Clustering Proteins by Structural Features

Bonnie Kirkpatrick, UC Berkeley

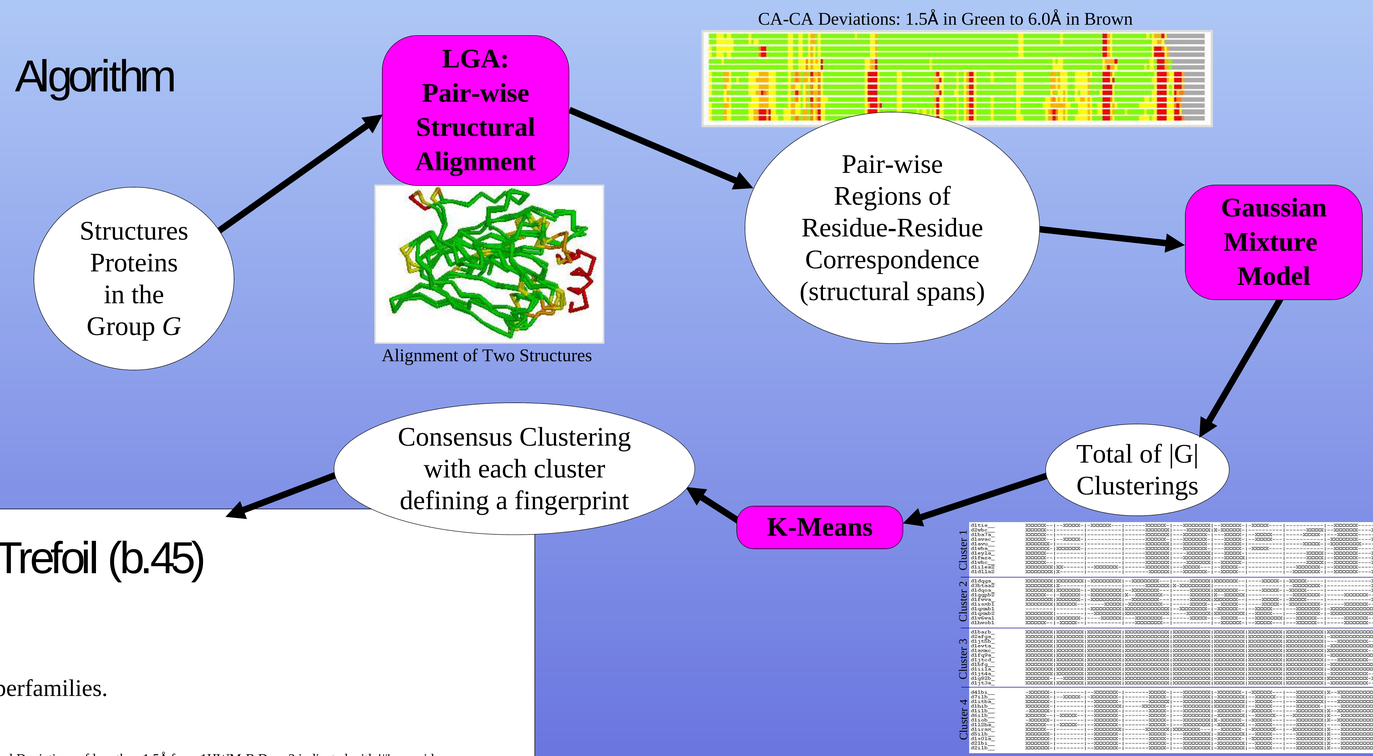
EEBI Division, Adam Zemla and Carol Zhou
Lawrence Livermore National Laboratory

Problem

Automatically insert new protein structures into the SCOP classification of proteins.

Our approach incorporates both clustering and prediction. We use clustering approaches to discover groups of related proteins along with the structural features that they share. Each cluster of structures has a maximal set of shared structural features (enumerated residue-wise) which we call a fingerprint. Once we have the structural fingerprint, we use it to determine the cluster membership of new structures.

Algorithm



A single clustering from our method: each cluster contains proteins with similar structural spans (indicated with 'X').

Clustering and Fingerprint Results for SCOP Fold Beta-Trefoil (b.45)

The Fold contains 6 Superfamilies split into 10 Families with a total of 99 unique structures

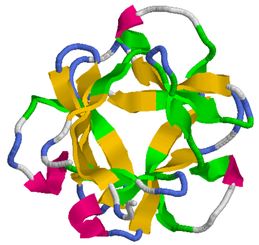
We clustered the structures into 6 clusters using the structural spans:

- One of the clusters correspond to a Superfamily in the Fold.
- Four clusters contained individual families which are the 4 Families belonging to 2 Superfamilies.
- One of the clusters contained a mixture of structures from several Superfamilies.

Accuracy	88.1%
False Positive Rate	6.6%
False Negative Rate	36.8%

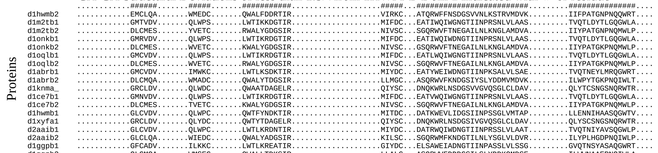
Error rates when taking the SCOP Superfamilies as the correct classification
(A random classification has 16% accuracy.)

Our algorithm found a Ricin-B-like structural fingerprint (upper far-right) that compares favorably to the fingerprint derived directly from the gold-standard SCOP classification (lower far-right).

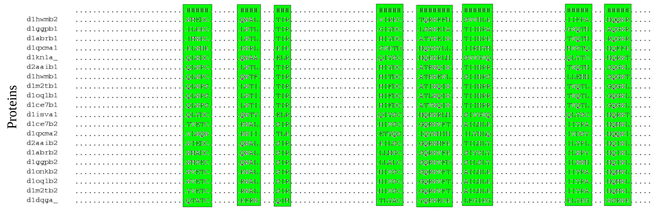


Ricin B Structural Spans (in green)
According to the SCOP Superfamily

CA-CA Structural Deviations of less than 1.5Å from 1HWM-B Dom 2 indicated with '#' or residue



Common Structural Regions of the Ricin-B Related Proteins
Found by the Gaussian Mixture Model



Common spans for SCOP's Ricin B Superfamily (b.45.2):
green corresponds to regions of small structural deviation also indicated in green on Ricin (at left).

Discussion

The data are very noisy. The PDB (Protein Data Bank) has about 23,000 unique structures ranging in quality from 0.54Å to 15Å. Despite this noise, the robustness of the consensus clustering allows our method to detect relationships on the level of superfamily.

Current work involves predicting the family and superfamily that a new structure belongs to. Our results indicate that the fingerprint derived from a SCOP family can predict membership with almost complete accuracy.

BK was funded for this work by the DOE CSGF fellowship under grant number DE-FG02-97ER25308.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.