Supporting Material
Experiment 2

Spectral Clustering of Protein Sequences

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

EXPERIMENT 2

Upon a suggestion of one of the referees we prepared a dataset that included protein sequences from the NAD(P)-binding Rossmann-fold domains super-family and also from the Triosephosphate isomerase (TIM) super-family.
This dataset consists of 430 sequences belonging to 5 super-families, namely NAD(P)-binding Rossmann-fold domains (208) Triosephosphate isomerase (TIM) (15) Nucleotide-binding domain (8) Globin-like (97) EF-hand (102). (Notice that the number of proteins in the Globin-like and EF-hand super-families have changed, because this is a slightly more recent version of SCOP.)

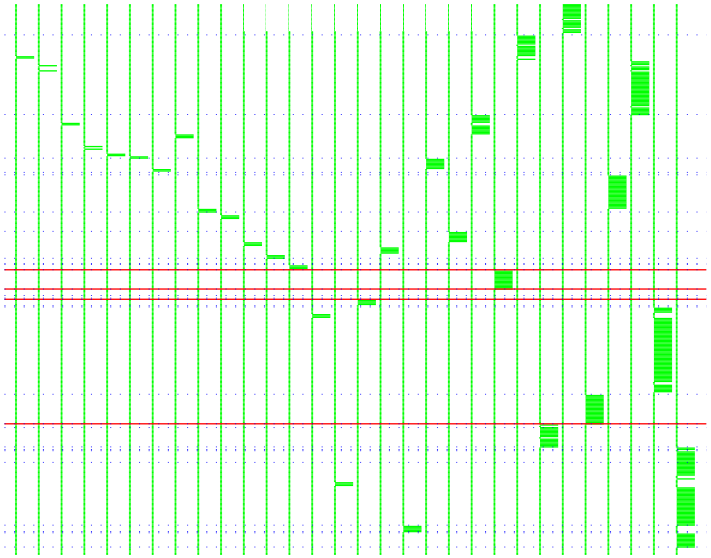
The results obtained applying to this dataset our implementation of GeneRAGE, the Hierarchical Clustering algorithm, TribeMCL and our spectral clustering algorithm are shown in the four diagrams of the figure.
The diagrams should be read as follows. Each row corresponds to a different cluster. Short (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 6 clusters; at the same time, our implementation of GeneRAGE detects 91 clusters, the hierarchical clustering detects 150 clusters, and TribeMCL 15 (with inflation parameter set to 1.60). The better quality of the clustering is quantified by the F-measure: for the spectral clustering it is equal to 0.8186; our implementation of GeneRAGE has a score of 0.5486, the hierarchical clustering 0.2854 and TribeMCL 0.6474.

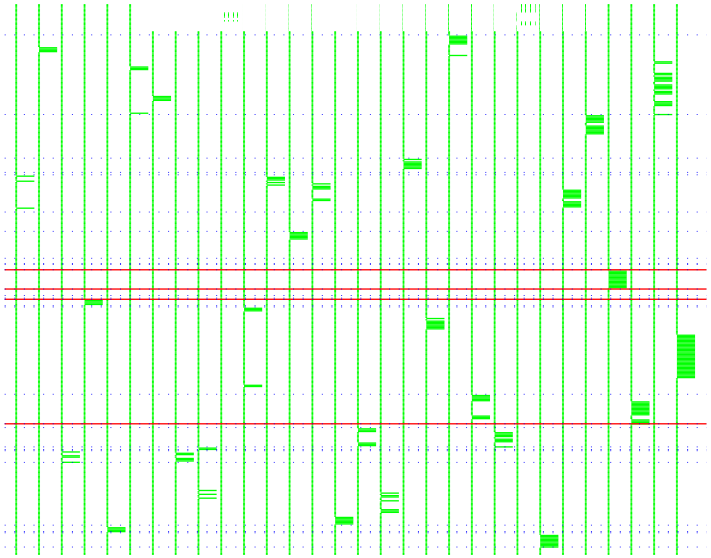
We observe that the GeneRAGE algorithm tends to cluster at the family level although it correctly assigns most of the EF-hand super-family to one cluster. The spectral method performs well including most of the NAD(P)-binding Rossmann fold domains in one cluster, although interestingly one family in this super-family (Tyrosine-dependent oxidoreductases) was assigned to a separate cluster. The spectral method also included the Nucleotide-binding domain super-

family in the cluster for the Rossmann fold domain super-family. TribeMCL is also able to group together most of the member of this family.
All four methods perform well at grouping the TIM super-family sequences although TribeMCL clusters it together with the EF-hand superfamily.

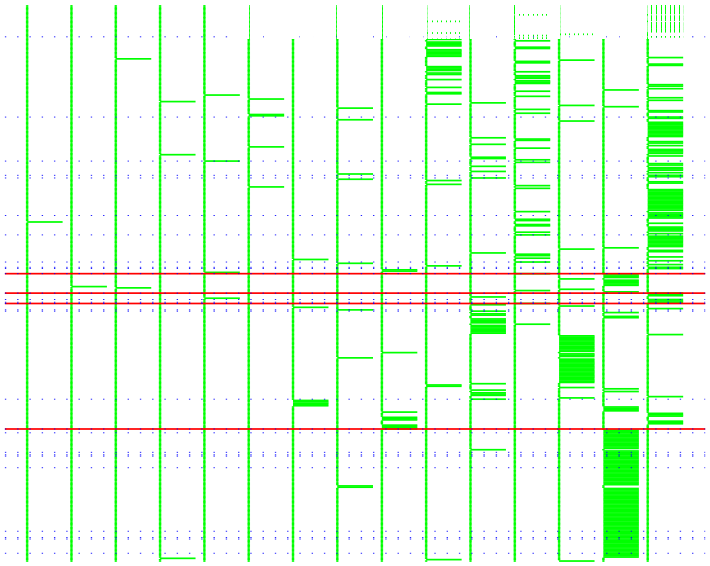
GeneRAGE



Hierarchical Clustering



TribeMCL



Spectral Clustering