Methods for detecting inter-protein covarying sites

Robert Ackermann and Liqing Zhang

Department of Computer Science, Virginia Tech
Blacksburg, VA

Abstract

Covarying sites are defined to be sites in a protein whose rate of evolution changes over time. We design software to group protein sites into three rate pools: conserved, variant, and temporary invariant. Other software is written to find sites which are closely correlated. The algorithms used by the software require a multiple sequence alignment and phylogenetic tree as input and rely heavily on tree-corrected information entropy. Through a study of the protein Cu, Zn Superoxide Dimutase it is shown that temporary invariant sites have interactions with at least one site which is either closely correlated or binary-switching. From this result it is reasonable to assume that temporary invariant sites which interact with no such intra-protein sites must be sites of protein-protein interaction. Temporary invariant sites are also shows to reflect the animal plant divergence.
Contact: *rackrmnn@vt.edu*

Introduction

Over evolutionary time, certain amino acid sites on a protein almost never mutate whereas others mutate very rapidly. These sites are called conserved and variant respectively, and from this classification the relative importance of the site can be inferred. If a site is conserved it is probably very important to making a functional protein, and thus very important to the fitness of the organism[1].

Some amino acid sites, however, do not remain conserved or variant for the whole period of evolution. For a time they are conserved, but at some point they switch to being variant (or vice versa). These sites are said to be covarying and are neither conserved nor variant, but rather are called "temporary invariant" (TI).

Temporary invariant sites could be significant in a several different ways. One hope is that temporary invariant sites sometimes occur at the locations of protein-protein interactions. Another is that temporary invariant sites represent major events in the evolution of a protein. Finally, taking into account covarying sites has been shown to allow for better simulations of evolution [2].

In this paper, we develop algorithms for finding temporary invariant sites across multiple proteins. Detection requires a multiple sequence alignment and a binary, rooted phylogenetic tree. Using these algorithms, we conduct a study of the protein Cu, Zn Superoxide Dimutase (Cu, Zn SOD) and show that temporary invariant sites have interactions with other sites which are either binary switching or closely correlated to the TI site.

Theory

The algorithms for finding temporary invariant sites make heavy use of information entropy. When applied to the column of a multiple sequence alignment (MSA), information entropy can be used as a measure of how conserved the amino acid site is. Specifically, the equation for information entropy is:

$$H(x) = \sum_{i=1}^{20} \frac{a_i}{|x|} \ln\left(\frac{a_i}{|x|}\right)$$

where x is the amino acid column, |x| is the total number of sequences, and a_i is the number of times amino acid i occurs in the column.

Often in our calculations, we compensate for entropy's dependance on the number sequences by dividing entropy by the maximum possible entropy:

$$G(x) = \frac{H(x)}{H(x_{\max})} = \frac{\sum_{i=1}^{20} \frac{a_i}{|x|} \ln\left(\frac{a_i}{|x|}\right)}{\ln\left(\frac{1}{|x|}\right)}$$

Although this calculation does not completely eliminate dependance on |x|, it does reduce it enough for the algorithms to work. Especially important is that this calculation gives a value ranging between zero and one. For this reason, we refer to G(X) as "percent entropy."

The software uses two different algorithms to detect TI sites. This is necessary due to the extremely unbalanced nature of some phylogenetic trees; one al-

gorithm is designed to be used on relatively balanced trees, while the other works on unbalanced trees. A good guideline for deciding whether a tree is relatively balanced or not is to look at the number of nodes in the root node's left and right subtrees. If the number of nodes in both subtrees is about equal, then the tree can be considered balanced.

Both algorithms use tree corrected entropy, inspired heavily by work done by Mihalek, Res, and Lichtarge[3]. Input to both algorithms is a phylogenetic tree and a MSA. Both algorithms are presented in figure 1 below.

n: a node
right: right subtree of *n*
left: left subtree of *n*
below: right and left subtree of *n*
above: all nodes not in the right or left subtree of *n*
threshold: user defined decimal between 0 and 1

TI Sites Balanced Tree
Input: A balanced, binary phylogenetic tree and MSA
Output: Listing of TI Sites

For each column in MSA
  Assign amino acids to leaf nodes
  For each *n* in the tree
    skip if *right* or *left* contain only leafs
    ent1 = percent entropy of leaf nodes in *right*
    ent2 = percent entropy of leaf nodes in *left*
    if |*ent1* - *ent2*| > *threshold*
      mark *n* as TI
  If any *n* marked as TI
    report column as TI

TI Sites Unbalanced Tree
Input: An unbalanced, binary phylogenetic tree and MSA
Output: Listing of TI Sites

For each column in MSA
  Assign amino acids to leaf nodes
  For each *n* in the tree
    skip if *right* or *left* contain only leafs
    ent1 = percent entropy of leaf nodes in *below*
    ent2 = percent entropy of leaf nodes in *above*
    if |*ent1* - *ent2*| > *threshold*
      mark *n* as TI
  If number of *n* makred as TI ≥ 2
    report column as TI

figure 1: Algorithms to find TI Sites

In both algorithms two entropy percent values are calculated for each node. These two values are subtracted and the result is compared to some user defined threshold for temporary invariance. If the threshold is exceeded, the node is marked as TI. Once all nodes have been marked, a decision as to whether the site as a whole is TI or not is made based on how many nodes marked TI were found.

The main point where the algorithms differ is in how each node gets its two entropy values. In a balanced tree, it makes sense to compare the left and right subtrees of a node. In an unbalanced tree, however, comparing the right and left subtree will often be meaningless. We compare the subtrees of the current node to all other subtrees instead.

```
Homo Sapiens    : A M
Bos Taurus      : A M
Mus Musculus    : Q K
Sus Scrofa      : N V
Equus Caballas  : A M
```

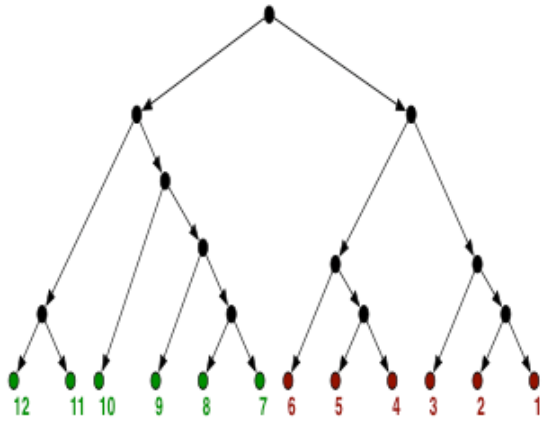
figure 2: Two closely correlated sites in a MSA

Closely correlated sites are sites which have mutations in the same species (figure 2). One final algorithm we created calculates how closely correlated pairs of sites are across several proteins. This was done using mutual information:

$$I(X;Y) = \sum_{i=1}^{20}\sum_{j=1}^{20} f(x_j, y_i)\log\frac{f(x_j, y_i)}{f(x_j)f(y_i)}$$

Where $f(x_j, y_i)$ is the frequency of amino acid i in column x occurring in the same sequence as the amino acid j in column y. Similarly, $f(x_j)$ and $f(y_i)$ are the frequencies of amino acid j and i occurring anywhere in column x and y respectively.

It would be too computationally intensive to simply calculate mutual information for every pair of sites across several proteins. In order to make run time reasonable, before mutual information is calculated the entropy of each column in the input sequence is calculated. Then the columns are placed into clusters using k-mean clustering on entropy (k = 9 by default, but this can easily be changed for large data sets). Once the clusters have been generated, mutual information is calculated for each pair within each cluster, rather than for each pair in the entire data set. All closely correlated sites are still reported, since any sites which are closely correlated will have similar entropies.



1: Sus Scrofa
2: Bos Taurus
3: Rattus Norvegicus
4: Mus Musculus
5: Homo Sapiens
6: Equus Caballas
7: Oryza Sativa
8: Arabidopsis Thaliuna
9: Ipomoea Batatas
10: Spinacia Oleracea
11: Pisum Sativum
12: Zea Mays

figure 3: Phylogenetic tree and species used for the protein Cu, Zn SOD.

Results

To show the importance of TI sites we did a study of the protein Cu, Zn Superoxide Dimutase (SOD). The software was run on a MSA containing twelve orthologous sequences of Cu, Zn SOD, six from plant species and six from animal species (figure 3).

Our software identified six sites which were strongly classified as temporary invariant. Inspection of the sequence alignment by hand showed that the software did not miss any TI sites in the alignment. Site interactions for each TI site were obtained by analyzing the structure of the human version of Cu, Zn SOD[4].

Each of the temporary invariant sites had several interactions with non-adjacent and adjacent sites. Of these interactions, at least one was always with a site that is either closely correlated with the target site or binary switching relative to the target column (Table 1). The term "binary switching" refers to sites which remain as one amino acid through the first few leaves of the phylogenetic tree, but switch to a different amino acid for the last few.

Table 1: TI sites in Cu, Zn SOD, and whether they have an interaction with a binary switching site (Bi), closely correlated site (MI), or an adjacent binary switching site (adj. B)

TI Site	Binary	MI	adj. B
9	Y		Y
21	Y		
24		Y	
27		Y	
136			Y
153			Y

Discussion

It can be concluded from the data that TI sites arise from interactions with sites that are binary switching or closely correlated. In our data, all these interactions are intra-protein because Cu, Zn, SOD is not known to be involved in any complexes or pathways. However, should a protein involved in a complex or pathway be used one could expect some sites to be TI due to protein-protein interactions. In this way, temporary invariant sites reflect not only structure but protein-protein interactions as well.

Also of note is that all the temporary invariant sites were conserved through the plant species and variant through the animal species, or vice versa. Clearly, the major evolutionary event that lead to the creation of plant and animal kingdoms is reflected in the nature of temporary invariant sites. Were the data set expanded to include more species, which will be possible as more sequences become available, it is likely that TI sites will reflect other major events in evolution as well.

We have created effective software for finding temporary invariant sites. In its present state the software is suitable for studies of TI sites if used with discretion. It is hampered by its heavy dependence upon the shape of the phylogenetic tree, and requires several user defined parameters. Another, screening algorithm will be created to remove this dependance on tree shape and allow the user to run the software without the knowledge necessary to set parameters. Further work is also being done to take into account branch lengths, if they are given in the phylogenetic tree.

TI sites have been shown to occur due to interactions between the target site and a site which is either closely correlated or binary switching. In this way TI sites are reflected in protein structure, and could also be indicators of where protein-protein interactions occur. It has also been shown that TI sites reflect the divergence of species into the plant and animal kingdoms.

We expect that they will reflect other major evolutionary events as well. Both of these predictions will be tested as more data becomes available.

Materials and Methods

The protein Cu, Zn SOD was chosen because of its use in a previous paper by Fitch and Miyamato[2]. The phylogenetic tree was also taken from this paper. For structural analysis, the specific protein structure used for this study is entry 1OZU in Protein Data Bank. Orthologous proteins were found by using using Blastp to search genebank, and aligned using ClustalW. Columns containing more than one gap were discarded from the sequence alignment. Intra-protein interactions were found by using contact of structural units software available at http://ligin.weizmann.ac.il/cgi-bin/lpccsu/LpcCsu.cgi[4]

References

1. Simon, Alexander L, Stone, Eric A, & Sidow, Arend. (2002). Inference of functional regions in proteins by quantification of evolutionary constraints. *PNAS.* 99(5):2912-2917
2. Miyamoto, Michael M. & Fitch, Walter M. (1995). Testing the Covarion Hypothesis of Molecular Evolution. *Mol. Biol. Evol.* 12(3):503-513
3. Mihalek, I, Res, I, & Lichtarge, O. (2004). A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance. *J. Mol. Biol.* 336:1265-1282
4. Sobolev, Vladimir & Sorokine, Anatoli et al. (1996). Automated analysis of interatomic contacts in proteins. *Bioinformatics.* 15(4):327-332