BMC Bioinformatics

RESEARCH Open Access

# SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model

Nung Kion Lee, Dianhui Wang*

*From* The Ninth Asia Pacific Bioinformatics Conference (APBC 2011)
Inchon, Korea. 11-14 January 2011

## Abstract

**Background:** Discrimination of transcription factor binding sites (TFBS) from background sequences plays a key role in computational motif discovery. Current clustering based algorithms employ homogeneous model for problem solving, which assumes that motifs and background signals can be equivalently characterized. This assumption has some limitations because both sequence signals have distinct properties.

**Results:** This paper aims to develop a Self-Organizing Map (SOM) based clustering algorithm for extracting binding sites in DNA sequences. Our framework is based on a novel intra-node soft competitive procedure to achieve maximum discrimination of motifs from background signals in datasets. The intra-node competition is based on an adaptive weighting technique on two different signal models to better represent these two classes of signals. Using several real and artificial datasets, we compared our proposed method with several motif discovery tools. Compared to SOMBRERO, a state-of-the-art SOM based motif discovery tool, it is found that our algorithm can achieve significant improvements in the average precision rates (i.e., about 27%) on the real datasets without compromising its sensitivity. Our method also performed favourably comparing against other motif discovery tools.

**Conclusions:** Motif discovery with model based clustering framework should consider the use of heterogeneous model to represent the two classes of signals in DNA sequences. Such heterogeneous model can achieve better signal discrimination compared to the homogeneous model.

## Background

Identification of transcription factor binding sites (TFBS) is fundamental of understanding gene regulations. Binding sites or motif instances are typically 10 ~ 15*bp* in length and degenerated in some positions. They are often buried in a large amount of non-functional background sequences, which causes low signal-to-noise ratio. Hence, using computational approaches to discriminate motif signals from background signals has not always brought satisfactory results. Development of advanced tools is necessary for more accurate motif predictions.

An essence of computational approaches for motif discovery is to search for motifs that are over-represented in the input sequences compared to the background sequences. Motif over-representation can be explained by the existence of segments that have been evolutionarily preserved due to their functional significance to gene regulation. Hence, appearances of motif instances are rather similar to each other despite having variability in some of their positions [1]. Two issues that are closely related to motif discovery problem are: (i) how to construct a model to represent the motifs and, (ii) how to define a suitable search strategy to find putative motifs from the solution space. Position-specific-scoring-matrix (PSSM) [2] and its variations are the most widely used motif model. This model defines the maximum-likelihood estimation on the probability of

* Correspondence: dh.wang@latrobe.edu.au
Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia