Prediction of functional residues and clustering for protein families using mixtures

Benjamin Georgi[1], Jörg Schultz[2], Alexander Schliep[1]

[1]Max-Planck-Institue for Molecular Genetics
[2]University of Würzburg

Email: georgi@molgen.mpg.de

Abstract

Protein families can be divided into subgroups with similar but distinct functionality. The analysis of these subgroups and the determination which residues convey substrate specificity is a central question in the study of these families. We present a clustering procedure for simultaneous inference of subgroups and prediction of these functional residues based on multiple sequence alignments of protein families. If prior knowledge on functional subfamilies is available it can be integrated into the clustering in a semi-supervised setup. Application of the method on several well studied families revealed a good clustering performance and ample biological support for the predicted positions.

Functional Residues

Example multiple sequence alignment (MSA) for two subfamilies **F1** and **F2** and three functional residues:

F1	R	G	S	H	L	N	T	V	K	.	V	N
	F	G	A	D	L	N	K	H	K	.	V	S
	Y	G	A	V	L	.	.	G	A	.	V	T
	V	G	L	H	L	P	A	V	N	Q	V	A
F2	V	R	V	A	T	G	A	A	.	Q	I	E
	L	R	A	L	T	S	N	L	.	G	I	D
	I	R	A	L	T	L	S	P	Q	S	I	K

In a MSA of protein sequences functional residues between families **F1** and **F2** can be recognized by a strong signal of subgroup specific conservation, indicated by different colors in the Figure.

Method

In the general case where the subgroup membership of the sequences is unknown, clustering can be performed to obtain a partition of the data which can be used as basis for functional resiude prediction.

Clustering method:
We employ the *context-specific independence* (CSI) mixture framework to obtain such a clustering. The CSI formalism augments the mixture by learning of a model structure which adapts model complexity to the degree of variability found in the data. A mixture model for a random variable X representing a MSA is defined as

$$P(X \mid \Theta) = \sum_{i=1}^{K} w_i f_i(X, \theta_i)$$

$$\text{s.t.} \sum_{i=1}^{K} w_i = 1, \; w_i \geq 0 \qquad \Theta = (\{w_i\}_{i=1..K}, \{\theta_i\}_{i=1..K})$$

Residue ranking:
After the clustering is completed, we can use the mixture to make predictions for functional residues by ranking the positions of the alignment by their importance for the characterization of a cluster. This ranking is obtained by computing an entropy-based score for each position in the alignment. The highest scoring positions are then taken as putative functional residues. The score for a position j for cluster i is given by

$$KL_{sym}(\theta_{ij}, \theta_0) = \frac{KL(\theta_{ij}, \theta_0) + KL(\theta_0, \theta_{ij})}{2}$$

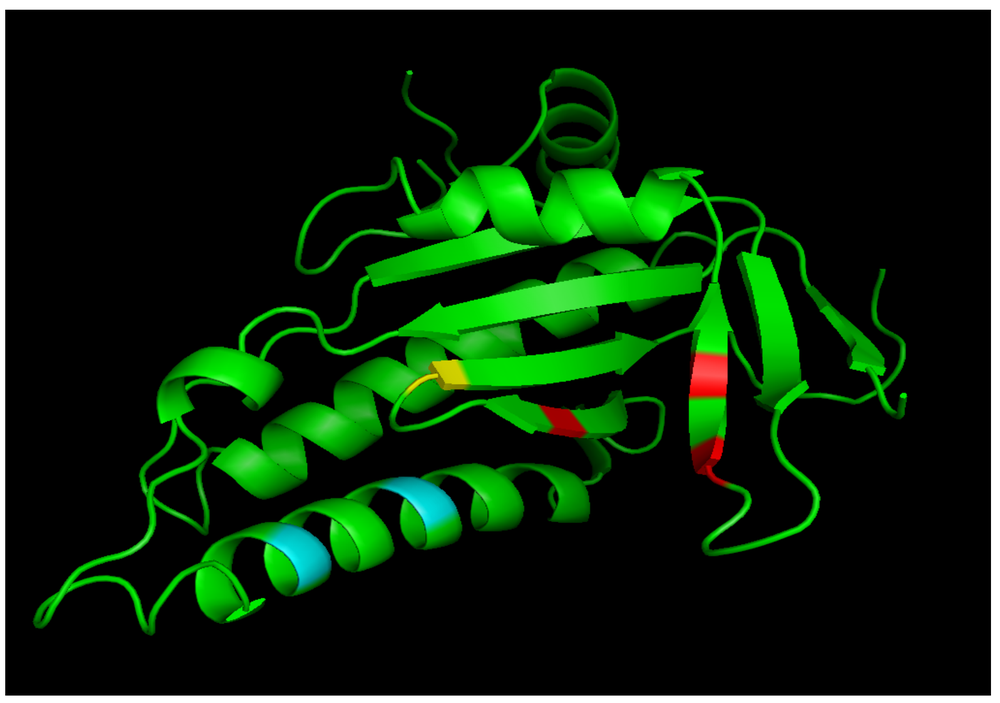
Where KL is the relative entropy, θ_{ij} are the parameters of position j in cluster i and θ_0 are the parameters for position j computed on the whole data set.

Semi-supervised learning:
For many data sets there are class labels, i.e. functional annotations for at least a subset of proteins in an alingment. By integrating this information into the parameter learning of the mixture, the clustering performance can potentially be increased considerably.

Results

Clustering & residue ranking:
We applied CSI mixture clustering on a data set of nucleotidyl cylcase sequences. The two subgroups have specificity for either guanine or adenine. The separation into guanylyl (GC) or adenylyl (AC) cyclases was recovered with an accuracy of 85% by the clustering.

The features were ranked with the entropy criterion and the top scoring positions mapped onto the C2 domain of the AC from rat (pdb: 1AB8)

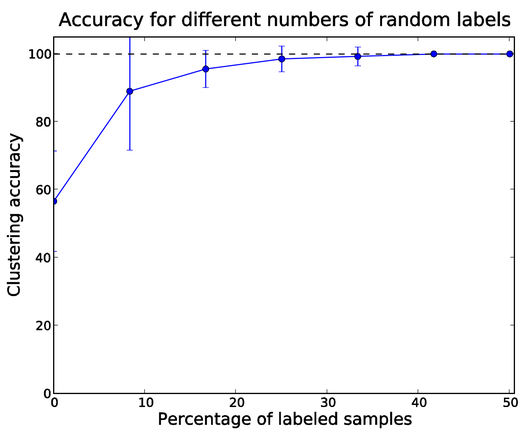


Among the ten highest scoring positions we found:
Positions 1018, 1016 and 938: experimentally confirmed specificity determining residues
Positions 911 and 919: Part of C1/C2 domain interface
Position 943: forsoklin interaction site

This abundance of biological annotation for the predicted positions as well as the performance of the clustering method shows the usefulness of the approach for the analysis of protein families.

Semi-supervised learning:

For a data set of α / β crystalins the unsupervised clustering had an accuracy of 57% with respect to α / β separation.
When adding prior knowledge in form of different amounts of randomly selected labels, the clustering accuracy increases considerably. From the variance in accuracy (esp. for only few labels) it can be seen that the choice of labels has an non-negligible impact on performance.



Crystalin data