A workflow for Protein Homology Inference using Sequence and Structural comparison

Anuj Sharma, Ioannis Z. Emiris, Elias S. Manolakos
Department of Informatics and Telecommunications, University of Athens, Greece
Graduate Program "Information Technologies in Medicine and Biology"

Abstract

We developed a work flow for improving Homology detection for Proteins. The proposed method combines protein sequence and structure comparison data for detecting Homology. The work flow involves reclassifying 'twilight zone' proteins, in the PSI-BLAST results, into 'true positives' and 'true negatives'. The reclassification is done using a kNN Classifier built from the structural data. In our preliminary test we correctly identified 61% of the "True Positives" and 91% of the "True negatives" that would other-wise be lost in the "twilight zone". We are currently investigating gridfication of the work flow.

Motivation

Homology detection involves identifying proteins with common ancestry.

Why is it important?

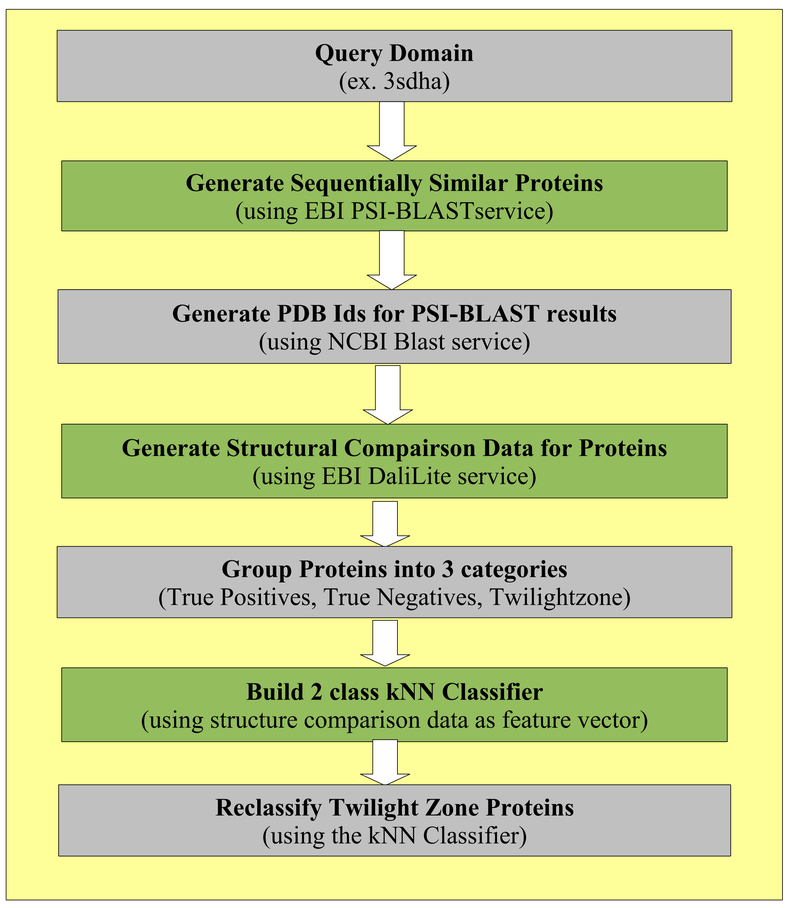
Homology relationship between two proteins allows inferring knowledge about the function of the unknown protein.

Why combine sequence and structure comparison?

- Methods based purely on sequence similarity have been shown to have inherent statistical limitations.
- Combined profile based techniques have been shown to outperform methods based on sequence or structure information individually [1].
- Sequence alignments are unambiguously accurate for protein pairs with high degree of identity. However large errors occur in the "twilight zone" (20-35%) of sequence identity [2].

Technical Methods

Proposed work flow



- Proteins with PSI-BLAST e-score < 0.0005 are considered True Positives and those with e-score > 10 are considered True Negatives.
- Only the top PDB domain, from sequence BLAST, of a protein is structurally compared to the query domain.
- Protein domain structure comparison is based on the following features [3]:
  - C score – indicates residue conservation in structural alignment.
  - Q value – indicates RMSD of structural alignment.
  - Nm – indicates percentage of aligned residues in structural alignment.
- The kNN classifier uses the structure comparison feature vectors for reclassification of Twilight Zone proteins.

Validation

- The SCOP Astral file will be used as the "golden standard".
- Domains having the same super family will be treated as Homologs.
  - Scop domain files contain structural classification for the domains.
  - Super family for each test domain and domains found through PSI-BLAST can be obtained from the SCOP Astral file.
- The PDB40D SCOP file will be used as the test set.

Preliminary Results

A preliminary test of the work flow was performed on 102 randomly selected domains from the PDB40D file. Scop – Astral file (release 1.73) was used as the "Gold-Standard" of protein homology.

Three one way ANOVA tests, one each for the three features, were performed for each of the domains at 5% significance level.
- The null hypothesis for the tests was that the compared classes have equal means.
- The three classes compared pair-wise were – True Positive, True Negatives and Twilight Zone.

What we observed?
- Taken independently Q-score had the highest discriminating ability of the three features.
- For most test domains clear distinction was observed only between Twilight Zone and True Negative proteins.
- Taken independently none of the feature can differentiate the classes accurately which was expected [3].

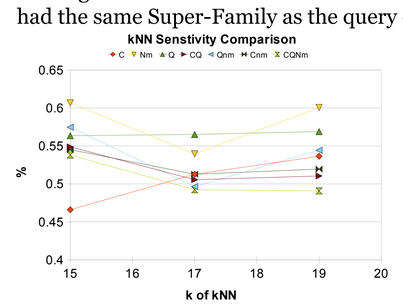
"Twilight Zone Protein" reclassification to "Positives" and "Negatives" was performed using kNNs. 7 kNN classifiers were generated for each test domain. In each kNN the feature vector used was different.
- Each structural parameter taken separately (3 classifiers).
- Each pair of parameters taken separately (3 classifiers).
- All three parameters considered together (1 classifier).

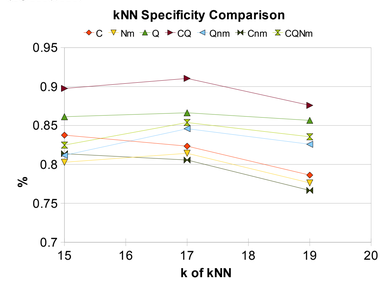
Performance of each kNN was measured in terms of the following metrics:

$$Sensitivity = \frac{No.\ of\ homologous\ protein\ domains\ correctly\ classified}{No.\ of\ homologous\ proteins\ contained\ by\ the\ twilight\ zone\ of\ the\ query\ domain}$$

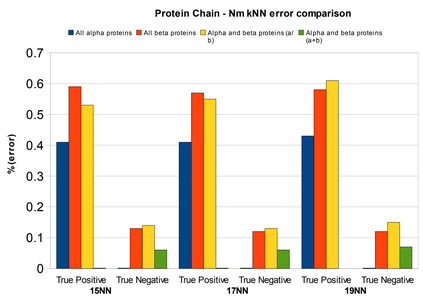
$$Specificity = \frac{No.\ of\ non\ homologous\ protein\ domains\ correctly\ classified}{No.\ of\ non\ homologous\ proteins\ contained\ by\ the\ twilight\ zone\ of\ the\ query\ domain}$$

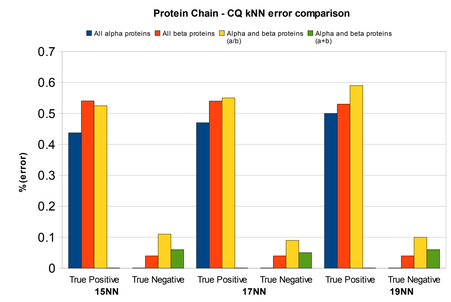
The graphs below compares the performance of the kNN classifiers with k varied from 15 to 19.
- A twilight zone domain was considered to be correctly classified as True Positive if it had the same Super-Family as the query domain.





Classification performance of the Nm and CQ - kNNs for the top 4 domain chain types in the test dataset was also compared. The following graphs show the comparison.





Discussion

- Classifiers based on the feature Nm performed best in terms of sensitivity. This was not expected since Q has been identified as the more "accurate" of the features [3].
- Classifiers based on the feature pair CQ performed best in terms of specificity. This was closer to expectations.
- It was interesting to note that the classifiers were able to re-classify domains with "Alpha and Beta proteins (a+b)" chain into "True Positives" correctly on most occasions.
- It was interesting to note that the classifiers were able to re-classify domains with "Alpha" chain into "True Negatives" correctly on most occasions.
- Reclassification of the "twilight zone" proteins by the classifiers was worst for protein domains with "Beta" type chain both in terms of identifying then as True Positives and True Negatives.
- In general the classifiers performed better in identifying the lack of homology between a pair of proteins.

Gridification

Why Gridify?
To establish the utility of the work flow comprehensive testing is required.
- The PDB40D dataset contains over 936 proteins.
- Large amounts of network data transfer is performed for each query.
- The gridified work flow is expected to perform much faster.

Road to Gridification!
- Build the work flow for a grid using TAVERNA.
- Setup the work flow on Hellas Grid.
- Use services available in the BioMed VO to run the work flow.

References

1. "Structural Genomics: Computational methods for structure analysis", Sharon Goldsmith-Fischman, Barry Honig, Protein Science, 2003.
2. "Twilight zone of protein sequence alignments", Burkhard Rost, Protein Engineering, 1999.
3. "On the relationship between sequence and structure similarities in proteomics", Evgeny Krissinel, Bioinformatics, 2007.