Supporting Material
Experiment 1

Spectral Clustering of Protein Sequences

Alberto Paccanaro, James A. Casbon and Mansoor A. S. Saqi

EXPERIMENT 1

The dataset used for this experiment consists of 511 sequences belonging to 7 super-families, namely Globin-like (88), Cupredoxins (78), Viral coat and capsid proteins (106), Trypsin-like serine proteases (73), FAD/NAD(P)-binding domain (64), MHC antigen-recognition domain (51), Scorpion toxin-like (51).

The results obtained applying to this dataset our implementation of GeneRAGE, the Hierarchical Clustering algorithm, TribeMCL and our spectral clustering algorithm are shown in the four diagrams of the figure.
The diagrams should be read as follows. Each row corresponds to a different cluster. Short (green) bars represent the assignment of each protein sequence to a cluster. Each protein has one of these bars in only one of the rows (clusters); the presence of the bar means that the protein is assigned to that cluster. Boundaries between super-families are shown by vertical thick (red) lines; boundaries between families within each super-family are shown by dotted (blue) lines. If an algorithm returned more than 30 clusters, only the most populated 30 are shown. The order of the super-families in each diagram from left to right is as given above.

The spectral clustering method clearly outperforms the other three. First of all, it detects a number of clusters which is close to the correct number of super-families, since it detects 11 clusters; at the same time, our implementation of GeneRAGE detects 121 clusters, the hierarchical clustering detects 153 clusters, and TribeMCL 28 (with inflation parameter set to 1.55). The better quality of the clustering is quantified by the F-measure: for the spectral clustering it is equal to 0.8181; our implementation of GeneRAGE has a score of 0.5376, the hierarchical clustering 0.4188 and TribeMCL 0.5188.

As expected from our analysis in section 2.1, the hierarchical clustering and our implementation of the GeneRAGE algorithm tend to cluster the sequences more at the family level, and all the clusters are pure.

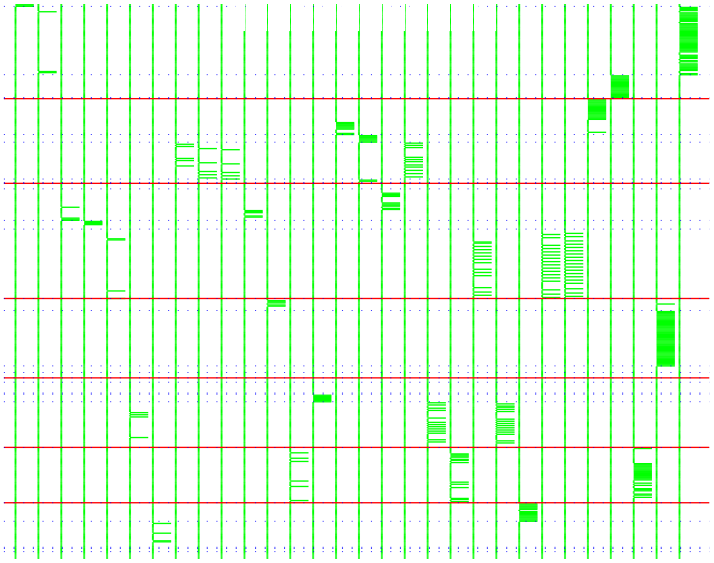
All four methods split the globin-like super-family into two predominant clusters, one containing globins and the other phycocyanins. The hierarchical method and TribeMCL showed some further splitting of the phycocyanin family.

We observe that the spectral clustering method successully groups together all the members of the 'trypsin-like serine protease' super-family which in our datset is composed of four families. The resulting cluster for the super-family is not entirely pure and includes 6 sequences from the viral coat and capsid super-family as well. TribeMCL also groups almost all the 'trypsin-like serine protease' super-family sequences in a single cluster although this cluster is contaminated to a greater extent than the corresponding cluster obtained using spectral clustering, containing several sequences from 5 other super-families. As expected our implementation of GeneRAGE splits this super-family leading to clusters representing commonality at the family rather than super-family level. The hierarchical method also shows some clustering at the family level.

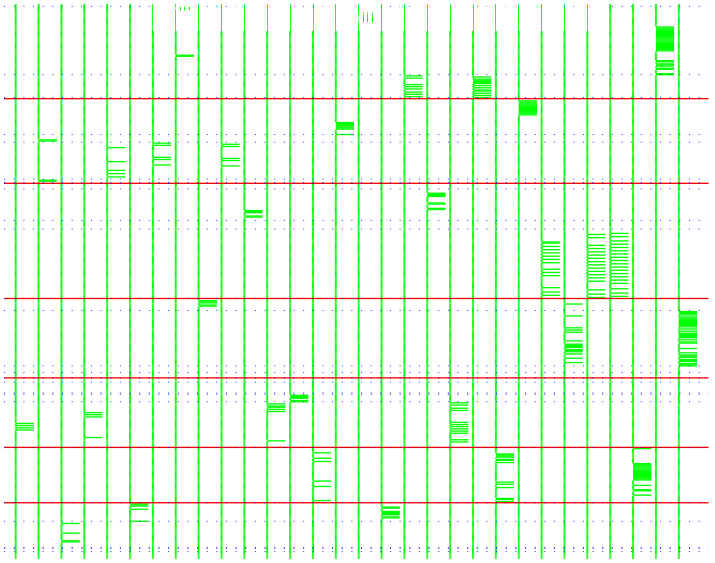
For the Cupredoxin and FAD/NAD(P)-binding domain super-families the spectral method showed a considerable improvement over the other methods assigning most of the sequences to one cluster.

Both TribeMCL and spectral clustering grouped most of the MHC sequences together. However, all methods failed to group the 'scorpion toxin-like' super-family sequences into one cluster. Similarly the 'viral coat and capsid proteins' super-family sequences were not clustered using any of the methods. (We note that later versions of the SCOP have changed the categorization of these proteins.)

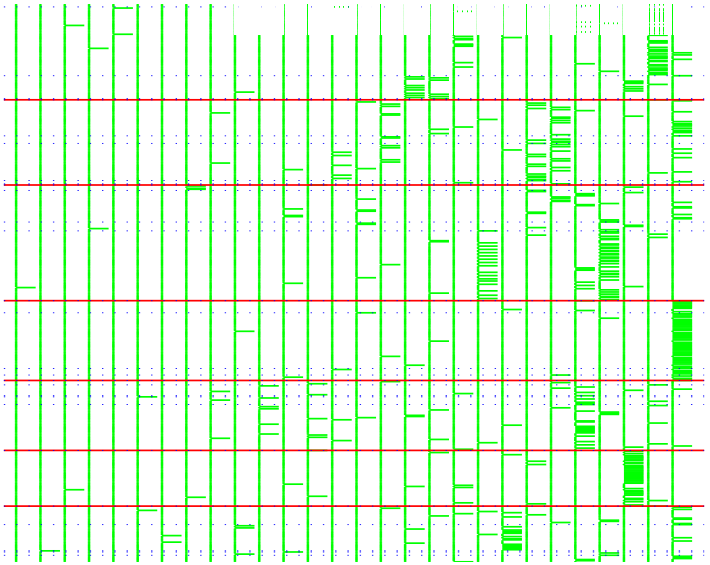
GeneRAGE



Hierarchical Clustering



TribeMCL



Spectral Clustering