Abstract

Multiple Protein Sequence is one of the most important problems in modern computational biology. The emphasis here is on the use of computers because most of the tasks involved in genomic data analysis are highly repetitive or mathematically complex. One of the largest areas of Bioinformatics and Data mining has been in the Protein Domain. These efforts have included protein Structure prediction, folding Pathway prediction, Sequence alignment, Substructure Detection and many others. Data storage became easier as the accessibility of large amount of computing power at low cost. The research in bioinformatics has accumulated large amount of data. As the hardware technology advancing, the cost of storing is decreasing. The biological data is available in different formats and is comparatively more complex. In the present work, data mining solution is provided for the problem of protein sequence alignment. Different formats of sequences are studied and plain text format is chosen for the problem under consideration. Clustering methods are based on expressing similarity or dissimilarity of such sequences. The similarity of two protein sequences can be assessed by score of the best alignment of the sequences. Scoring matrix accesses the replacement of one amino acid by another, accepted by natural selection. The replacement can be due the result of two distinct
Implementing Protein Sequence Alignment using PAM-250 Matrices

processes: i) occurrence of mutation in the portion of the gene template producing one amino acid of a protein. ii) acceptance of the mutation by the species (similar function). PAM (Accepted Point Mutations) is the scoring matrix that is used for the different computations. PAM-250 matrix is used for the problem under consideration. The matrix is frequently used to score aligned peptide sequences to determine the similarity of those sequences. The numbers given above were derived from comparing aligned sequences of proteins with known homology and determining the "accepted point mutations" (PAM) observed. Global and Local alignments are predicted along with the alignment score.

References

- Sierk, et al. (2010) "Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments"; Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908 USA, pp. 2-15, vol: 146.
- Luscombe, N. M. , Greenbaum, D. , Gerstein, M. (2001), "What is Bioinformatics? A proposed definition and Overview of the field"; Luscombe group publications, pp 346-
- Myers, Eugene and Miller, Webb "Optimal Alignments in Linear Space"; Department of Computer Science, University of Arizona, Tucson, AZ 85721, NFS Grant DCR-8511455, pp. 1-13
Keywords

Index Terms

Computer Science

Applied Sciences