Assessment of 3D Protein Domain Predictions

Project Report for CSE 527 Computational Biology

Autumn 2004

University of Washington, Seattle.

Luca Cazzanti and Robyn Greaby

Abstract

We apply Gaussian mixture models to the problem of classifying three-dimensional protein domain structure predictions generated with the Rosetta structure prediction method. A standard log-likelihood ratio test distinguishes good predictions from lower quality ones with 83% accuracy. Good separation between higher and lower scores also suggests that the best predictions can be identified with a high degree of confidence. The results are significant for inferring protein function from de novo protein domain structure predictions.

1 Background

The number of gene sequences in databases is growing rapidly. Proportionately, the number of sequences whose function is unknown is also increasing. Accurate estimation of protein function is key to understanding and designing cellular processes. In many cases the newly determined sequences do not exhibit sufficient homology to known sequences, thus methods like Pfam produce poor protein function estimates. Structures, on the other hand, are conserved across greater evolutionary distances than sequences. Thus, methods that infer the function of new sequences from their three-dimensional structure similarity to known domains are a powerful alternative to sequence-based methods.

Rosetta is the most successful de novo protein structure prediction methods available today [1, 2]. It generates several thousand candidate structures from each sequence by applying a Monte Carlo search to the set of conformations that can be built from smaller, local structures derived from sequence segments [3, 4]. Two optimization paths find compatible combinations of local and global structures, and the resulting candidate predictions are stable in the sense that they have low free energy both locally and globally.

A strategy is needed to infer the function of the structural predictions produced by Rosetta. Following the approach in [1], for each three-dimensional structure, the best match in the Protein Data Bank (PDB) is found using MaxSub [5], a sequence-independent structural alignment procedure. Our work uses Gaussian mixture models (GMMs) to estimate the likelihood that a prediction and its PDB match are functionally similar. Based on our estimated likelihood, we classify the structural predictions as an accurate functional match or not, resulting in a functional prediction for the sequence.

Protein domains are commonly assumed to be functionally similar if they are in the same SCOP superfamily [6, 7, 8]. Since there is no structural and functional ground truth for newly discovered sequences, we adopt the strategy in prior work [1] and use test sequences from the PDB with known function that can be used as ground truth to determine prediction success. Figure 1 shows a diagram of our approach. We hypothesize that an approach that performs well on known sequences will generalize to the prediction of sequences of unknown function.

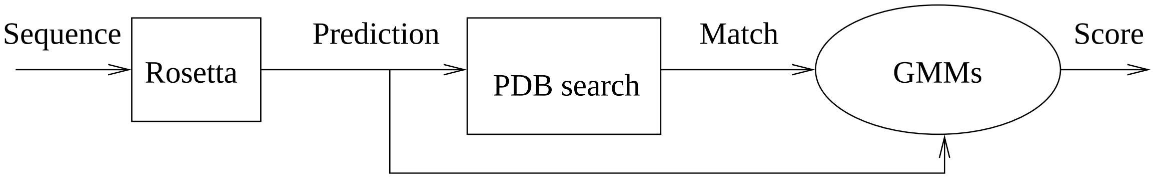


Figure 1: Diagram of proposed approach to assessing Rosetta three-dimensional protein domain structure predictions. GMMs estimate the likelihood that the prediction and its closest PDB match are functionally similar.

2 Classification with Gaussian Mixtures

We consider two classes, C_s and C_d, which respectively include the predictions that belong to the same SCOP superfamily as their PDB match, and the predictions that do not. The probability distribution of each class is modeled by a linear combination of multivariate Gaussians,

$$p(\mathbf{x}|C) = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \mathbf{\Sigma}_k|C)\,,\ \sum_{k=1}^{K} w_k = 1 \tag{1}$$

where $\mathbf{x}$ is the vector of features used for classification, $C = \{C_s, C_d\}$, $\mathcal{N}(\mu_k, \mathbf{\Sigma}_k|C)$ is the k-th Gaussian component of the mixture with mean μ_k and covariance matrix $\mathbf{\Sigma}_k$, and w_k is its associated weight. Each feature vector, corresponding to a pair Rosetta prediction/PDB match, is presented to each model, producing two likelihood scores, $p(\mathbf{x}|C_s)$ and $p(\mathbf{x}|C_d)$. The decision rule for classifying the prediction, formulated in terms of the log-likelihood ratio is,

$$\log\left(\frac{p(\mathbf{x}|C_s)}{p(\mathbf{x}|C_d)}\right) > T, \tag{2}$$

where T is a threshold value.

We use diagonal $\mathbf{\Sigma}_k$ in our implementation to reduce the number of parameters to train and thus speed up model training, which is carried out with the EM algorithm[1]. The Bayesian Information Criterion (BIC) [10] was used to determine that C_s is well modeled by $K = 10$ and C_d is well modeled by $K = 70$. The GMMs are trained and tested with 5-way cross validation

3 Measures of Structural Similarity

Based on theoretical and data analysis, we use the following features:
Mammoth z-score. Mammoth is a sequence-independent structure-to-structure comparison approach which is widely used in protein structure studies [11]. The Mammoth z-score is based on the root-mean-squared deviation (RMSD) of structural alignments and takes into account the number of residues in a structure. Figure 2a shows the distributions of Mammoth z-scores for the Rosetta protein structure predictions in our data set. The z-score distribution for predictions in C_s is more heavily weighted toward the higher values than that for predictions in C_d. However, a large overlap between the two curves remains, indicating that a good Mammoth z-score is not always a good indicator of functional similarity.
α-helices and β-sheets. Tertiary protein structures are made up of smaller, secondary structures which reflect the chemical interactions between the residues. These secondary structures are linked to protein function. Among them are α-helices and β-sheets. An α-helix is a right handed helix composed of 3.6 residues per turn. It is formed by a series

[1] We used a subset of the LNKNet software package from MIT Lincoln Laboratory [9], adapted for MATLAB.

of hydrogen bonds between the peptide C=O bond of an amino acid with the peptid N-H bond on the amino acid four residues away. This series of bonds forms a tightly packed 3D cylindrical structure, or helix, which encloses the hydrophobic residues and exposes the hydrophilic ones. A β-helix is formed by hydrogen bonds between neighboring peptide strands. The strands can be oriented in either a parallel or antiparallel structure. The bonded sections ripple into a characteristic pleated sheet [12].

The percentage of α-helices and β-sheets in the prediction and its match is a measure of structural similarity. Figures 2b and 2c show plots of the percentages of α-helices and β-sheets for C_s and C_d Note that for predictions that have a PDB match in the same SCOP superfamily the percentages are more tightly clustered.

Sequence length. Predictions that are close in length to their PDB match are more likely to be in the same superfamily than those that are far apart. Figure 2d shows the distribution of the prediction length and the PDB match length. Note how the length of the predictions in C_s are much more tightly correlated than those in C_d.

Each Rosetta prediction and its PDB match are represented by a feature vector $\mathbf{x}$ with 4 elements. Let r be the ratio of the prediction length to its PDB match length. Then $\mathbf{x}_1 = r - 1$. Let (α_p, β_p) and (α_m, β_m) be the percentages of α-helices and β-sheets in the Rosetta prediction and its PDB match respectively. Then, $\mathbf{x}_2 = \alpha_p - \alpha_m$ and $\mathbf{x}_3 = \beta_p - \beta_m$. Lastly, $\mathbf{x}_4$ is the Mammoth z-score of each prediction to its PDB match.

4 Results and Discussion

Our data set consists of 192,240 total Rosetta prediction/PDB match pairs for 8,560 domains; 4,745 pairs are in C_s and the remaining 187,495 are in C_d. The PDB matches are determined by comparing the Rosetta predictions to the ASTRAL compendium, which organizes the structures listed in the PDB in domains of low functional redundancy [13, 14]. The search is limited to a subset of PDB domain structures that have less than 40% sequence homology. Higher homology within the set of possible PDB matches indicates very close structural similarity, and would needlessly match Rosetta predictions to structurally redundant domains. Furthermore, care is taken to ensure that the subset of allowed PDB matches does not include the same sequences used as test structures in Rosetta. This would create self-matches, and the likelihood score would be biased. Finally, only prediction/match pairs with a Mammoth z-score greater than 4.5 are considered, as lower scores indicate a very poor structural match.

Figure 3 shows the distribution of the log-likelihood scores for C_s and C_d. The two distributions are well separated, with the higher scores corresponding to Rosetta predictions deemed functionally very similar to their PDB matches, and lower scores indicating poor functional matches.

Binary classification of the predictions is achieved by comparing the scores to a threshold value, as in equation 2. Figure 4 shows the percent of false positive, false negative, and total error as a function of the threshold T. For T=0, the classifier achieves a false positive rate of 17.31% and a false negative rate of 22.36%. The total error is 17.28%. Due to the high C_d/C_s ratio, the total error tracks the false positive error.

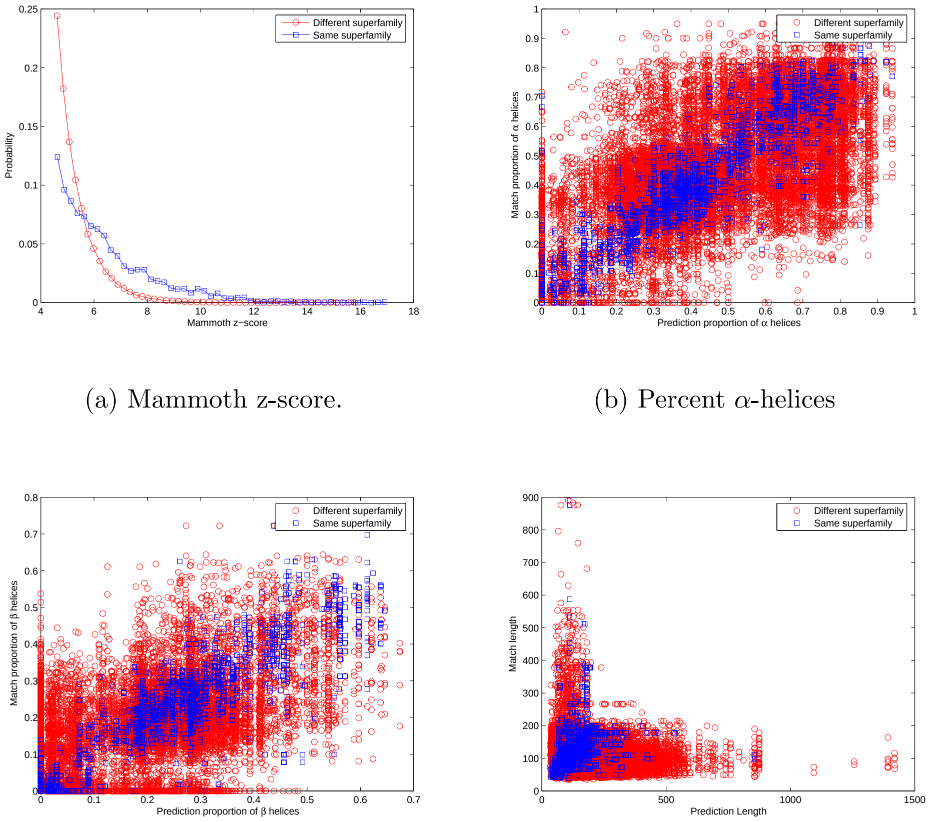
These results are significant because they allow us to make likely functional predictions in spite of the high number of structural predictions that cannot be tied to function. In particular, even for sequences with weak homology, protein function can be predicted accurately.

5 Summary

We have described an approach for assessing the quality of three-dimensional protein domain structure predictions produced by Rosetta. Our approach, based on Gaussian mixture models, estimates the likelihood that a prediction and its PDB match are functionally similar. Performance curves show that our approach allows successful functional predictions even when a high number of structural predictions cannot be tied to function. In the particularly interesting case of newly discovered sequences with no link to known domains, de novo methods for structural prediction combined with statistical inference for functional prediction provide a powerful approach to understanding and designing cellular processes.

Acknowledgments

We thank The Baker Lab, Dept. of Biochemistry, University of Washington, for providing the data. In particular we thank Prof. David Baker and Lars Malström for the helpful discussions. We also thank Prof. Maya Gupta, Dept. of Electrical Engineering, University of Washington, for her feedback on early versions of this report.



(a) Mammoth z-score.

(b) Percent α-helices

(c) Percent β-sheets.

(d) Length.

Figure 2: Four features used to measure structural similarity and predict the function of Rosetta predictions. The blue squares are prediction/match pairs in the same SCOP superfamily (C_s); the red circles are pairs in different superfamilies (C_d). To limit file size, only 1 in 10 predictions in C_d are plotted. The overall character of the scatter plots is not affected.

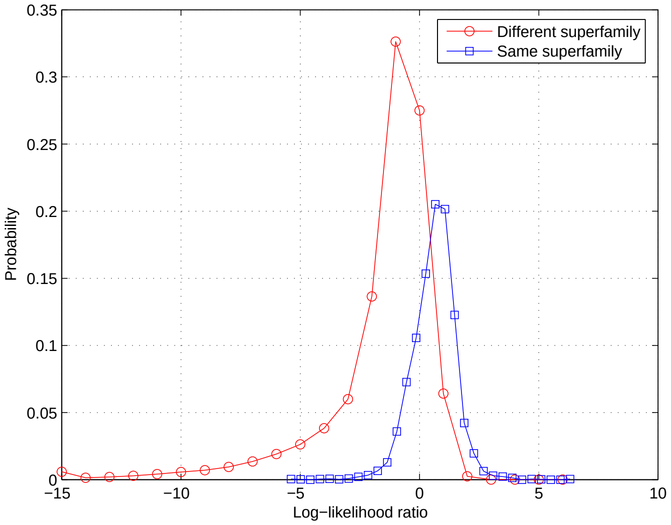


Figure 3: Log-likelihood ratio distributions for C_s and C_d. Scores lower than -15 are truncated for display purposes.

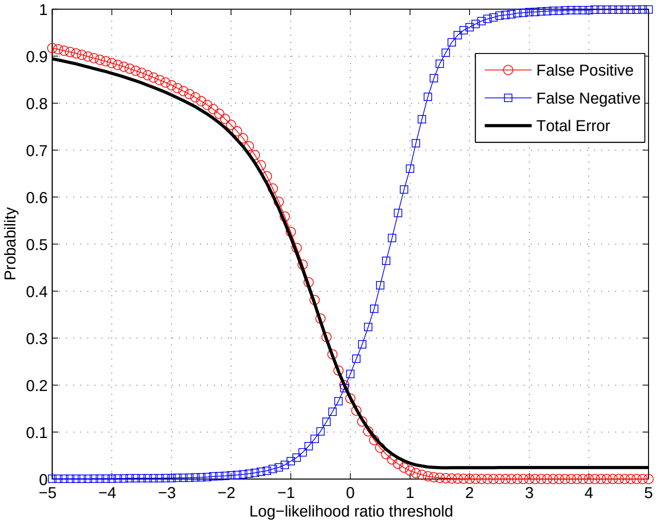


Figure 4: Error performance for different values of the threshold T.

References

[1] R. Bonneau, C. E. M. Strauss, C. A. Rohl, D. Chivia, P. Bradley, L. Malström, T. Robertson, and D. Baker, "De novo prediction of three-dimensional structures for major protein families," *Journal of Molecular Biology*, no. 322, pp. 65–78, 2002.

[2] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. A. Rohl, C. E. M. Strauss, and D. Baker, "Rosetta in CASP4: Progress in ab initio protein structure prediction," *Proteins: Structure, Functions, and Genetics Supplement*, no. 5, pp. 119–126, 2001.

[3] K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, "Ab initio protein structure prediction of CASP III targets using ROSETTA," *Proteins: Structure, Function, And Genetics*, no. Supplement 3, pp. 171–176, 1999.

[4] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein structure prediction using rosetta," *Methods in Enzymology*, no. 383, pp. 66–93, 2004.

[5] N. Siew, A. Elofsson, L. Ryclewski, and D. Fischer, "MaxSub: an automated measure for the assessment of protein prediction quality," *Bioinformatics*, no. 16, pp. 776–789, 2000.

[6] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, no. 247, pp. 563–540, 1995.

[7] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2002: Refinements accommodate structural genomics," *Nucleic Acids Research*, vol. 30, no. 1, pp. 264–267, 2002.

[8] A. M. Lesk, *Introduction to Protein Structure.* Oxford University Press, 2001.

[9] "LNKnet Pattern Classification Software," http://www.ll.mit.edu/IST/lnknet/index.html.

[10] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[11] A. R. Ortiz, C. E. M. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): An automated method for model comparison," *Protein Science*, no. 11, pp. 2606–2621, 2002.

[12] D. Voet, J. G. Voet, and C. W. Pratt, *Fundamentals of Biochemistry.* John Wiley & Sons, Inc., 2002.

[13] S. E. Brenner, P. Koehl, and M. Levitt, “The ASTRAL compendium for protein structure and sequence analysis,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 254–256, 2000.

[14] J.-M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner, “The ASTRAL compendium in 2004,” *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D189–D192, 2004.