Abstract

A common task in bioinformatics is the comparison of biological sequences to probabilistic models in order to evaluate their similarity. Completion of genomes of most of the organisms lead to profitable comparative analyses, providing insights into non-coding regions as well as into protein coding regions of DNA. In the present work we propose a method for finding similar sequence in a database of upstream sequences of DNA. For testing purpose, we have extracted upstream sequences of different mammals of citrate synthase and actin genes and also that of cab gene in different plants. The promoter sequences are extracted from NCBI database. Motifs/ TFBS of the upstream sequences are extracted using the software tool ‘TF search’. Then probabilistic models are obtained for motif sequences by HMM method. Query motif sequence can be compared with all the motif sequences in the data base and based on maximum likelihood procedure, degree of similarity between query and all the motif sequences
Promoter Database Search using Hidden Markov Model

is obtained.

Reference

- L. E. Baum, J. A. Eagon, 1967. An inequality with application to statistical estimation for probabilistic functions of Markov processes and to a model of ecology, Bull American Society,
Index Terms

Computer Science

Artificial Intelligence

Key words

Database

Hidden Markov

Model

Promoter sequence

pattern matching

Transcription factors (TFs)

Transcription factor binding sites (TFBS)

Similarity measure