Clustering Protein Sequence and Structure Space with Infinite Gaussian Mixture Models

A. Dubey, S. Hwang, C. Rangel, C.E. Rasmussen, Z. Ghahramani, and D.L. Wild

Pacific Symposium on Biocomputing 9:399-410(2004)

CLUSTERING PROTEIN SEQUENCE AND STRUCTURE SPACE WITH INFINITE GAUSSIAN MIXTURE MODELS

A. DUBEY, S. HWANG, C. RANGEL
Keck Graduate Institute, 535 Watson Drive, Claremont CA 91711, USA

C.E. RASMUSSEN
Max Planck Institute for Biological Cybernetics, Spemann Strasse 38 72076 Tuebingen, Germany

Z. GHAHRAMANI
Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London, WC1N 3AR, UK

D.L.WILD
Keck Graduate Institute, 535 Watson Drive, Claremont CA 91711, USA

Abstract

We describe a novel approach to the problem of automatically clustering protein sequences and discovering protein families, subfamilies etc., based on the theory of infinite Gaussian mixtures models. This method allows the data itself to dictate how many mixture components are required to model it, and provides a measure of the probability that two proteins belong to the same cluster. We illustrate our methods with application to three data sets: globin sequences, globin sequences with known three-dimensional structures and G-protein coupled receptor sequences. The consistency of the clusters indicate that our method is producing biologically meaningful results, which provide a very good indication of the underlying families and subfamilies. With the inclusion of secondary structure and residue solvent accessibility information, we obtain a classification of sequences of known structure which both reflects and extends their SCOP classifications.

A supplementray web site containing larger versions of the figures is available at http://public.kgi.edu/~wild/PSB04/index.html

1 Introduction

The clustering of protein sequences into families and superfamilies is a common approach for both comparative genomics and the prediction of protein function. With the advent of structural genomics projects, the clustering of protein sequences with those of known structure has also

been proposed as a method of target selection for structure determination. Newly determined protein structures must then be classified, both to assess their novelty, and in the case of proteins of unknown function, as a first step in functional annotation.

Most methods for clustering protein sequences begin with an all-against-all pairwise similarity search and use the pairwise score as a measure of similarity of the two sequences. A variety of approaches have been described to construct clusters from these scores: GENERAGE [1] uses recursive single linkage hierachical clustering, and PROTOMAP [2] constructs hierarchical clusters in a similar manner but using the means of all pairwise scores. SYSTERS[3] uses heuristics derived from set-theoretic considerations to obtain a set of disjoint clusters. Abascal and Valencia [4] describe a method for clustering protein families which uses the Ncut algorithm derived from graph theory. All these methods rely on the setting of some score theshold to distinguish members of a particular cluster from non-members, making the determination of the number of clusters arbitrary and subjective. Approaches based on single linkage hierarchical clustering can give results which are highly dependent on small changes to the data (such as adding or removing a single sequence). Moreover, non-probabilistic approaches do not provide a measure of uncertainty about the clustering, make it difficult to compute the predictive quality of the clustering and to make comparisons between clusterings based on different model assumptions (e.g. numbers of clusters, shapes of clusters, etc). Krogh et al. [5] provided an alternative probabilistic approach which used hidden Markov models (HMMs) to cluster protein sequences from the globin family into subfamilies. They fit a mixture of HMMs (which is itself a special kind of HMM) using maximum likelihood methods. The results of these experiments were promising for this particular example, yielding clusters that correspond to known globin subfamilies. Little work has followed up on this area. Methods for automatically clustering sequences into hypothesized classes will be increasingly useful as amounts of sequence and structural data continue to grow.

An important issue that must be addressed in any clustering method is the question of how many clusters to use. Bayesian statistics can provide a solution to model selection questions of this kind (e.g.[6,7]). Within the Bayesian framework, an elegant alternative approach is to assume that the data was in fact generated from an *infinite* number of Gaussian clusters. Any actual clusters in the protein sequence data will surely not be Gaussian distributed[a]. Infinite mixtures are a sensible way to capture the fact that we don't really believe that protein sequence data is well modeled by a finite number of Gaussians. An infinite Gaussian

[a] We discuss below how one can derive vectorial representations of sequences so that questions about Gaussianity are well-defined.

mixture model can readily model a finite number of non-Gaussian clusters. Finally, in an infinite Gaussian mixture model there is no need to make arbitrary choices about how many clusters there are in the data; nevertheless, after modeling one can ask questions such as how probable it is that two protein sequences or structures belong to the same cluster?

We describe a novel approach to the problem of automatically clustering protein sequences and discovering protein families, subfamilies etc. based on the theory of infinite mixtures [8]. This theory is based on the observation that the mathematical limit of an infinite number of components in an ordinary finite mixture model (i.e. clustering model) corresponds to a Dirichlet process prior [9,10,8]. Such a Dirichlet process prior allows the data itself to dictate how many mixture components are required to model it. That is, a diverse family may require several components whereas a simpler family may require only one. Although in theory the infinite mixture has an infinite number of parameters, surprisingly, it is possible to sample from these infinite mixture models efficiently since only the parameters of a few of the models need to be represented. The theory of infinite mixture models is laid out by Rasmussen[8], who showed that the procedure works effectively with mixtures of Gaussians. It has since been applied to the clustering of gene expression profiles by Medvedovic and Sivaganesan [11].

2 Infinite Gaussian Mixture Models

One commonly used computational method of non-hierarchical clustering based on measuring Euclidean distance between feature vectors is given by the k-means algorithm. However, the k-means algorithm is inadequate for describing clusters of unequal size or shape. A generalization of k-means can be derived from the theory of maximum likelihood estimation of Gaussian mixture models[12]. In a Gaussian mixture model, the data (e.g. features of protein sequences or gene expression profiles which can be arranged into p-dimensional vectors $\mathbf{y}$) is assumed to have been generated from a finite number (k) of Gaussians, $P(\mathbf{y}) = \sum_{j=1}^{k} \phi_j P_j(\mathbf{y})$ where ϕ_j is the mixing proportion for cluster j (fraction of population belonging to cluster j; $\sum_j \phi_j = 1$; $\phi_j \geq 0$) and $P_j(\mathbf{y})$ is a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j. The clusters can be found by fitting the maximum likelihood Gaussian mixture model as a function of the set of parameters $\theta = \{\phi_j, \mu_j, \Sigma_j\}_{j=1}^{k}$ using the EM algorithm [12]. Euclidean distance corresponds to assuming that the Σ_j are all equal multiples of the identity matrix.

Starting from a finite mixture model (2), we define a prior over the mixing proportion parameters ϕ. The natural conjugate prior for

mixing proportions is the symmetric Dirichlet distribution: $P(\phi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^{k} \phi_j^{\alpha/k-1}$ where α controls the distribution of the prior weight assigned to each cluster, and Γ is the gamma function.

We then explicitly include indicator variables c_i for each data point (i.e. protein sequence) which can take on integer values $c_i = j$, $j \in \{1, \ldots, k\}$, corresponding to the hypothesis that data point i belongs to cluster j. Under the mixture model, by definition, the prior probability is proportional to the mixing proportion: $P(c_i = j|\phi) = \phi_j$. A key observation is that we can compute the conditional probability of one indicator variable given the setting of all the other indicator variables after *integrating over* all possible settings of the mixing proportion parameters:

$$P(c_i = j|\mathbf{c}_{-i}, \alpha) = \int P(c_i = j|\mathbf{c}_{-i}, \phi) P(\phi|\mathbf{c}_{-i}, \alpha)\, d\phi = \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha} \quad (1)$$

where $\mathbf{c}_{-i}$ is the setting of all indicator variables except the i^{th}, n is the total number of data points, and $n_{-i,j}$ is the number of data points belonging to class j not including i. By Bayes rule,

$$P(\phi|\mathbf{c}_{-i}, \alpha) = P(\phi|\alpha)/P(\mathbf{c}_{-i}|\alpha) \prod_{\ell \neq i} P(c_\ell|\phi) \quad (2)$$

which is also a Dirichlet distribution, making it possible to perform the above integral analytically. We now can take the limit of k going to infinity, obtaining a Dirichlet Process with differing conditional probabilities for clusters with and without data: for clusters where $n_{-i,j} > 0$: $p(c_i = j|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}$, for all other clusters combined: $p(c_i \neq c_{i'} \text{ for all } i' \neq i|\mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}$. This shows that the probabilites are proportional to the occupation numbers, $n_{-i,j}$. Using these conditional probabilities one can Gibbs sample from the indicator variables efficiently, even though the model has infinitely many Gaussian clusters. Having integrated out the mixing proportions one can also Gibbs sample from all of the remaining parameters of the model, i.e. $\{\mu, \Sigma\}_j$. The details of these procedures can be found in Rasmussen (2000)[8].

We have used infinite Gaussian mixtures to model protein sequence data with the intention of answering queries of the kind: what is the probability that two proteins belong to the same cluster? Unlike previous methods based on a single clustering of the data, this approach computes this probability while taking into account all sources of model uncertainty (including number of clusters and location of clusters). We use the probability p_{ij} that two proteins i and j belong to the same cluster in the infinite mixture model as a measure of the similarity of these protein sequences. Conversely $1 - p_{ij}$ defines a dissimilarity measure

which for the purposes of visualization can be input to one of the standard linkage algorithms used for hierarchical clustering (see Figure 3). We illustrate our methods with application to three data sets: globin sequences, globin sequences with known three-dimensional structures and G-protein coupled receptor sequences.

3 Methods

To be able to cluster protein sequences, we need to be able to obtain a vector representation of the protein in a suitable metric space. We use the Fisher score vector respresentation described by Jaakkola et al [13], which provides an appropriate measure of similarity between sequences. The Fisher score vector for a particular protein X is obtained by evaluating the derivative of the log-likelihood with respect to a vector of parameters (θ) of a hidden Markov model (HMM) trained on the set of protein sequences: $U_X = \nabla_\theta \log P(X|\theta)$. Each component of the vector U_X is the derivative of the log-likelihood for the sequence X with respect to a particular parameter (the emission probabilities of the HMM).

In the work described below, we first train an HMM on the set of protein sequences of interest and then calculate a Fisher score vector as described above. In the case of sequences of known structure, we use the Bayesian network model of Raval et al. [14], which can be thought of as an extension of a hidden Markov model to incorporate multiple observations of primary sequence, secondary structure and residue solvent accessibility, calculated from the three-dimensional coordinates by the DSSP method of Kabsch and Sander [15]. For all data sets the dimensionality of the Fisher score vector was then reduced by principal components analysis and we used this reduced dimension vector as the $\mathbf{y}$ vector input into the infinite Gaussian mixture model. We used the first 10 principal components, which captured most of the variance in the U_X vectors. The mixture model was initialized with all data belonging to a single Gaussian, and a large number of Gibbs sampling sweeps are performed, updating all variables and parameters, i.e. $\{\{\mu_j, \Sigma_j\}, \{c_i\}, \alpha\}$, in turn by sampling from the conditional distributions derived in the previous sections and described in more detail in Rasmussen (2000)[8]. We typically run the chain for 110,000 iterations, discarding the initial 11,000 steps as "burn-in" and keeping every 1000th step after that, generating 100 roughly independent samples from the posterior distribution.

4 Results

4.1 Globin Sequences

The mixture of HMMs method of Krogh et al [5] discovered 7 clusters in a set of 628 globin sequences, corresponding to:

1. **Class 1** 233 sequences: principally all α, a few ζ (an α-type chain of mammalian embryonic hemoglobin), $\pi/\pi^{'}$ (the counterpart of the α chain in major early embryonic hemoglobin P), and $\theta - 1$ chains (early erythrocyte α-like).
2. **Class 2** 232 sequences: almost all β, a few δ (β-like), ϵ (β-type found in early embryos), γ (comprises fetal hemoglobin F in combination with two α chains), ρ (major early embryonic β-type chain) and θ chains (embryonic β-type chain).
3. **Class 3** 71 myoglobins.
4. **Class 4** 58 sequences. The 13 highest scoring in this cluster were leghemoglobins. This class contained a variety of sequences including 3 non-globins in the original data set.
5. **Class 5** 19 sequences. Midge globins.
6. **Class 6** 8 sequences. Globins from agnatha (jawless fish).
7. **Class 7** 7 sequences. varied.

Our results, using an updated version of the same data set (630 globin sequences, distributed with the HMMER2 software package) are shown in Figure 1. In this plot we show the number of times, out of 100 samples, that the indicator variables for two sequences were equal. As shown above, this may be interpreted as the probability p_{ij} that two proteins i and j belong to the same cluster. It is evident that our model has discovered a larger number of clusters that the method of Krogh et al.[5]. The granularity of this clustering is determined by the data and not by some user-defined threshold. Large solid blocks of color along the diagonal correspond to homogeneous clusters. Note that in our method, sequences may belong to more than one cluster with a defined probability: off-diagonal elements indicate 'cross-clustering'. For comparison, we also clustered the sequences using BLASTCLUST, which clusters the sequences according to a sequence identity threshold and a single linkage algorithm. With a 90% sequence identity threshold, 261 clusters were obtained. The first large homogeneous cluster in Figure 1 (bottom right hand corner) comprises 37 hemoglobin β sequences plus two δ sequences (HBD_COLPO and HBD_PANTR) (Figure 1). Although a number of these sequences are contained within the same cluster in the BLASTCLUST output, indicating that they have $> 90\%$ sequence identity, we note that the clusters are by no means identical.

The BLASTCLUST cluster containing many of these hemoglobin β sequences also contains 8 hemoglobin δ sequences and one Hemoglobin β-2 chain (HBB2_PANLE). Figure 1 indicates that all sequences within this cluster also 'cross-cluster' with another group of β sequences with a probability of around 20-30%. The next cluster from the bottom right (Figure 1) contains all α sequences and cross clusters with another group of α sequences with a probability of around 40-50%. Although a detailed analysis of these results is beyond the scope of this paper, we identify at least 11 distinct α and 13 distinct β clusters (plus some additional smaller ones). Although some of the variant sequences cluster with α and β sequences, we identify a number of clusters composed only of variant sequences: 3 clusters comprising only γ, ϵ and θ sequences, one cluster of δ and one cluster of ζ sequences. We identify 3 distinct clusters of leghemoglobins and 1 cluster of midge hemoglobins (6 sequences), a small cluster of fish hemoglobins and a small cluster comprising clam and earthworm sequences. Myoglobins, which Krogh et al (1994) found in one cluster, form 10 distinct clusters, mainly comprising proteins from related species. BLASTCLUST groups these into 6 clusters plus 9 singletons at a 90% identity theshold. We identify only 11 singletons (proteins which never cluster with another), none of which are myoglobins. The largest cluster comprises 40 hemoglobin beta sequences.

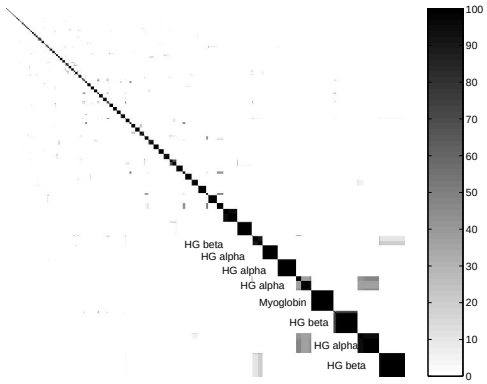

Figure 1: Clustering of the 630 globin sequences. The gray scale indicates the number of times, out of 100 samples, that the indicator variables for two sequences were equal, or the probability that two sequences belong to the same cluster

These results indicate that our method is capable of producing biologically meaningful results and correctly classifies the main globin subfamilies. In addition, it provides a finer level of clustering within these subfamilies than either the use of BLAST alignments and sequence identity or the method of Krogh et al.[5]

4.2 Globin Sequences of Known Structure

For this experiment we obtained globin sequences from the Strucural Classification of Proteins (SCOP) database [16] using the ASTRAL resource [b]. Sequences with $> 95\%$ sequence identity were excluded, leaving 91 proteins. According to the SCOP classification, these conprised representatives of 4 globin structural subfamilies (a.1.1.1: truncated hemoglobins (4 sequences) , a.1.1.2: glycera globins, myoglobins, hemoglobin I, flavohemoglobins, leghemoglobins, hemoglobin α and β chains, a.1.1.3: phycocyanins, allophycocyanins, phycoerythrins and a.1.1.4: nerve tissue mini-hemoglobin (1 sequence)). The sequences were clustered using feature vectors derived from two models: a sequence-only HMM and a Bayesian net model (structural HMM). The results are shown in Figure 2 and Figure 3.

The results from the sequence only clustering (Figure 2 left) show a similar pattern to those obtained with the 630 globin sequences. Fairly homogeneous clusters are mainly composed of related sequences, eg: β hemoglobin chains, α hemoglobin chains, myoglobins, phycocyanin a and b, phycoerythrin and b and allophycocyanin a and b chains (which all form separate clusters). Glycera globins form a separate cluster, as do leghemoglobins. Three or four heterogeneous (loosely associated) clusters are observed, which include truncated hemoglobins, hemoglobin I's, dehaloperoxidase etc.

The results from the model which includes secondary structure and residue accessibility information shows fewer clusters; 12 in all, plus two singletons (dehaloperoxidase and pig roundworm hemoglobin, domain 1) (Figure 2 right). Again α and β hemoglobin chains form distinct and fairly homogeneous clusters, as do the myoglobins, with the exception of 1MYT (this is a myoglobin which lacks the D helix), which clusters more strongly with β hemoglobins, as well as weakly with the myoglobin cluster, and 1MBA (a mollusc myoglobin), which clusters with clam hemoglobins and glycera globins from bloodworms. Phycocyanins, allophycocyanins and phycoerythrins (which are all classified by SCOP into the same subfamily a.1.1.3) form two distinct large joint clusters. Within these clusters one can detect subfamilies corresponding to the allophycocyanins, phycoerythrins and phycocyanins, which cluster amongst themselves with a higher probability. Leghemoglobins cluster strongly with a single non-symbiotic plant hemoglobin from rice, and weakly with a clam hemoglobin I. Truncated hemoglobins, which SCOP classifies into a different subfamily (a.1.1.1), form two distinct clusters, and the sole member of subfamily a.1.1.4 (nerve tissue mini-hemoglobin), clusters with 1CH4 (chimeric synthetic hemoglobin beta-alpha). In comparison, 13 clusters are produced with BLASTCLUST only at a 29% sequence

[b] http://astral.stanford.edu

identity threshold or lower. These comprise a single cluster for a.1.1.1, nine separate clusters for a.1.1.2 (including 4 singletons), a single cluster for a.1.1.3 and a singleton for a.1.1.4. Our results, which do not require a predefined threshold to be specified, provide a reflection the underlying SCOP classifications, but the biologically meanigful sub-clusters also suggest that a further level of subfamily subdivision is possible.

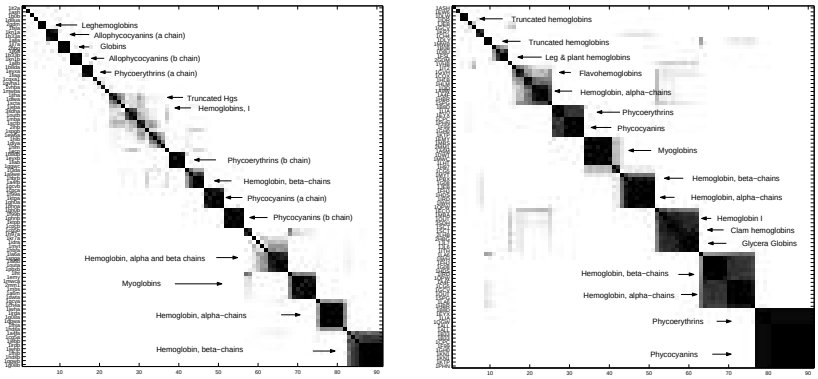

Figure 2: Clustering of the 91 SCOP globin sequences:left, by sequence information only; right, with the inclusion of structural information. Sequence labels on the y-axis are ordered optimally for each plot.

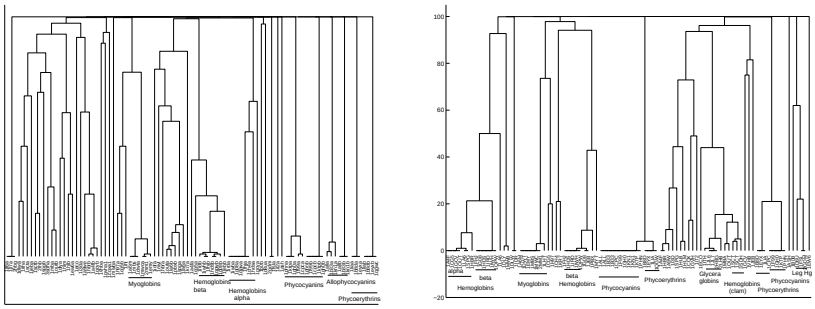

Figure 3: Dendrogram resresentation of the clustering of the 91 SCOP globin sequences shown in Figure 2: left, by sequence information only; right, with the inclusion of structural information.

4.3 G-Coupled Protein Receptors (GPCRs)

According to the GPCRDB classification system [17], the G-protein coupled receptor (GPCR) superfamily is classified into 5 major classes: Class A (related to rhodopsin and adrenergic receptors), Class B (related to

calcitonin and PTH/PTHrP receptors), Class C (related to metatropic receptors), Class D (related to pheromone receptors) and Class E (related to cAMP receptors). The classes share $\geq$ 20% sequence identity over predicted transmembrane helices [17]. Each class is further divided into level 1 subfamilies (eg: Amine, Peptide, Opsin etc. for Class A) and further into Level 2 subfamilies (Muscarinic, Histamine, Serotonin etc. for the Amine subfamily). A number of putative GPCRs have no identified natural ligand and are dubbed 'orphan' receptors. The sequence diversity of the GPCR classes makes subfamily classification a challenging problem. The problem of recognizing GPCR subfamilies is compounded by the fact that the subfamily classifications in GPCRDB are defined chemically (that is, according to the differential binding of ligands to the receptors) and not necessarily by either sequence similarity or the post ligand-receptor binding pathways.

A number of other authors have described computational approaches to classifying GPCRs. Karchin et al [18] trained 2-class support vector machines (SVMs) using Fisher score vectors derived from HMMs [13]. Joost and Methner [19] used a phylogenetic tree constructed by neighbor joining with bootstrapping. Lapinsh et al [20] translated amino acid sequences into vectors based on the physicochemical properties of the amino acids and used and autocross-covariance transformation followed by principal components analysis (PCA) to classify GPCRs.

For our experiments, sequences were obtained from the GPCRDB database [17] [c]. Because of the smaller number of sequences in Classes B-E, we have focussed our analysis of Class A sequences. Our dataset comprised 946 sequences, of which 303 were "orphan" receptors, with no family classification. A portion of the clustering results using the infinite Gaussian mixture model are shown in Figure 4. Because of the sequence diversity of this superfamily, a larger number of smaller clusters are evident around the diagonal than were observed with the globin sequences. Most of the homogeneous clusters (solid color) comprise sequences from the same subfamily (level 3 in the GPCRDB hierarchy), and appear to be orthologs of the same protein from related species. Whilst a detailed analysis of these is beyond the scope of the present paper, as an illustration, we note that the largest cluster (bottom right hand corner), comprises Rhodopsin (Rhodopsin Vertebrate type 1) sequences from mammals and reptiles (plus lamprey), whilst the second cluster is composed entirely of fish Rhodopsins. Some unexpected associations also appear. Although in some case our results indicate assignments for certain orphan receptors which agree those of the authors cited above, in other cases our predictions are novel. A detailed analysis of these will be published in an extended version of this paper.

[c] http://www.gpcr.org

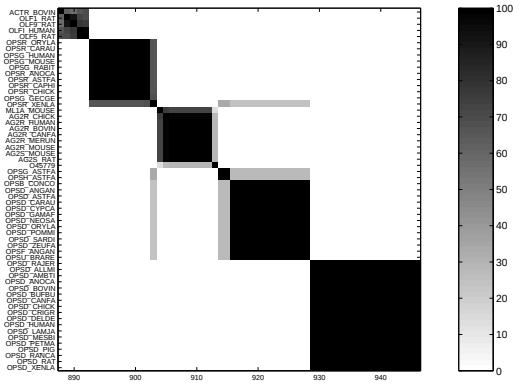

Figure 4: Part of the clustering of the GPCR Class A sequences.

5 Discussion

The consistency of the clusters we obtain with a well annotated superfamily of protein sequences such as the globins gives us confidence that our method is producing biologically meaningful results, which provide a very good indication of the underlying families and subfamilies. Homogeneous clusters tend to consist of orthologs of the same protein and paralogs appear to be separated into distinct clusters. This pattern appears to be repeated in our clustering of the GPCR sequences, with the potential of providing functional annotations for certain orphan receptors. Whilst some of these agree with predictions derived from neighbor-joining phylogenetic trees and principal component analysis, a number are novel. In all cases, our method provides a finer level of granularity than the method of Lapinsh et al. [20], clustering orphan receptors with members of particular GPCRDB subfamilies, rather than a broad family classification. With the inclusion of secondary stucture and residue solvent accessibility information in the HMM on which our method is based, the clustering of the SCOP globin sequences changes from a large number of small clusters of functionally related sequences to a smaller number of clusters, in which the members of the SCOP globin families are clearly separated. However, once again we achieve an even finer level of classification, clearly separating α, β and myoglobins, as well as other members of SCOP class a.1.1.2. This suggests that our method also has the potential to provide a novel automated method for the structural classification of proteins. In order to achieve a large scale clustering of sequence or structure space we will investigate the use of Fisher scores obtained from from a "mixture model" which combines individual mod-

els for different superfamilies as described in [14].

Acknowledgments

This work was supported by the National Institutes of Health (NIH) under Grant Number 1 P01 GM63208. CER was supported by the German Research Council (DFG) through grant RA 1030/1.

References

1. A.J. Enright and C.A. Ouzounis, Bioinformatics **16**, 451-457 (2000)
2. G. Yona and N. Linial and M. Linial, Proteins **37**, 360-378 (1999)
3. A. Krause and M. Vingron, Bioinformatics **14**, 430-438 (1998)
4. F. Abascal and A. Valencia, Bioinformatics **18**, 908-921 (2002)
5. A. Krogh A and M. Brown and I.S. Mian and K. Sjolander and D. Haussler, J. Mol. Biol. **235**, 1501-1531 (1994)
6. Y. Barash and N. Friedman,J. Comput. Biol. **9**, 161-191 (2002)
7. S. Richardson and P. Green (1997), J. Roy. Stat. Soc. **B59**, 731-792 (1997)
8. C. E. Rasmussen in *Advances in Neural Information Processing Systems 12*, ed. S. A. Solla, T. K. Leen, and K.-R. Muller (MIT Press, 2000)
9. C.E. Antoniak, Annals of Statistics **2**, 1152-1174 (1974)
10. R. M. Neal, J. Comp. and Graphical Statistics **9**, 249-265 (2000)
11. M. Medvedovic and S. Sivaganesan, Bioinformatics **18**, 1194-1206 (2002)
12. G. McLachlan and D. Peel,*Finite Mixture Models*,(Wiley, New York, 2000).
13. T. Jaakkola and M. Diekhans and D. Haussler J Comput Biol. **7**, 95-114 (2000)
14. A. Raval and Z. Ghahramani and D.L. Wild, Bioinformatics **18**, 788-801 (2002)
15. W. Kabsch and C. Sander Biopolymers **22**, 2577-2637 (1983)
16. A.G. Murzin and S.E. Brenner and T. Hubbard and C. Chothia,J. Mol. Biol. **247**, 536-540 (1995)
17. F. Horn and J. Weare and M.W. Beukers and S. Hoersch and A. Bairoch and W. Chen and O.Edvardsen and F. Campagne and G. Vriend,Nucleic Acids Res. **26**, 277-281 (1998)
18. R. Karchin and K. Karplus and D. Haussler, Bioinformatics **18**, 147-159 (2002)
19. P. Joost and A. Methner, Genome Biol. **3**, RESEARCH0063 (2002)
20. M. Lapinsh and A. Gutcaits and P. Prusis and C. Post and T. Lundstedt and J.E. Wikberg, Protein Sci. **11**, 795-805 (2002)