

A Bayesian Approach to Clustering Protein Sequence and Structure Space

David Wild

Keck Graduate Institute

February 17, 2004

Outline

- Bayesian Learning and Model Selection
- Clustering Protein Sequences *(using Infinite Mixtures)*
- Clustering Protein Structures *(using Infinite Mixtures)*
- Conclusions and Future Work

Bayesian Learning

Consider a data set \mathcal{D} , and a model m with parameters θ .

Likelihood of model parameters for data set \mathcal{D} : $p(\mathcal{D}|\theta, m)$

Prior over model parameters: $p(\theta|m)$

Prior over model class: $p(m)$

The likelihood and parameter priors are combined into the posterior for a particular model using [Bayes Rule](#):

$$p(\theta|\mathcal{D}, m) = \frac{p(\mathcal{D}|\theta, m)p(\theta|m)}{p(\mathcal{D}|m)}$$

Predictions are made by integrating over the posterior

Bayesian Model Comparison

A data set \mathcal{D} , and a model m with parameters θ .

Likelihood of model parameters for data set \mathcal{D} : $p(\mathcal{D}|\theta, m)$

Prior over model parameters: $p(\theta|m)$

Prior over model class: $p(m)$

To compare models m and m' , we again use Bayes' rule and the prior on models

$$\frac{p(m|\mathcal{D})}{p(m'|\mathcal{D})} = \frac{p(\mathcal{D}|m) p(m)}{p(\mathcal{D}|m') p(m')}$$

This also requires an integral over θ :

$$p(\mathcal{D}|m) = \int d\theta p(\mathcal{D}|\theta, m) p(\theta|m)$$

For interesting models, these integrals may be difficult to compute. *Approximations.*

Clustering Protein Sequences and Structures using Infinite Gaussian Mixture Models

Approaches to Clustering Protein Sequences

- Useful for target selection in structural genomics experiments
- Pairwise sequence alignment scores and hierarchical clustering (single linkage)
 - BLASTCLUST (NCBI)
 - PROTOMAP (1999) Yona, Linial and Linial
 - SYSTERS (1998) Krause and Vingron
 - GENERAGE (2000) Enright and Ouzounis
- All require setting threshold to distinguish cluster members from non-members
 - Krogh et al. (1994) Mixture of HMMs - requires number of clusters to be specified

Results - Globin Sequences by Krogh et al

The mixture of HMMs method of Krogh et al:

1. **Class 1** 233 sequences: principally all α , a few ζ (an α -type chain of mammalian embryonic hemoglobin), π/π' (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.

Finite Gaussian Mixture Models

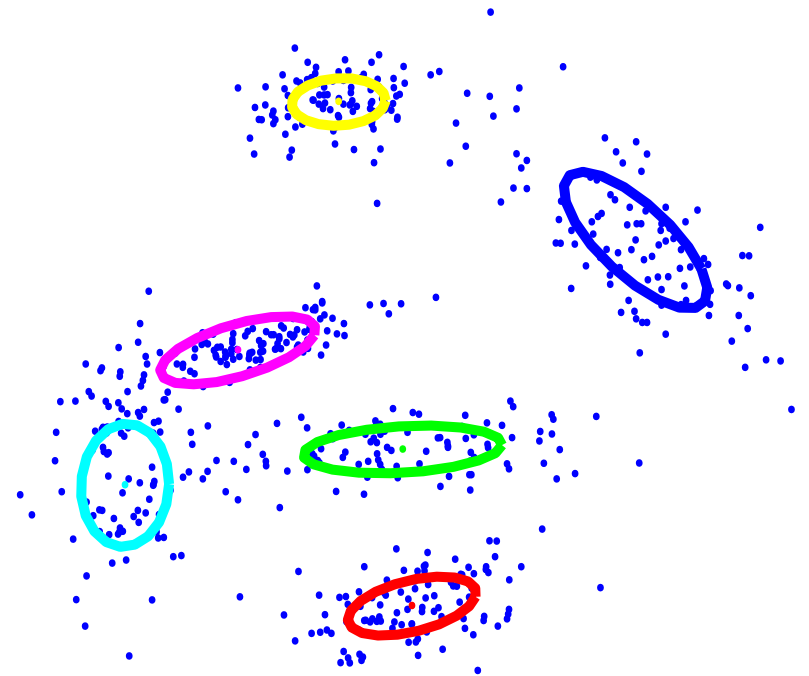
- **Data:** features of protein sequences or gene expression profiles which can be arranged into p -dimensional vectors \mathbf{y} ;
- **Model:** a Gaussian mixture model with a finite number (k) of Gaussians, with parameters ϕ_j , $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$:

$$P(\mathbf{y}) = \sum_{j=1}^k \phi_j P_j(\mathbf{y})$$

ϕ_j : the mixing proportion for cluster j

$$(\sum_j \phi_j = 1; \phi_j \geq 0)$$

$P_j(\mathbf{y})$: a Gaussian with mean $\boldsymbol{\mu}_j$ and cov matrix $\boldsymbol{\Sigma}_j$



Infinite Gaussian Mixture Models

- We can choose a flexible model with *infinitely* many components, as long as we do Bayesian average over parameters! The data automatically tells us how many components we need.
- Let $c_i = j$ denote “data point i belongs to cluster j ”
- Before observing \mathbf{y} : $P(c_i = j | \phi) = \phi_j$, proportional to the mixing proportion;
- Choose prior over ϕ_j to be the symmetric Dirichlet distribution, where α controls

how evenly mass is spread between clusters:
$$P(\phi | \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^k \phi_j^{\alpha/k-1}$$

Infinite Gaussian Mixture Models

- *Integrating over* all possible settings of ϕ we get the conditional probability:

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

- In the extreme case $k \rightarrow \infty$ we get infinite Gaussian mixtures (a.k.a. Dirichlet process mixtures).

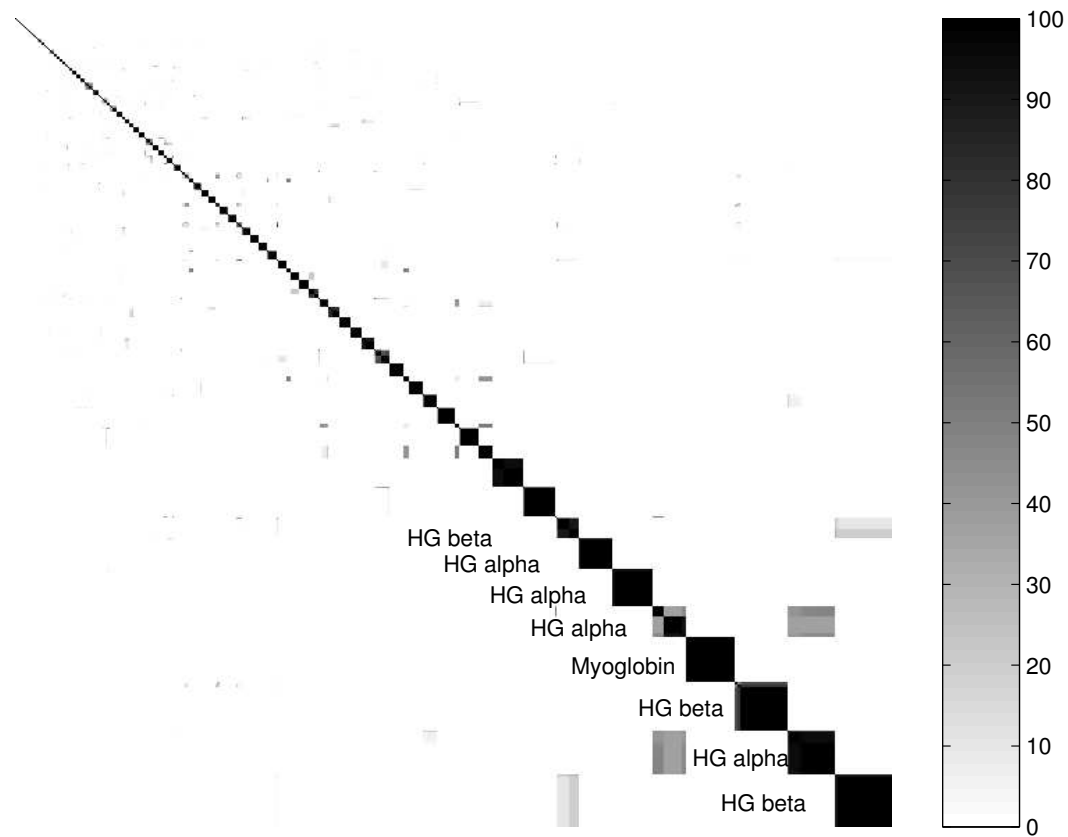
Infinite Gaussian Mixture Models

- **Goal:** to infer $P(c_i = c_\ell | \mathbf{y}_1, \dots, \mathbf{y}_n, \alpha) = p_{i\ell}$: “the probability that proteins i and ℓ belong to the same cluster”
- $1 - p_{i\ell}$ represents a distance that can be input into linkage algorithm for hierarchical clustering

Methodology

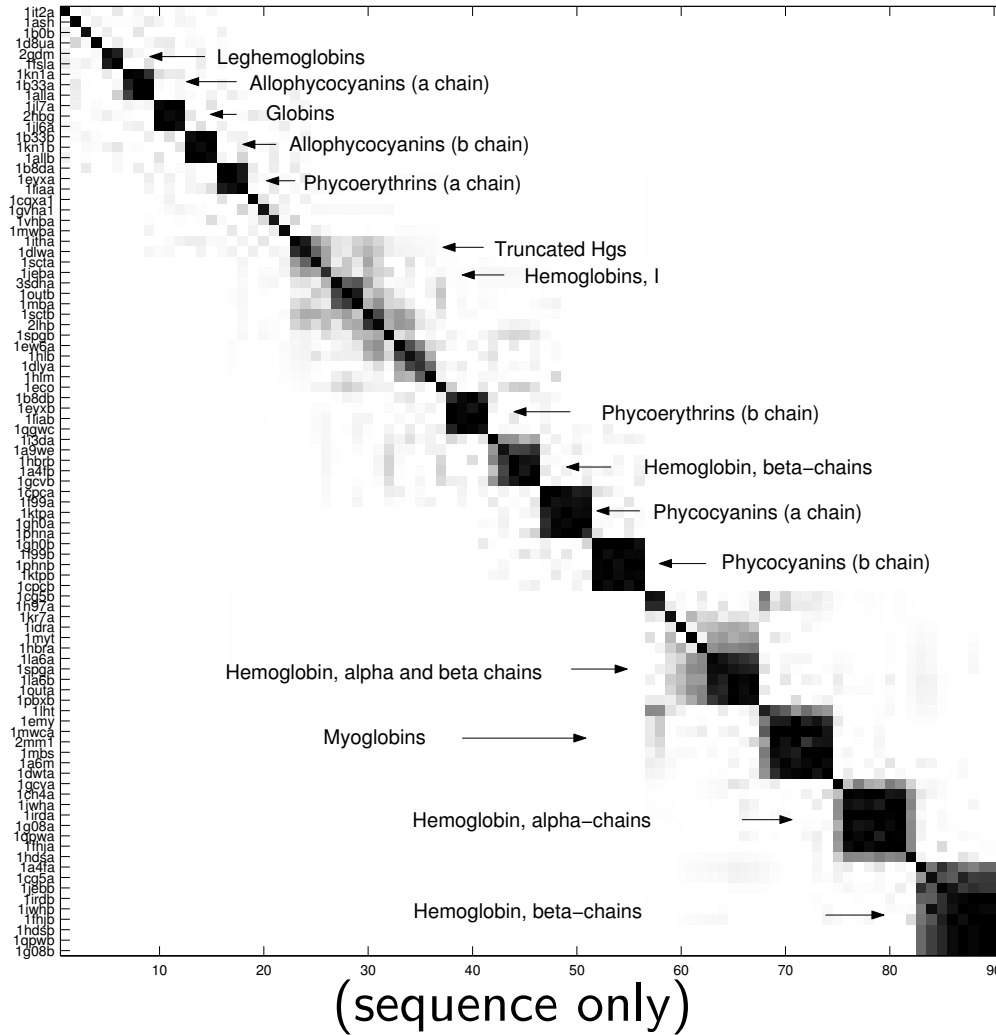
- vector representation: Fisher score vector representation described by Jaakkola et al (2000): $U_X = \nabla_{\theta} \log P(X|\theta)$;
- structural information included using structure-based hidden Markov models (Raval et al. 2000);
- principal components analysis: the dimensionality of the Fisher score vector was reduced to 10;
- infinite Gaussian mixture models: a large number of Gibbs sampling sweeps are performed.

Results - Globin Sequences

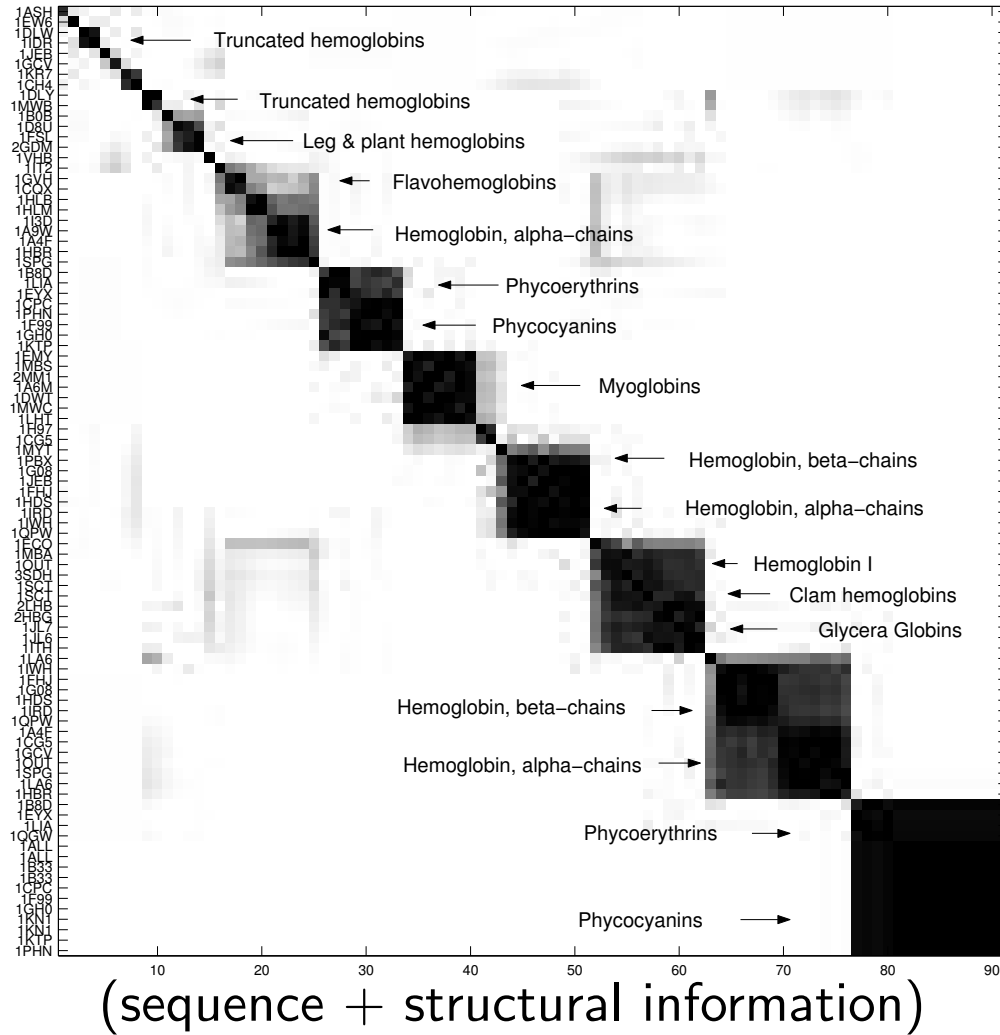


Finds larger number of clusters including biologically meaningful sub-clusters than Krogh et al.

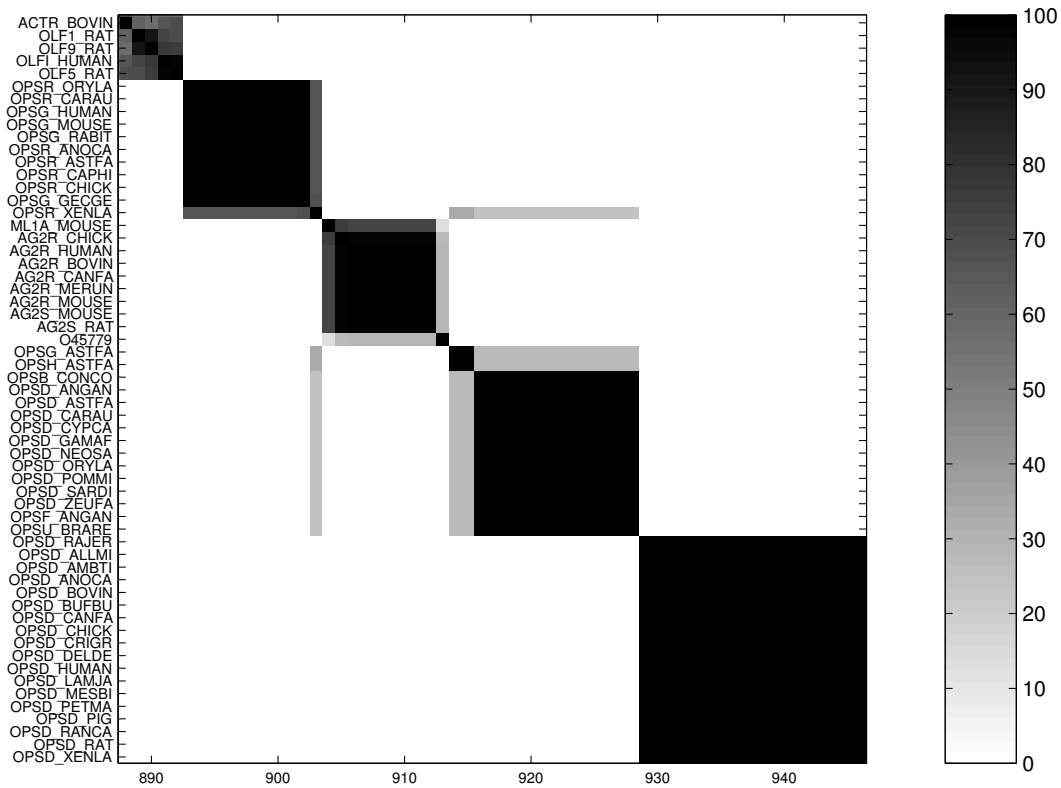
Results - SCOP Globin Sequences



Results - SCOP Globin Sequences



Results - GPCR Sequences

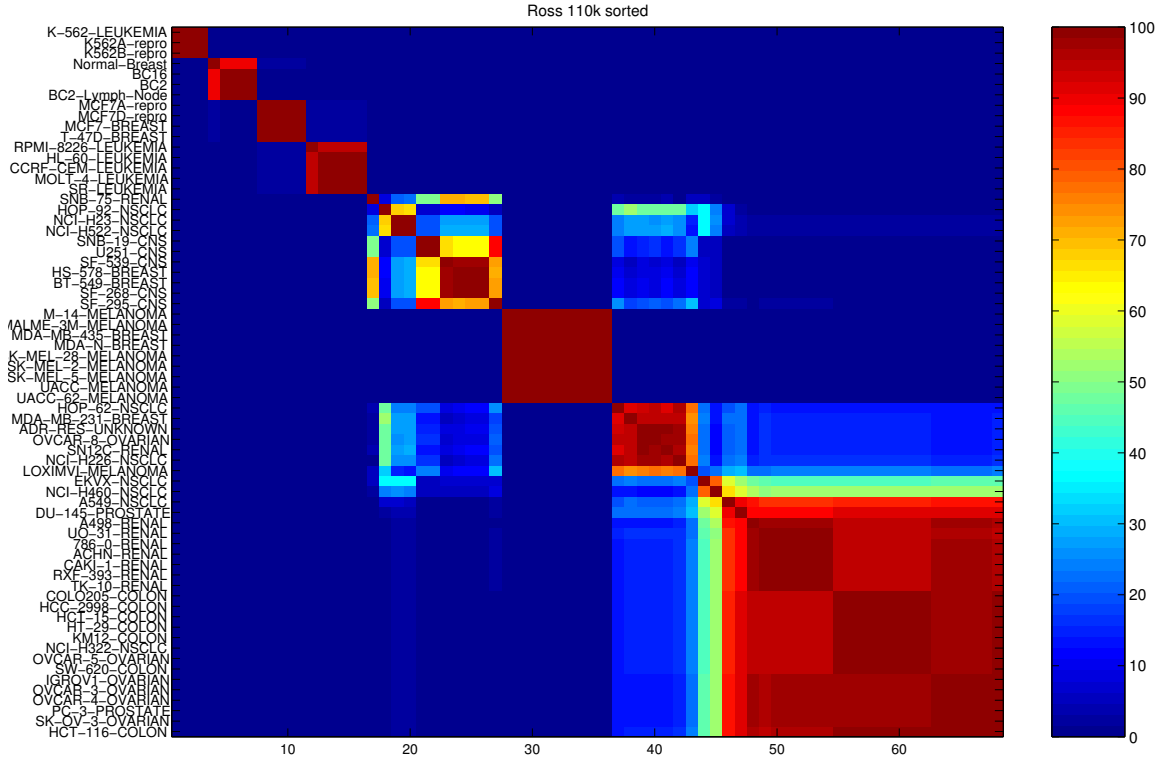


(sequence only)

Some are orphan receptors with no known function
 We can use the clustering to suggest putative functions.

NCI60 microarray gene expression data

We've also used infinite mixtures for clustering microarray gene expression data as a way of modeling uncertainty.



Conclusions

- Bayesian methods allow large-scale statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- Clustering protein sequences using infinite Gaussian mixture models produces biologically meaningful clusters and sub-clusters
- The data itself determines the optimal number of clusters
- Inclusion of secondary structure and residue accessibility information reflects and extends the SCOP classification
- Infinite mixtures also provide a good solution for clustering microarray gene expression data

Future Work

Extension to Superfamily clustering:

- Compute Fisher scores from a 'mixture model' M^* , combining HMMs for the different superfamilies:

$$P(X|M^*) = \sum_{i=0}^N \pi_i P(X|\theta_i)$$

where the mixing proportions π_i are the prior probabilities of superfamily i

- The Fisher score vector for a particular protein X is given by

$$\frac{\delta \log P(X|M^*)}{\delta \pi_k} = \frac{r_k}{\pi_k} - 1$$

where the r_k are the previously calculated posterior probabilities of X , i.e $r_k = P(M_k|X)$.

Acknowledgements

Collaborators:

Keck Graduate Institute of Applied Life Sciences:

[Ananya Dubey](#), [Seungwoo Hwang](#), [Claudia Rangel](#),

Gatsby Computational Neuroscience Unit, University College London, UK:

[Zoubin Ghahramani](#),

Max Planck Institute for Biological Cybernetics, Tübingen, Germany:

[Carl E. Rasmussen](#),

This work is supported by the National Institutes of Health (NIH) and its National Institute of General Medical Sciences (NIGMS) division under Grant Number 1 P01 GM63208.