BBSI Journal Club

Jeffrey Kidd
with
Jennifer Dietrich and Kristin Wheeler
Mentor: Dr. Hugh Nicholas
12 June 2003

Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments

Elisabeth R.M. Tillier and Thomas W.H. Lui

So... what does that mean?

Protein Alignments

- A protein is defined by a specific sequence of amino acids
- There are 20 different amino acids
- The sequence of amino acids is ultimately dictated by the genome (DNA)
- Related proteins will have a similar sequence
Protein Alignments

- The similarity of a set of homologous sequences gives insight into the evolutionary relationships among a group of organisms
  - The fewer differences between two sequences, the more directly related they are
  - Evolutionary changes can result in the substitution of one amino acid for another or in the insertion or deletion of an amino acid

Various algorithms and scoring methods have been developed to compare two sequences and align them so that the best match is achieved.
When searching a database for homologous sequences, statistical calculations are used to decide if the alignment is more likely to result because of chance or because the sequences are related.

Phylogenetic Correlations

- Phylogeny: the evolutionary history of a kind or class of organisms
- One expects organisms descended from a common ancestor to show similarity
- Related protein sequences should also be similar
  - The less similar the sequences, the more distant the relationship
- Sequence alignments can be used to create phylogenetic trees

Functional relationships

- A protein will have some biological function
- This function is ultimately dictated by the sequence of the amino acid residues that make up the protein
  - In turn, is determined by the genome
- The different amino acids have different chemical properties
- The interplay between the different amino acids determines the structure/function
This paper describes a novel technique for separating the functional correlations from similarities caused by the phylogenetic history.

Information theory

- A statistical method is developed for identifying functionally constrained and correlated residues in a multiple sequence alignment
- Concepts from information theory are used

Information theory

- Entropy
  - A measure of the average uncertainty of an outcome
  - Entropy is maximized when the probabilities of each event are equal, and the uncertainty of the outcome is greatest
  - If we are certain of the outcome (P = 1), the entropy is zero
Information theory

- Information
  - Can be considered as the difference in entropy
  - A measure of a reduction in uncertainty after some "message" is received
  - \( I = \text{Entropy}_{\text{after}} - \text{Entropy}_{\text{before}} \)

- Mutual Information
  - One often wants to know "how independent" two random variables are
  - This can be measured by the "relative entropy distance" between \( P(X,Y) \) and \( P(X)P(Y) \)
  - This is the mutual information
  - \( \text{MI}(X,Y) \) can be interpreted as the amount of information acquired about \( X \) when we are told the outcome of \( Y \)

- Multiple Significant Interdependency
  - Interdependency—interactions, or relationships between 2 sites
  - \( \text{MS}(X) \) gives the total amount of interdependencies that site \( X \) has with all other sites in the sequence

- Sites with high MS (multi. signif. interdependencies) values reflect the phylogenetic pattern of substitutions
- To find out how much correlation between sites is not caused by the phylogeny, the mutual information between two sites is divided by their multiple significant interdependencies
- This is the Dependency Ratio
Information theory

- The dependency ratio also identifies sites that are evolving unusually slowly
- The dependency ratio is weighted by entropies
- This entropy-weighted dependency ratio ($R$) may indicate functional correlation
- Standard statistical techniques are used to see if $R$ values are significant

Putting it All Together

- A statistical method (based on Information Theory) was developed to separate phylogenetic from functional correlations
- Sites following the phylogenetic trend will strongly correlate and should have many interdependencies
- Sites with low multiple significant interdependencies do not follow the phylogenetic trend
- May be functional correlations
- Dependency Ratios seem to be a good way to identify functional correlations

Applying this Technique

- Consider the family of Peptide Methionine Sulfoxide Reductase enzymes
- Dr. Nicholas aligned 49 sequences from this family

Bovine Peptide Methionine Sulfoxide Reductase (1FVG)
Applying the Algorithm

- The authors provide a program which carries out the analysis described in the paper.
- The program provides values for Entropy, Multiple Significance Interdependencies, Mutual Information, and the Dependency Ratio.

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$X_j$</th>
<th>$H(X_i)$</th>
<th>$H(X_j)$</th>
<th>$H(X_i) + H(X_j)$</th>
<th>$H(X_i, X_j)$</th>
<th>$R(X_i, X_j)$</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>18</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>14</td>
<td>19</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>16</td>
<td>21</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>17</td>
<td>22</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>18</td>
<td>23</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>19</td>
<td>24</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>21</td>
<td>26</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>22</td>
<td>27</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>23</td>
<td>28</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>24</td>
<td>29</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>26</td>
<td>31</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>27</td>
<td>32</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>28</td>
<td>33</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>29</td>
<td>34</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>31</td>
<td>36</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>32</td>
<td>37</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>33</td>
<td>38</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>34</td>
<td>39</td>
<td>8.9969</td>
<td>6.9031</td>
<td>15.9000</td>
<td>2.0828</td>
<td>6.0987</td>
<td>&lt;0.95</td>
</tr>
</tbody>
</table>
Applying the Algorithm

- Let's only consider the 5 pairs with the highest R values

<table>
<thead>
<tr>
<th>Xi</th>
<th>Xj</th>
<th>H(Xi)</th>
<th>MS(Xi)</th>
<th>H(Xj)</th>
<th>MS(Xj)</th>
<th>MI(Xi,Xj)</th>
<th>R(Xi,Xj)</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>89</td>
<td>0.468065</td>
<td>6.21078</td>
<td>0.362802</td>
<td>6.0987</td>
<td>0.159502</td>
<td>13.9088</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>16</td>
<td>123</td>
<td>0.64777</td>
<td>6.83967</td>
<td>0.660888</td>
<td>14.3255</td>
<td>0.146895</td>
<td>12.9379</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>27</td>
<td>30</td>
<td>0.629301</td>
<td>19.4805</td>
<td>0.600746</td>
<td>18.7008</td>
<td>0.233992</td>
<td>10.9715</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>30</td>
<td>45</td>
<td>0.600746</td>
<td>18.7008</td>
<td>0.640913</td>
<td>24.9238</td>
<td>0.28079</td>
<td>11.604</td>
<td>&lt;0.95</td>
</tr>
<tr>
<td>35</td>
<td>129</td>
<td>0.674117</td>
<td>13.1721</td>
<td>0.676451</td>
<td>8.71802</td>
<td>0.125635</td>
<td>10.8403</td>
<td>&lt;0.95</td>
</tr>
</tbody>
</table>

References