Poster I-12

Predicting co-evolving pairs in Pfam based on Mutual Information of mutation events measured along the phylogenetic tree

Authors:
Scooter willis (*Department of Computer and Information Science and Engineering, University of Florida*)

Short Abstract: 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. The predictive quality with MI(Z>=4), a sequence distance >10 and within 12 angstroms using the RPE method is on average 81% in a Pfam family.

Long Abstract:

<B>Introduction</B>

<P>The accurate prediction of co-evolving pairs in protein sequences plays an important role in tertiary protein structure prediction and protein engineering. The use of mutual information or entropy measures to detect co-evolving pairs is an actively researched topic [1],[2],[3]. The phylogenetic effect on sequence data in a protein family impacts accurate probability measurements when calculating entropy for a sequence position. In this paper, a method is introduced to reduce the phylogenetic effect (RPE method) on probability calculations used to determine Mutual Information (MI) between two sequence positions. Two sequence positions that have a MI score that is equal to or greater than four standard deviations from the mean (Z>=4) for that family are considered statistically significant and potential co-evolving pairs or highly correlated. A co-evolving pair can be defined that when one amino acid mutates a neighboring amino acid in 3D space may also mutate to preserve protein structure or function. The predictive quality of co-evolving pairs that have a sequence distance > 10, with high mutual information (Z>=4) and are less than 12 angstroms apart using the RPE method is on average 81% in a Pfam family. The accuracy of detecting co-evolving pairs in Pfam using MI, where Z>=4 and standard probability measurements is 56% (STA method). This study represents the first known analysis of MI to detect co-evolving pairs in the full Pfam data set. Results for each protein family in Pfam and corresponding ribbon models graphing the MI relationships can be found at <A HREF="http://www.proteinx3d.com">www.proteinx3d.com</A>.</P>

<B>Information Theory Approach</B>

<P>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. 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 for that sequence position. In Figure 1, a binary tree represents a hypothetical phylogenetic tree where the circles are the theoretical parent and the boxes represent a sequence position in a protein family. Without

taking into consideration the phylogenetic influence the probability of the sequence position is p(A)=1/6, p(D)=1/6 and p(C)=2/3. The RPE method calculates probability of the sequence position by starting at the root node and counting all children nodes where the child does not equal the parent (represented by the dashed lines in Figure 1) resulting in probability calculations of p(A)=1/3, p(D)=1/3 and the p(C)=1/3. The first approach accurately represents the population sample and the latter represents the probability of transition to a different amino acid.</P>

<P>In Figure 2, the tree represents a pair of amino acids found at sequence positions 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 pairs between two sequence positions 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 we start at the root node and count children nodes (represented by the dashed lines) that are different with the additional rule that if an internal node is XX it takes on the value of its parent node and is not counted, we get the following probabilities: p(AE)=1/4, p(DE)=1/4, p(CD)=1/4 and p(CE) =1/4. Using the STA method the CD sequence pair occurs 50% of the time but using the RPE method it is reduced to 25%.</P>

<P align="center"><img src="http://www.proteinx3d.com/3dsig/figure.png"></P>

<B>Mutual Information in Protein Families</B>

<P>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. The primary focus is on mutual information scores with a value four times or greater than the standard deviation from the mean or Z>= 4 and sequence distance between pairs greater than 10. The average prediction percentage score for each method 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. The prediction accuracy of MI to detect co-evolving pairs in each protein family is determined and then averaged for all families to determine the overall prediction accuracy of each method. To improve the prediction accuracy, various data attributes of the measurements are used as cutoffs to eliminate false positives. By including an additional filter or constraint on the overall number of predicted co-evolving pairs in a family where the number of pairs with Z>=2 is less than 500 increases the Z>=4 percentages to 56.2% for the STA method. For the RPE method it was determined that if the number of mutation events between co-evolving pairs was less than 40 the prediction accuracy was low. For the RPE method the additional filter is applied where the number of mutation events for a sequence pair must be > 40 and #MI pairs < 500 results in overall improved prediction accuracy of 81.3%. Using the filter criteria, the accuracy of the STA and RPE methods are compared in Table 1.</P>

<P align="center">
<table border="1">
<caption>Table 1- #MI scores per family < 500 and RPE MC > 40</caption>
<tr>
<th colspan="3">STA</th><th colspan="2">RPE</th>

</tr>
<tr>
<td></td><td>%<12A</td><td>%Pfam</td><td>%<12A</td><td>%Pfam</td>
</tr>
<tr>
<td>Z>=4</td><td>56.2</td><td>18.3</td><td>81.3</td><td>15.8</td>
</tr>
<tr>
<td>Z=3</td><td>42.6</td><td>55.8</td><td>56.4</td><td>46.4</td>
</tr>
<tr>
<td>Z=2</td><td>33.2</td><td>79.7</td><td>36.9</td><td>71.6</td>
</tr>
</table>
</P>

<B>Conclusion</B>

<P>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. Future work will focus on using predicted co-evolving pairs using the RPE method in a protein family for tertiary protein prediction and detecting potential regions of structural or functional importance.</P>

[1]Crooks,G., Wolfe,J., and Brenner S. (2004), Measurements of Protein Sequence Structure Correlations, PROTEINS: Structure, Function, and Bioinformatics, 57, 804-810

[2]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.

[3]Hamilton,N., Burrage,K., Ragan,M.A. and Huber,T. (2004) Protein contact prediction using patterns of correlation. Proteins, 56, 679-684.