yet they were not separated by this analysis, likely because too few protein sequences for the homodimeric kinases are available in the database to allow a separate model for these proteins to develop. Also the estimated proportion of sequences in each subclass is a reflection of the numbers of these in the non-redundant database. For the present, these estimates have little meaning because they are biased in the direction of the interests of the community of sequencing labs. As more complete genomic sequences become available, these proportions will gain biological meaning. The results shown here are an application of the Bayesian Classifier to a set of proteins that constitute a relatively small superfamily. Future tests of the Classifier will include test cases of larger collections of proteins.

Additionally, for the cNMP-binding protein example, uninformed priors were used, i.e., the probability of each sequence belonging to each class was treated as equally likely. In many cases, however, information about a target protein's biological characteristics is available from experimental data. Therefore, for future work, informed priors will be integrated into the classification procedure, which could lead to more precise results.

Recent reports indicate that Bayesian inference is a tool of considerable value for many bioinformatics problems. The results presented here indicate that it shows good promise in the classification of protein sequence collections into subclasses.

Acknowledgements

This work is partially supported by grants DEFG0296ER62266 from DOE, 5R01HG0125702 from NIH to CEL. The authors would express thanks to the Computational Molecular Biology and Statistics Core at Wadsworth Center for Laboratories and Research.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J.D. eds. 1994. *Molecular Biology of the Cell*. 3rd ed. New York, NY: Garland Publishing, Inc.

Bairoch, A., Bucher, P., Hofmann, K. 1997. The PROSITE database, its status in 1997. *Nuc Acids Res* 25:217-221.

Casari, G., Sander, C., Valencia, A. 1995. A method to predict functional residues in proteins. *Nat Struct Biol* 2:171-178.

Fischer, H-M. 1994. Genetic regulation of nitrogen fixation in Rhizobia. *Microbio Rev* 58: 352-386.

Golding, B., and Felsenstein, J. 1990. A maximum likelihood approach to the detection of selection from a phylogeny. *J Mol Evol* 31:511-523.

Henikoff, S., and Henikoff, J.G. 1994. Position-based sequence weights. *J Mol Biol* 243:574-578.

Henikoff, J.G., and Henikoff, S. 1996. Blocks database and its applications. *Methods Enzymol* 266:88-105.

Kolb, A., Busby, S., Buc, H., Garges, S., Adhya, S. 1993. Transcriptional regulation by cAMP and its receptor protein. *Annu Rev Biochem* 62:749-795.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 235:1501-1531.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. 1993. Detecting subtle

sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208-214.

Lawrence, C.E. 1997. Tutorial, http://www.wadsworth.org/res&res/bioinfo/tut1/index.htm .

Little, R.J.A., and Rubin, D.B. 1987. *Statistical analysis with missing data.* New York, NY: John Wiley & Son.

Liu, J.S., Neuwald, A., Lawrence, C.E. 1995. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J Amer Statist Assoc* 90:1156-1170.

Liu, J.S., and Lawrence, C.E. 1996. Statistical models for multiple sequence alignment: unifications and generalizations. *Proc Amer Statist Assoc*, Statistical Computing Section, 1-8, 1996.

Liu, J.S., Neuwald, A., Lawrence, C.E. 1998. Markovian structures in biological sequence alignments. Submitted to *J Amer Statist Assoc*.

Neuwald, A., Liu, J.S., Lawrence, C.E. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science* 4:1618-1632.

Neuwald, A., Liu, J.S., Lipman, D.J., Lawrence, C.E. 1997. Extracting protein alignment models from the sequence database. *Nuc Acids Res* 25:1665-1677.

Schneider, T.D., and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nuc Acids Res* 18:6097-6100.

Sonnhammer, E.L., Eddy, S.R., Durbin, R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405-420.

Su, Y., Dostmann, W.R.G., Herberg, F.W., Durick, K., Xuong, N-h., Ten Eyck, L., Taylor, S.S., Varughese, K.I. 1995. Regulatory subunit of protein kinase A: structure of deletion mutant with cAMP binding domains. *Science* 269:807-813.

Wu, C.H. 1996. Gene classification artificial neural system. *Methods Enzymol* 266:71-88.

Zhu, J., Liu, J.S., Lawrence, C.E. 1997. Bayesian alignment and inference. *ISMB-97* 5:358-368.