Assessment of 3D Protein Domain Predictions

Luca Cazzanti and Robyn Greaby

CSE 527

Autumn 2004

Goal

- Use Gaussian mixture models to estimate the likelihood that a 3D protein domain structure prediction produced by Rosetta and its PDB match are functionally similar

Background

- Number of sequences whose function is unknown in increasing
- Accurate estimation of protein function is key to understanding a designing cellular processes
- Many newly determined sequences do not have sufficient sequence homology to known sequences to used methods like Pfam to estimate function
- Structure is better conserved than sequence
- Infer function based on structural similarity

Rosetta

- Most successful de novo protein structure prediction method currently available
- Generates several thousand candidate structures for each sequence
- Uses a Monte Carlo search to find the conformations that can be built from smaller local structures derived from sequence segments
- Uses two optimization paths to find compatible combinations of global and local structures
- Predictions have low global and local free energies
- A strategy is needed to infer the function of the structural predictions produced by Rosetta

Methods

- Find the best 3D structural match for each prediction to a domain in the PDB
  - Sequence-independent structural alignment procedure
- Use Gaussian mixture models (GMMs) to estimate the likelihood that a prediction and its PDB match are functionally similar
  - Considered similar if they are in the same SCOP superfamily
- Test using sequences from the PDB with known function

Methods cont…

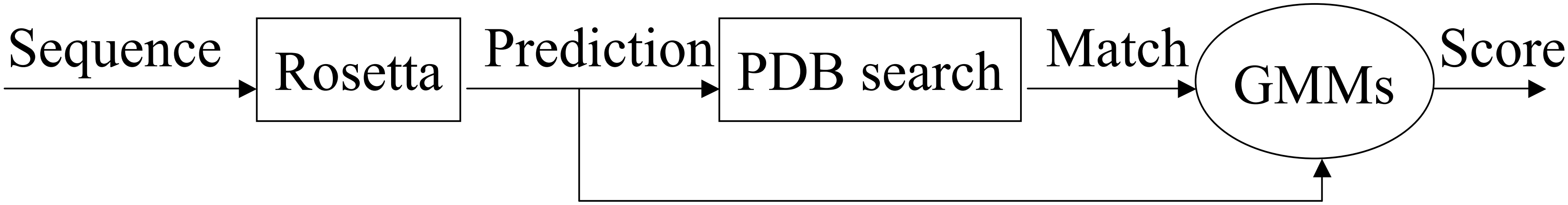


Figure 1: Diagram of proposed approach to assessing Rosetta 3D protein domain structure predictions. GMMs estimate the likelihood that the prediction and its closest PDB match are functionally similar.

Measures of Structural Similarity

- Mammoth z-score
- α-helices and β-sheets
- Length

Mammoth z-score

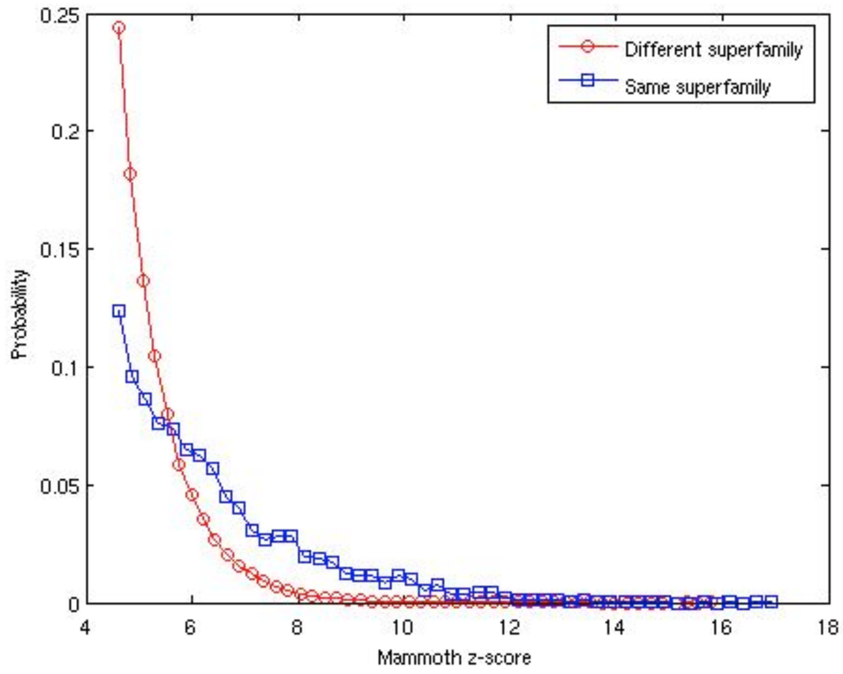


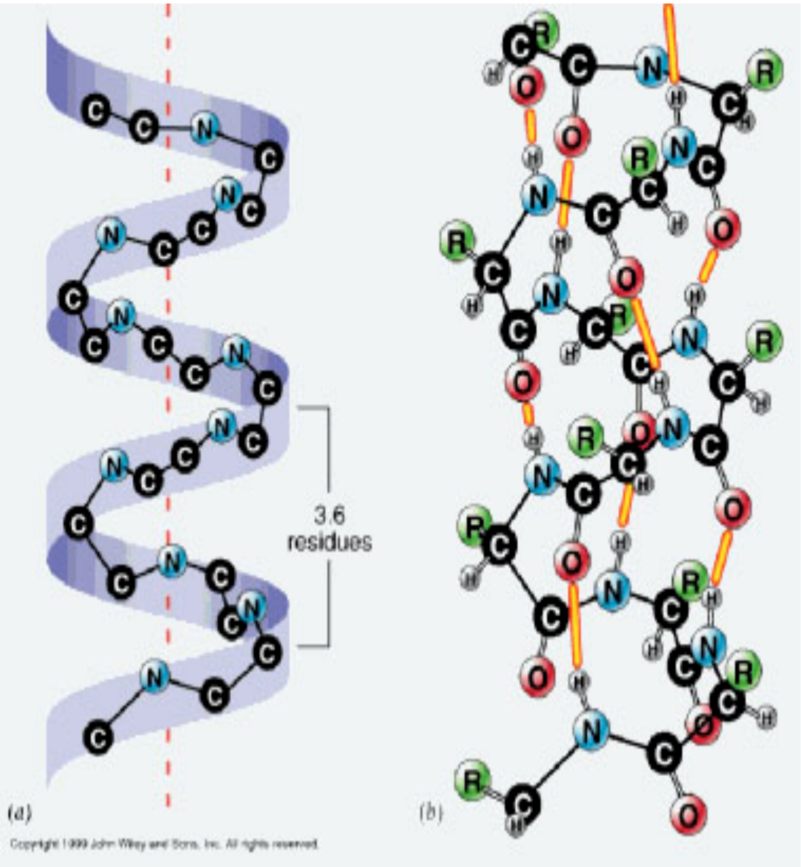
Figure 2: Distributions of Mammoth z-scores for the Rosetta protein structure predictions

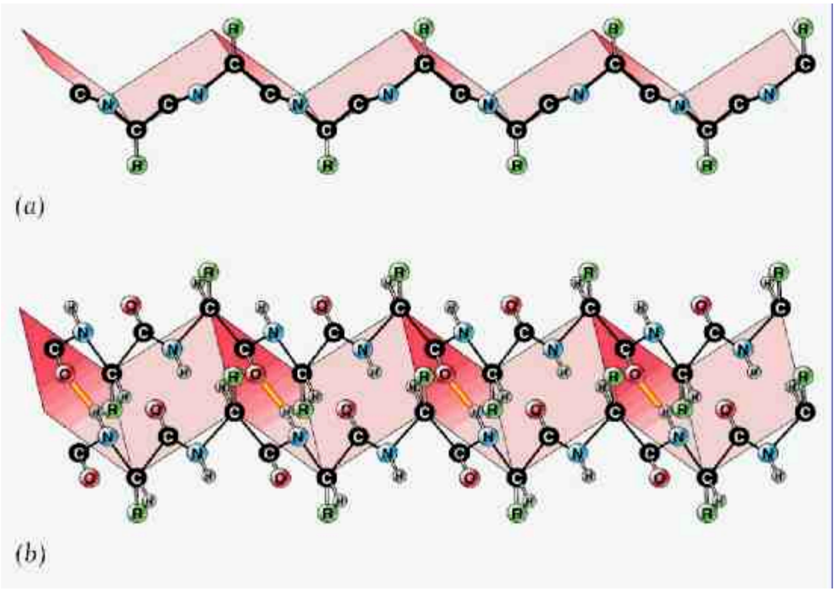
- Sequence-independent structure-to-structure comparison
- Based on RMSD of structure alignments
- Takes number of residues into account
- Rosetta predictions in the same SCOP superfamily as their PDB match have higher z-score.
- Large overlap between curves

α-helices and β-sheets

- Tertiary protein structures made of smaller secondary structures that are linked to protein function
- α-helices
  - Right handed helix with 3.6 residues per turn
  - Hydrogen bonds between peptide C=O in an amino acid and the peptide N-H bond four residues away
- β-sheets
  - Hydrogen bonds between neighboring peptide strands
  - Oriented either parallel or antiparallel

α-helices and β-sheets cont…


3.6
residues
(a)
(b)
Copyright 1999 John Wiley and Sons, Inc. All rights reserved.



α-helix

β-sheet

Images from: http://www.uic.edu/classes/bios/bios100/lectf03am/lect02.htm

α-helices and β-sheets cont…

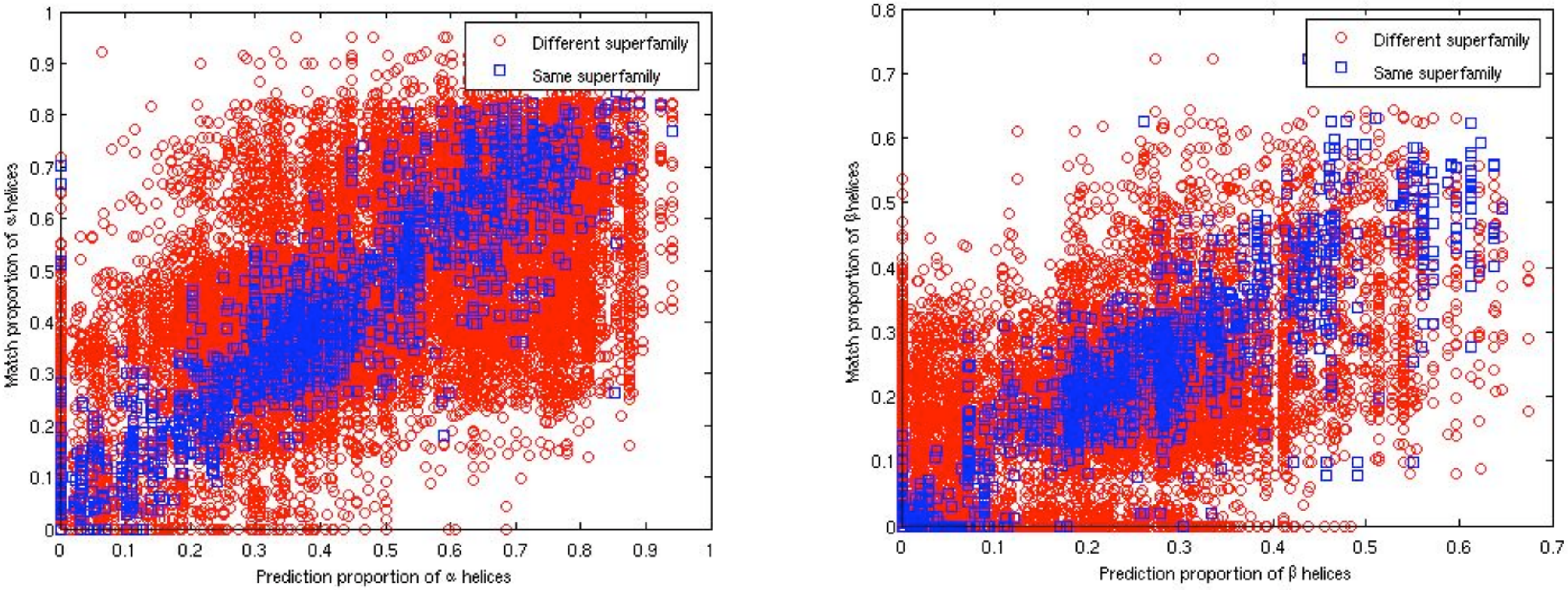


Figure 3: Plots of the percentages of α-helices and β-sheets for C_s and C_d. Prediction that have a PDB match in the same SCOP super family have more tightly clustered percentages

Length

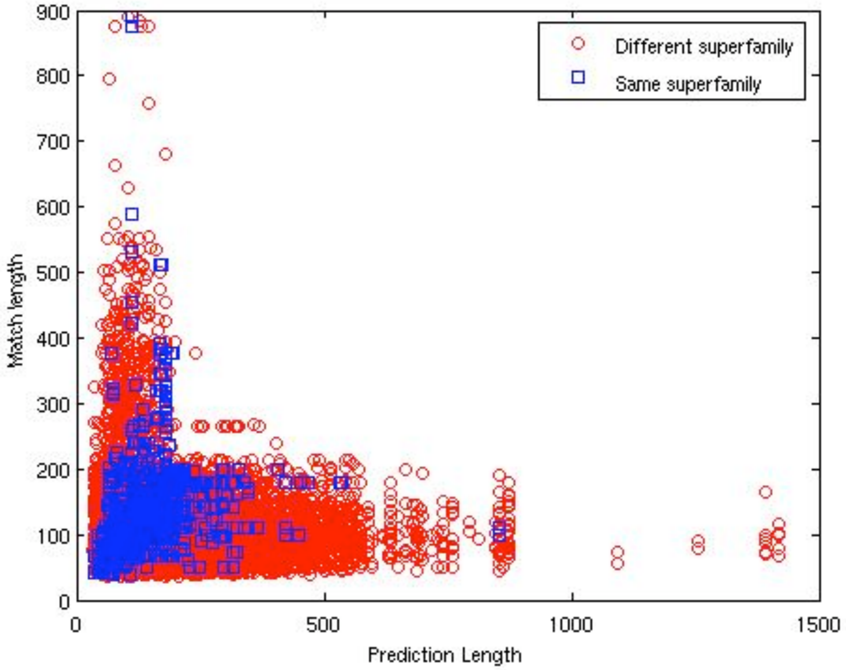


Figure 4: Distribution of the prediction length and the PDB match length

- Predictions that are close in length to their PDB match are more likely to be in the same superfamily than those far apart

Features for Classification

- 4 features based on z-score, secondary structure, and length

$$x_1 = \frac{prediction\ length}{match\ length} - 1$$

$$x_2 = \%\alpha_p - \%\alpha_m$$

$$x_3 = \%\beta_p - \%\beta_m$$

$$x_4 = Mammoth\ z - score$$

Gaussian Mixtures

- Two classes
  - C_s: belongs to same SCOP superfamily as PDB match
  - C_d: does not belong to same superfamily of PDB match

$$p(x \mid C) = \sum_{k=1}^{K} w_k N(\mu_k, \sum k \mid C), \sum_{k=1}^{K} w_k = 1$$

$$C = \{C_s, C_d\}$$

$$\log\left(\frac{p(x \mid C_s)}{p(x \mid C_d)}\right) > T$$

- Use diagonal Σ_k to reduce number of parameters
- Trained and tested with EM algorithm with 5-way cross validation

Data

- 192,249 Rosetta prediction/PDB match pairs for 8,560 domains
  - 4,745 pairs in C_s
  - 187,495 pairs in C_d
- Matches determined by comparing Rosetta prediction to ASTRAL compendium
  - Sub list of PDB domains with low functional redundancy and low sequence homology
  - Test/match sequences have less than 40% homology
- Only match pairs with Mammoth z-score greater than 4.5 considered

Log-likelihood ratio

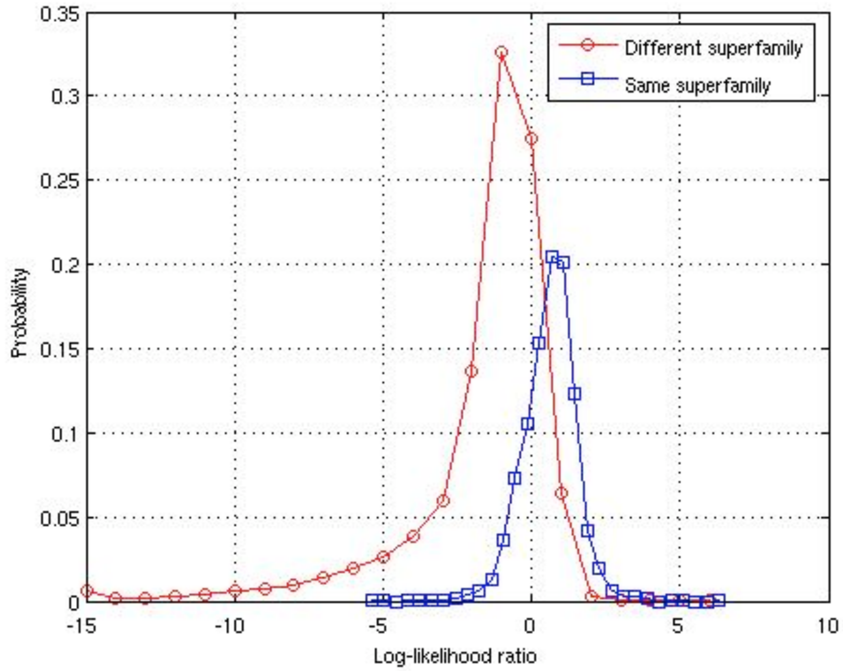


Figure 6: Log-likelihood ration distributions for C_s and C_d. Scores lower than –15 are truncated for display purposes.

- Two distributions are well separated
- High scores correspond to Rosetta predictions deemed functionally similar to their PDB matches

Error performance

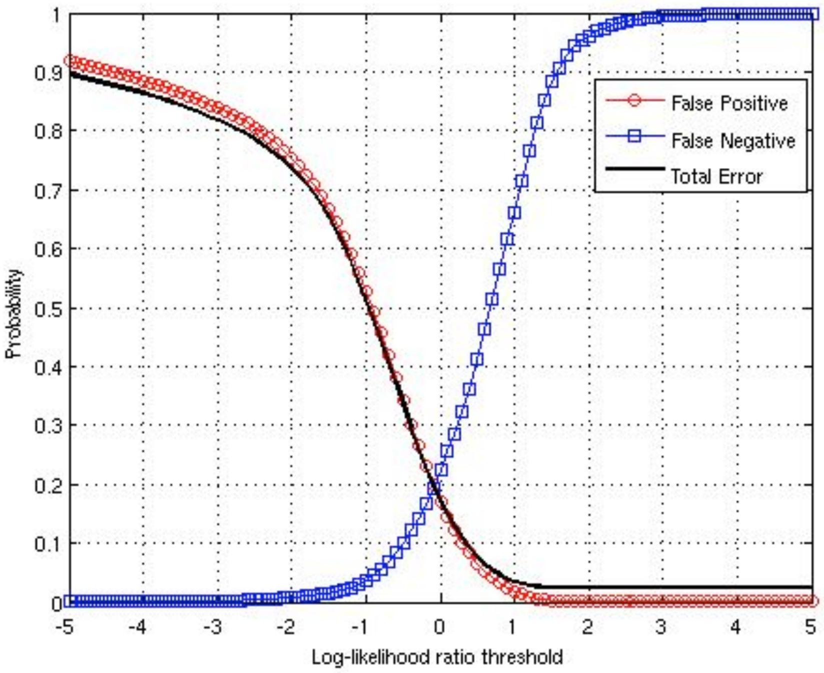


Figure 7: Error performance for different values of the threshold T

- Binary classification of the predictions achieved by comparing scores to a threshold
- For T = 0
  - False positive: 17.31%
  - False negative: 22.36%
  - Total error: 17.28%

Summary and Conclusions

- Described an approach for assessing the quality of 3D protein domain structure predictions produced by Rosetta
- Uses Gaussian mixture models to estimate the likelihood that a prediction and its PDB match are functionally similar
- Can be used to make functional predictions for newly discovered sequences

Acknowledgements

- The Baker Lab, Dept. of Biochemistry, UW
  - Prof. David Baker and Lars Malström
- Prof. Maya Gupta, Dept EE, UW