

Predicting co-evolving pairs in Pfam using information theory where entropy is determined by phylogenetic mutation events

Scooter Willis

Department of Computer and Information Science and Engineering,
University of Florida, Gainesville, FL 32611, USA
willishf@ufl.edu

Abstract. The accurate prediction of co-evolving pairs in protein sequences plays an important role in tertiary protein structure prediction and protein engineering. Using information theory to detect co-evolving pairs is impacted by the phylogenetic effect on entropy measurements. Mutual Information (MI) is used to detect co-evolving pairs in a protein family by re-sampling based on mutation events in the phylogenetic tree (RPE). The predictive quality of co-evolving pairs with high mutual information ($Z \geq 4$), a sequence distance > 10 and within 12 angstroms using the RPE method is on average 81% in a Pfam family. The accuracy of detecting co-evolving pairs without RPE is 56%. This study represents the first known analysis of mutual information to detect co-evolving pairs in the full Pfam data set. Results for each protein family in Pfam and corresponding ribbon model graphing the MI relationships can be found at <http://www.proteinx3d.com>.

1 Introduction

Mutual information is the measure of mutual dependence between two variables. In protein structures this is referred to as co-evolving pairs [1]. When two amino acids are distant (>10 positions) in a sequence the fold of the protein could place them at contact points or near neighbors (<12 angstroms - Å) in 3D space. If one amino acid mutates then it is possible that a neighboring amino acid in 3D space will also mutate to preserve function or structure of the protein. The use of mutual information or entropy measures to detect co-evolving pairs is an actively researched topic [2],[3],[4],[5],[1],[6]. The ability to predict secondary structure of helix, strands or loops by homology is considered 70-80 percent accurate [7]. The ability to take these predicted secondary structures and develop an accurate tertiary model is considered a challenging an unsolved problem.

Developing an algorithm that allows the use of mutual information to detect co-evolving pairs that are close in 3D space but distant in a protein sequence would play an important role in tertiary structure prediction for that protein family. With a reliable method to predict co-evolving pairs this allows for the initial relative placement of secondary structures in a tertiary structure. This narrows the possible number of protein folding solutions and provides a potentially accurate base model which can then be further refined by already accepted methods.

With a growing database of sequences associated with a particular protein family the challenges of small sample size are becoming less of an issue. The small sampling problem is replaced by the impact of phylogenetic noise from closely related sequences. This represent the first known review of an algorithm to detect co-evolving pairs against the complete Pfam database which represents 85% coverage of the Swiss-Prot database and 75% coverage of the SP+TrEMBL database[8].

2 Information Theory Approach

Application of Information Theory and the analysis of sequence data are impacted by a sampling bias from targeted research on proteins of medical interest and the introduction of noise from the phylogenetic impact on probability calculations. The field of Information Theory was introduced by Shannon (1948), "A Mathematical Theory of Communication", which outlined the statistical measure of information and the detection of noise in a communication channel. Mutual information is based on Shannon Entropy (H)

which is derived from the probabilities of occurrences of individual and combined events between two discrete random variables. The entropy of a discrete random variable is given by formula (1) and for two discrete random variables in formula (2). Mutual Information is given by formula (3) and represents the sum of entropy between two variables minus the joint entropy between two variables. Entropy is maximized when the variable is completely random. Standard units for Entropy and Mutual Information are bits where the log base is 2.

$$H(x) = -\sum p(x) \log(p(x)) \quad (1)$$

$$H(x, y) = -\sum p(x, y) \log(p(x, y)) \quad (2)$$

$$MI(x, y) = H(x) + H(y) - H(x, y) \quad (3)$$

When calculating probabilities one underlying assumption is that the samples used to calculate the probabilities are randomly picked samples from the population. If the samples are not randomly picked then a bias is introduced towards the grouping of the samples [5]. In a protein family each sequence represents a sample of the population of sequences as a member of that family. The first introduced bias is that the sequences tend to come from research studies that have medical or pharmaceutical interests. Given common evolutionary history of all genetic sequences, the ability to survey the entire population for all members of a particular protein family is not practical. These two factors can contribute statistical bias or noise when calculating entropy or mutual information relationships.

Using a phylogenetic tree for a protein family it is possible to detect mutation events at a particular sequence position. This reduced set of mutations could then be used as the basis for probability calculations to calculate the entropy. In Figure 1, a binary tree represents a phylogenetic tree where the circles are the hypothetical parent and the boxes represent a sequence position in a protein family. Without taking into consideration the phylogenetic influence the probability of the set $p(A)=1/6$, $p(D)=1/6$ and $p(C)=2/3$. If we calculate the probability of observed mutation events starting from the parent node then $p(A)=1/3$, $p(D)=1/3$ and the $p(C)=1/3$. This is done by starting at the root node and counting all children nodes where the child does not equal the parent. One approach accurately represents the population sample and the latter represents the probability of transition to a different amino acid. This has the impact of reducing or compensating for the phylogenetic influence on probability calculations.

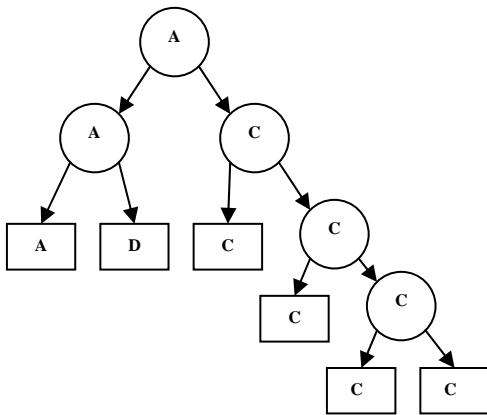


Fig. 1. – Phylogenetic tree for a single sequence position

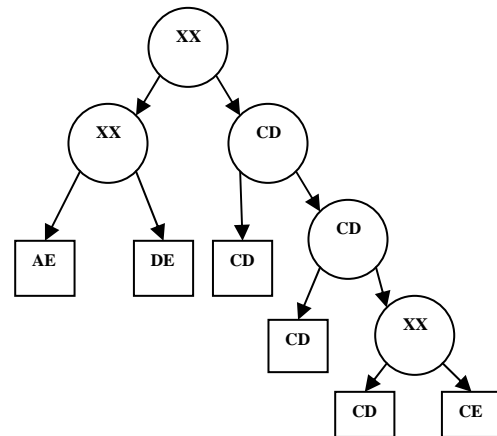


Fig. 2. – Phylogenetic tree for co-evolving pairs

In Figure 2, the tree represents a pair of amino acids found at sequence position x and y . The phylogenetic tree is used to detect mutation events between pairs which becomes the population sample used for probability calculations. The probability based on the number of observed sequences would result in $p(AE)=1/6$, $p(DE)=1/6$, $p(CD)=3/6$ and $p(CE)=1/6$. By using the method described above where we start at the root node and count children nodes that are different with the additional rule that if an internal node is XX it takes on the value of its parent node we get the following probabilities: $p(AE)=1/4$, $p(DE)=1/4$, $p(CD)=1/4$ and $p(CE)=1/4$. The impact of having CD occur 50% of the time is now reduced to 25% which serves as an adjustment to the phylogenetic influence of a mutation that occurs early in the tree and overall

only four distinct mutations occur. The motivation for this approach is to increase the reliability of detecting co-evolving pairs that are $< 12 \text{ \AA}$ apart when measured from the closest non-Hydrogen atoms. The phylogenetic effect on the overall probability scores is reduced by calculating probabilities using observed state transitions.

3 Mutual Information in Protein Families

To illustrate the two approaches mutual information was calculated using log base 2 (STA) for Pfam family PF04055.9. A total of 24 amino acid pairs that are greater than 10 sequence positions apart have a Z score of 4 or greater. Eight of these pairs have a near-contact point $< 12 \text{ \AA}$ and four pairs $> 16 \text{ \AA}$. This does not mean the other 16 MI scores are not significant but it does make it difficult to use the data in determining relative 3D position of two secondary structures. An indication of a false positive will be defined as two amino acid pairs greater than 16 \AA apart. We would like to increase the percentage of MI scores that are $< 12 \text{ \AA}$ apart and reduce the number of well defined false positives.

Using the reduced phylogenetic effect (RPE) method for PF04055.9, 23 data points have a MI score with $Z \geq 4$, 16 of the data points are $< 12 \text{ \AA}$ apart with an average distance of 7.4 \AA . The average overall distance for the RPE method is 9.6 \AA with a standard deviation of 3.8 compared to an average of 12.3 \AA and a standard deviation of 5.0 for the STA method. Reducing the phylogenetic impact in the probability scores in this one example has doubled the number of identified pairs that are $< 12 \text{ \AA}$ with a lower average distance between amino acid pairs.

The two methods of calculating Mutual Information are applied to the Pfam data set where proteins are grouped by family based on Hidden Markov Models. Of the 8,183 protein families in Pfam 19.0, 2,765 families have one or more referenced PDB structures and were used to test the prediction accuracy of detecting co-evolving pairs using Mutual Information. The phylogenetic tree used for each protein family was provided with the Pfam data set and is assumed to be optimal.

Once mutual information is calculated for a protein family the referenced PDB models for that family are used to determine actual 3D position/distance between amino acid pairs as the closest non-Hydrogen atoms. For a PDB model to be used as a reference it was required that the Pfam sequence should align with the PDB sequence by at least 90%. From the 2,765 Pfam families with known PDB structures, 783 families were used that had a family size > 100 and < 5000 with at least one PDB structure that has 90% sequence alignment.

The primary focus is on mutual information scores with a value four times or greater the standard deviation from the mean or $Z \geq 4$ and sequence distance between pairs greater than 10. The percentage of MI scores grouped by Z score that indicate a pair is $< 12 \text{ \AA}$ is calculated for each family. This average prediction percentage represents the likelihood that if we use the same approach in Pfam families that do not have solved PDB models that the MI scores would represent co-evolving pairs. There are numerous attributes associated with the data and by clustering the different dimensions; the goal is to detect additional filter criteria that can be used to increase the quality of the predicted co-evolving pairs that are $< 12 \text{ \AA}$ apart.

By including an additional filter or constraint on the overall number of predicted co-evolving pairs with MI scores, where $Z \geq 2$ is less than 500 increases the $Z \geq 4$ percentages to 56.2% for the STA group. It was determined that if the average number of mutation events between co-evolving pairs for a protein family was less than 40 the prediction of co-evolving pairs was poor. For the RPE group and additional filter is applied where the number of mutation events for a column pair must be > 40 results in improved prediction accuracy of 81.3%. Using the filter criteria, the accuracy of the STA and RPE methods are compared in Figure 3 and Table 1. The RPE average prediction accuracy improved from a reduction in the number of low scoring families for both the $Z \geq 4$ and $Z=3$ groups. The average percentage of predicted co-evolving pairs with a distance greater than 16 \AA (false positive), $Z \geq 4$ and the number MI scores < 500 for the STA group is 26% and for the RPE group is 10.9%.

Table 1- #MI scores per family < 500 and RPE MC > 40

	STA		RPE	
	%<12A	%Pfam	%<12A	%Pfam
Z>=4	56.2	18.3	81.3	15.8
Z=3	42.6	55.8	56.4	46.4
Z=2	33.2	79.7	36.9	71.6

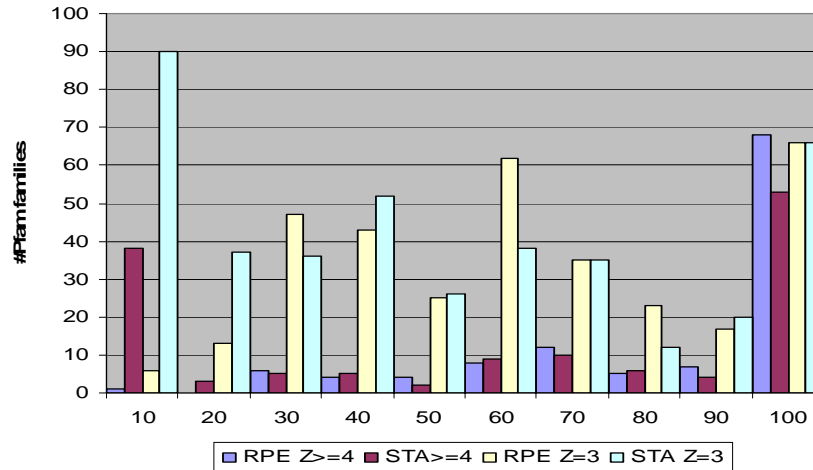


Fig. 1 – Percentage accuracy per Pfam family with #MI scores per family < 500

4 Conclusion

The RPE method of re-sampling sequence data based on mutation events along the phylogenetic tree is an effective approach in improving the quality of predicted co-evolving pairs. Information theory when applied to multiple sequence alignments in protein families can play an important role in detecting co-evolving pairs for tertiary protein prediction, protein engineering and regions of structural and functional importance.

References

- [1] Martin, L., Gloor, G., Dunn, D. and Wahl, L. (2005) Using information theory to search for co-evolving residues in proteins, *Bioinformatics*, 21, 4116-4124.
- [2] Dimmic, M., Hubisz, M., Bustamante, C. and Nielsen, R. (2005) Detecting coevolving amino acid sites using Bayesian mutation mapping, *Bioinformatics*, 21, i126-i135.
- [3] Crooks, G., Wolfe, J., and Brenner S. (2004), Measurements of Protein Sequence Structure Correlations, *PROTEINS: Structure, Function, and Bioinformatics*, 57, 804-810
- [4] Pritchard, L., Bladon, P., Mitchell, J. and Dufton, M. (2001) Evaluation of a novel method for the identification of coevolving protein residues, *Protein Engineering*, 14, 549-555.
- [5] Atchley, W.R. et al. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol. Evol.*, 17, 164-178.
- [6] Hamilton, N., Burrage, K., Ragan, M.A. and Huber, T. (2004) Protein contact prediction using patterns of correlation. *Proteins*, 56, 679-684.
- [7] Crooks, G. and Brenner, S. (2004) Protein secondary structure: entropy, correlations and prediction, *Bioinformatics*, 20, 1603-1611.
- [8] Chandonia, J. and Brenner, S. (2005) Implications of Structural Genomics Target Selection Strategies: Pfam5000, Whole Genome, and Random Approaches, *PROTEINS: Structure, Function, and Bioinformatics*, 58, 166-179