Abstract

The clustering of biological sequence data is a significant task for biologists. The reason is that sequence clustering assists molecular biologists to group sequences based on the ancestral traits or hereditary information that are hidden in sequences. To accomplish the similarity detection and clustering tasks, several clustering algorithms, similarity and distance measures have been proposed. Most of these algorithms and similarity measures manifest some form of inefficiency in the detection of sequences based on their structural similarity as was observed in the course of this study. In this paper, the codon-based scoring method (COBASM) is developed to handle this inefficiency. COBASM employs the codon principle, by the application of triplet nucleotides, in the clustering of nucleic acid sequences. The results obtained show that COBASM is able to produce compact and well-separated clusters based on the structural similarity of sequences.

References

A Scoring Method for the Clustering of Nucleic Acid Sequences

A Scoring Method for the Clustering of Nucleic Acid Sequences


  Sultana, L. Shade, J. E. Wray, K. D. Wells, and J. Quackenbush. 2002. "Analysis of
  bovine mammary gland EST and functional annotation of the Bos Taurus gene index,"
  interpretation of the Friedman test in the analysis of ordinal- scale data in repeated measures

**Index Terms**

Computer Science  
Bioinformatics

**Keywords**

Codon  
Scoring Method  
Similarity Measure  
Clustering