Discovering protein functional sites with unsupervised techniques

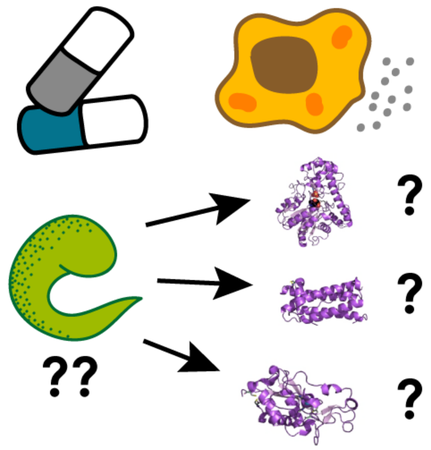
Shirley Wu[1], Russ Altman[2]

1 Program in Biomedical Informatics, 2 Department of Bioengineering

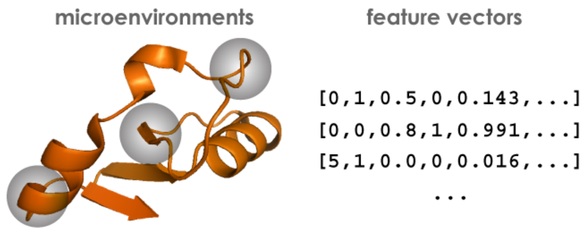
Background

Characterizing protein function - what reactions they carry out, what molecules they bind, etc - is important for understanding biological processes. We can use this knowledge to engineer therapeutics and other beneficial biology.

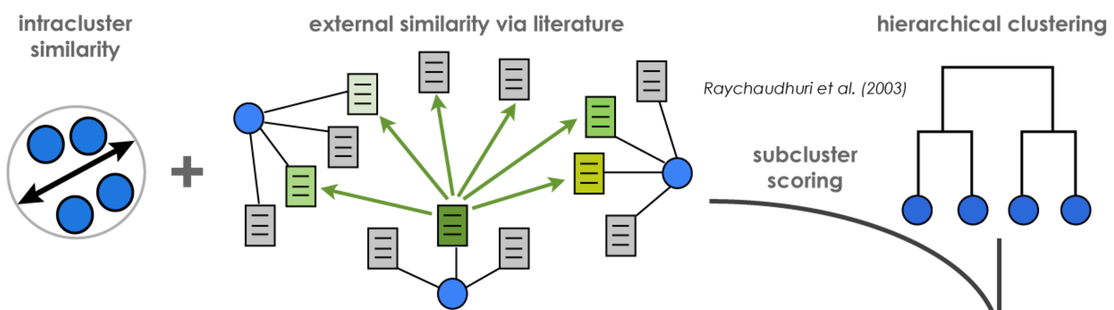
Computational methods are fast and inexpensive, allowing high-throughput prediction of protein function. Most methods are supervised approaches, i.e. they use available data about known proteins and functions to make predictions. Large-scale genomics efforts, however, are increasing the rate at which we discover novel organisms and proteins, so we also need methods to identify new biological functions.



Methods



To identify functional sites without prior knowledge, we use k-means clustering to group together protein structure microenvironments. Microenvironments are described by vectors of physical and chemical features found around the site of interest. We have done this for a representative subset of the Protein Data Bank (PDB), resulting in thousands of clusters.

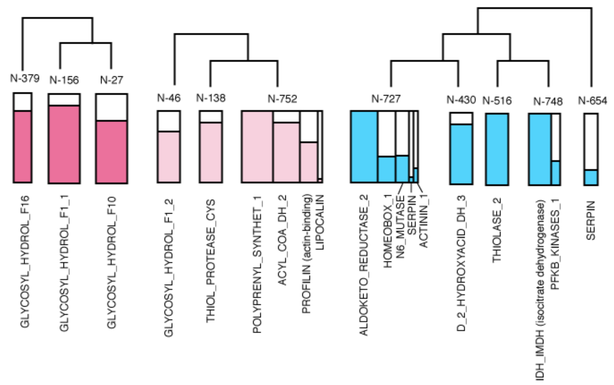


To prioritize the clusters produced by k-means, we use a scoring function that combines internal and external measures of coherence. For larger or less coherent clusters, we first employ hierarchical clustering to group the microenvironments into a tree based on their similarity to each other. We then use the scoring function to select the optimal subclusters within that tree.

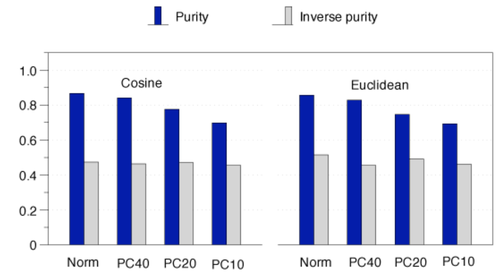
Evaluation

To evaluate different distance metrics and normalizations of vectors for subcluster selection, we used a set of 1400 microenvironments with assignments to 168 known functions. We found that cosine similarity produced subclusters with slightly better purity with regards to known assignments (external coherence), and better silhouette widths, which balance intracluster tightness and intercluster separation (internal coherence). In addition, reducing the number of features using principal component analysis results in subclusters that are more internally coherent but less externally coherent.

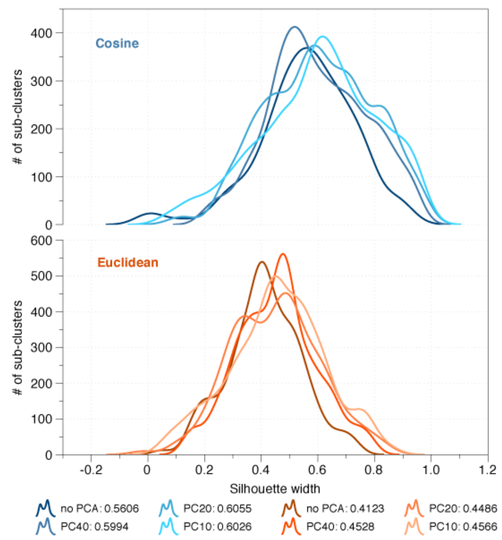
Examples of subclusters with known functional assignments



Purity and inverse purity of validation set



Distribution of silhouette values for validation set



Application

We are currently applying the subcluster selection strategy to the whole-PDB k-means clustering to determine compelling candidates for further analysis. We use a number of term enrichment methods to gain insight into the possible biological role of the microenvironment represented by each candidate subcluster.

Cluster 257: 30 proteins

MeSH terms	Raw text terms	Gene Ontology terms (Boyle et al. (2004))
Insulin	structur monomer	hormone activity
Hydrogen Bonding	sequenc c-peptid	glucose metabolic process
Ribosomal Proteins	monomer insulin	receptor binding
Peptides	conform insulin	hexose metabolic process
Pancreas	hexam crystal	insulin receptor binding
Amino Acids	2zn insulin	monosaccharide metabolic process
Chymotrypsin	protein-protein	negative regulation of catabolic process
Electrophoresis	coordin zn	positive regulation of cytokine secretion
Protein Folding	zn atom	insulin-like growth factor binding

References

Yoon S, Ebert JC, Chung EY, De Micheli G, Altman RB. (2007) BMC Bioinformatics, 8:Suppl 4:S10.
Raychaudhuri S, Chang J, Imam F, Altman RB. (2003) Nucleic Acids Res 31(15):4553-60.
Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. (2004) Bioinformatics 20(18):3710-5.