DETECTION OF CORRELATED MUTATIONS

Tomáš Ižák
Master Degree Programme (3), FIT BUT
E-mail: xizakt00@stud.fit.vutbr.cz

Supervised by: Tomáš Martínek
E-mail: martinto@fit.vutbr.cz

Abstract: This article describes new algorithm for detection of correlated mutations (positions which mutate together only) in a protein using a phylogenetic tree. The detection is based on physicochemical properties of amino acids, namely charge, polarity, and hydropathy.

Keywords: correlated mutations, proteins, phylogenetic tree, physicochemical properties, amino acids, algorithm

1 PREFACE

Correlated mutations in proteins are amino acids, which mutate together only during the evolution. There are many existing tools using statistical or probabilistic methods, but physicochemical properties of amino acids are usually ignored. Detected correlated pairs are exploited in protein engineering or can give us important knowledge about protein structure or its function. They can indicate structural contact (which helps to fold protein into a tertiary structure) or functionally important site (which binds important substances or other protein molecules). If single amino acid from correlated pair mutates to different one (with different physicochemical properties), protein will lose its function partially or completely. That is why these mutations are not asserted during the evolution.

There are two types of detections of correlated mutations according to input data. First type uses model of tertiary protein structure and is determined especially for structural correlated mutations. This 3D model of protein is gained using X-ray crystallography or NMR.

Other types of detection use a multiple sequence alignment (MSA) as an input. In this case, MSA contains hundreds or thousands sequences with aligned amino acids (positions). To distinguish functional signal (correlated mutations) from phylogenetic signal, phylogenetic tree (PT) is in some methods used (as shown on the picture 1). MSA and PT is used in the algorithm described in the next section.

Physicochemical properties of amino acids are ignored in correlated mutations detection often. Charge of a side chain is important for binding of two amino acids and hydropathy property (if amino acid is hydrophilic or hydropathic) is responsible for protein folding. Hydropathic amino acids are usually located inside proteins and hydrophilic on the surface of proteins. Many current methods treats with serine and threonine as with different amino acids, however better option would be treating with them as same amino acids, because they have same polarity, charge, and almost same hydropathy index.

2 THE ALGORITHM

Algorithms for detection of correlated mutations should meet some requirements:

- elimination of conserved positions - positions in MSA, which do not mutate at all or mutate very little. These positions are not interesting in detection because of small information value.

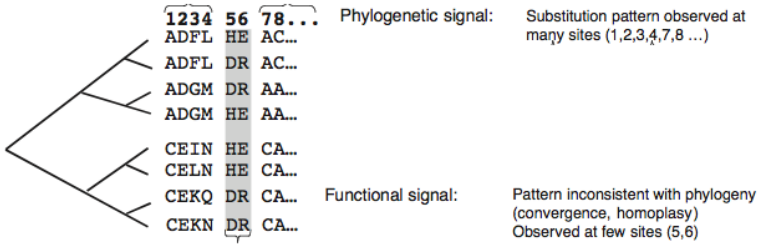

Figure 1: Demonstration of a phylogenetic noise and correlated mutations. Picture taken from [5].

- elimination of a phylogenetic noise - case, when only few mutations took place near root of a PT, which implies a lot of affected sequences. Positions, which contain only this noise, are not interesting in a detection because of small information value.

Approach described below eliminates these two types of positions in one step, which is denoted as evolutionary conservation elimination in this article. As written in preface, MSA and phylogenetic tree is used as an input in the developed algorithm. Output is defined as positions of detected correlated pairs and their correlation ratio.

2.1 ALGORITHM DESCRIPTION

In the first step, amino acids from MSA are mapped onto the inserted phylogenetic tree for all positions. The second step consists of inferring amino acids to predecessors using the Sankoff algorithm with the McLachlan substitution matrix. A phylogenetic tree is available in this moment with amino acid in all nodes for each position in MSA.

The next step is to filter gapped positions, conserved positions, and positions with a phylogenetic noise. Gapped positions are positions with more percentage representation of gaps in MSA than a defined threshold. Conserved positions and positions with a phylogenetic noise are removed during the elimination of evolutionary conservation developed in my work. Evolutionary conservation of position is defined as the sum of mutation flags (1 - if parent and child have same amino acid; 0 - if parent and child have different amino acid) on all edges in the phylogenetic tree, attached to this position, divided by number of all edges. This percentage value is then compared to a threshold and position is eliminated or kept for further analysis.

Unwanted positions are eliminated and every position is represented by PT with inferred ancestors, now. There is also information about presence of a mutation on edges in PT. There is added mutation type information (change of polarity, charge or hydropathy property) for mutation flag also.

Last step is to compare these types of mutations in phylogenetic trees for all pairs of remaining positions. If mutation types on the same edge in compared positions (phylogenetic trees) are different, penalization function will reduce the correlation score of these two positions. Correlation score is computed as follows:

$$correlation\_score = \frac{\sum_{edge \in edges\_in\_PT} 1 - penalization}{number\_of\_edges\_in\_PT} \quad (1)$$

, where penalization value lies in the interval $< 0, 1 >$ and is derived from the McLachlan substitution matrix. Appropriate candidates for correlated mutations are then selected based on computed correlation score. There are still problems (as like as in other methods) with stochastic noise and that's why threshold values can't be very strict.

2.2 COMPLEXITY OF THE ALGORITHM

Time and space complexity of the developed algorithm depends on length of sequences m and number of sequences n. Time complexity is then $T(m,n) \in O(m^2 n)$ and space complexity $S(m,n) \in O(mn)$.

3 TESTING

Testing of the developed algorithm isn't simple at all in this case. Fundamental problem is nonexistence of an accurate definition of correlated mutations, which results in nonexistence of general test sets. Success evaluation depends on subjective decision of each end user (usually biologist). That's why few testing options remain. First option is to compare results of developed algorithm with results of other tools (e.g. CAPS [2] and CMAT [4]), which gives first preview on divergence of results from these tools. According to preliminary tests on six protein families taken from Pfam, CAPS produces similar results to the developed algorithm and CMAT is slightly different sometimes. Next option is to examine detected correlated pairs in the PDB tertiary structure of a protein and to compare these positions with the distance map of the corresponding protein. The less distance detected pair has, the better detection is. But this assumption is not valid for some functional correlated pairs, which can be very distant. Algorithm was tested also for its behaviour in the case of random sequences. For selected protein family, random sequences were added to original sequences. There was an assumption that none of these positions correlates to any other. This assumption was confirmed. Final tests confirmed correct functionality of essential parts of the algorithm as detection of gapped positions and evolutionary conserved positions. Algorithm was tested for certain number of model cases also.

4 CONCLUSION

Detection of correlated mutations is very important for gaining more information about protein structures. It helps to find amino acids, which should not be changed and thus it reduces time required for designing new proteins or editing existing ones. A new algorithm for detection of correlated mutations based on a multiple sequence alignment, a phylogenetic tree and physicochemical properties of amino acids was developed for better results than usual methods like MI [1] or SCA [3]. Output of developed algorithm should fit to its purpose more than other methods due to using physicochemical and phylogenetic information during the whole process of detection, not as supplementary data only.

Web-based prototype is implemented and available on `http://bioware.fit.vutbr.cz/`. User interface is based on HTML and PHP; detection itself is processed by a Perl script through CGI.

REFERENCES

[1] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. `http://bioinformatics.oxfordjournals.org/content/21/22/4116.short`, 2005.

[2] M.A. Fares and D. McNally. Caps: coevolution analysis using protein sequences. `http://bioinformatics.oxfordjournals.org/content/22/22/2821.full`, 2006.

[3] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. `http://www.sciencemag.org/content/286/5438/295`, 1999.

[4] J. Chan-Seok and K. Dongsup. Reliable and robust detection of coevolving protein residues. Protein Eng. 25, 705–713. 2012

[5] E.R.M. Tillier and T.W.H. Lui. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. `http://bioinformatics.oxfordjournals.org/content/19/6/750.short`, 2002.