PROTEIN PRIMARY SEQUENCES AS MARKOV CHAINS

Mitra Chanchal K, Sen Arusharka

1Department of Biochemistry, School of Life Sciences University of Hyderabad, India
e-mail: ckmsl@uohyd.ernet.in
2Department of Mathematics and Statistics, School of Mathematics and Computer Information Sciences, University of Hyderabad, India
e-mail: asensm@uohyd.ernet.in

*Corresponding author

Keywords: protein, primary sequence, Markov chain, sequence analysis, regularity, protein structure

Resume

Motivation:
The protein primary sequence is believed to contain all necessary information for the overall three dimensional folded structure and the functional properties of the protein. Although several empirical algorithms exist that predict the overall folded structure of the protein based on its primary sequence, none has a theoretical basis. We have tried to locate and identify some of the order or regularity present in protein primary sequence using a simple Markov model. A first order Markov model does not show any order or regularity and therefore a higher order model has to be considered. The analysis has been performed using the SwissProt protein sequence databank. The primary motivation came from the simple observation that the amino acid composition differ significantly at various positions of the protein sequence, particularly in the initial region of the sequence and tend to stabilise afterwards (Figure 1).

Results:
We have already computed the first order Markov dependence for all the protein sequences available in the database. We do not find the expected behaviour and the results suggest that first order dependence is not sufficient to explain the observations. Statistical analysis of the results are shown in Figure 2.

Results for the higher order dependence are yet to be compiled and will be presented.

Methods and algorithms
The computations to check for Markov dependence has been carried out as follows:
All fragments (partial sequences) have been eliminated. All sequences smaller than 512 residues have been ignored. The frequencies for all possible pairs are computed \((n_{ij})\) and the frequencies for all the residues are computed \((n_i)\). A test statistic

\[
Z = \frac{\sum_{i,j=1}^{20} (n_{ij} - n_i n_j / n)^2}{n_i n_j / n}
\]

is computed as where \(n\) is the total number of residues in the given sequence. The distribution of \(Z\) is shown in Figure 2.

The computations for equilibrium distributions have been done as follows: All fragments (partial sequences) have been eliminated. All sequences smaller than 512 residues have been ignored. All possible pairs (i.e., 20 x 20) have been counted and stored in a matrix. This matrix is referred to as the pair-frequency matrix. A row with all zeros is eliminated. Each element of this pair frequency matrix is divided by the sum of the corresponding row. The matrix is transposed. The solution of the transposed matrix using the following equation:

\[
\begin{bmatrix}
(P' - I) \cdot x \\
I' \cdot x
\end{bmatrix} = \begin{bmatrix}
0 \\
1
\end{bmatrix}
\]

where \(P'\) is the transposed matrix gives us the equilibrium distribution of the 20 different amino acid residues. The equilibrium distributions are all equal to 1/20, i.e., 0.05, suggesting that all the amino acid residues are equally likely to appear in a sequence in the steady state (near the tail end of the sequence). This result is contrary to common experience.

Discussion

The first order Markov analysis is not suitable for analysis of protein sequences and higher order preferences must be invoked. Alternatively, we can confidently say that it is the long range preferences that play the major role in the protein structure and function.

Acknowledgement

This work has been made possible by a grant from the University Grants Commission, Government of India. Receptor Database (RDB) as an analytical tool for the Drug Design