Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 97 (2013) 602 – 611

Procedia
Social and Behavioral Sciences

The 9th International Conference on Cognitive Science

Filtering of background DNA sequences improves DNA motif prediction using clustering techniques

Nung Kion Lee*, Allen Chieng Hoon Choong

Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, 94300, Malaysia

Abstract

Noisy objects have been known to affect negatively on the performance of clustering algorithms. This paper addresses the problem of high false positive rates in using self-organizing map (SOM) for DNA motif prediction due to the noisy background sequences in the input dataset. We propose the use of sequence filter in the pre-processing step to remove portion of the noisy background before applying to the SOM. Our method is motivated by the evolutionary conservation property of binding sites as opposed to randomness of background sequences. Our contributions are: (a) propose the use of string mismatch as filtering threshold function; and (b) two filtering methods, namely sequence driven and gapped consensus pattern, are proposed for filtering. We employed real datasets to evaluate the performance of SOM for DNA prediction after the filtering process. Our evaluation results show promising improvements in term of precision rates and also data reduction. We conclude that filtering background sequences is a feasible solution to improve prediction accuracy of using SOM for DNA motif prediction.

Keywords: Sequence filter; self-organizing map; DNA motif discovery

1. Introduction

Identification of regulatory elements or binding sites bound by transcription factor (TF) proteins are important to the understanding of genetic diseases, protein-DNA interaction, as well as for medical purposes. The interaction of transcription factor proteins and their binding sites regulate which, when, and how rapid proteins are produced. Binding sites are short sequences about 6-25bp long, located in the upstream, downstream or distal locations of genes they regulate. A sequence *motif* is a characteristic nucleotide or amino acid sequence that is conserved in a group of sequences. The problem of computational DNA motif prediction is to predict the exact or approximate locations of binding sites in a set of carefully collected DNA sequences from a genome of species under study and infers the sequence specificities of TFs. The sequence regions a TF's binding sites are likely to be found can be established through wet-lab techniques such as ChIP-seq and microarray analysis or by using comparative genomic approaches. Apparently, each input DNA sequence to a computational tool contains two types of regions: binding and background sequence. Background sequences are noisy because they are highly unstructured due to exposure to random evolutionary events in cells such as substitution, mutation, insertion, or deletion. They are generally regarded as generated by a Markov process in the literature.

* Corresponding author. Tel.: +60 82 584152; fax: +60 82 581579.
E-mail address: nklee@fcs.unimas.my

doi:10.1016/j.sbspro.2013.10.279