SOMIX: Motifs Discovery in Gene Regulatory Sequences Using Self-Organizing Maps

Nung Kion Lee and Dianhui Wang*

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

Abstract. We present a clustering algorithm called Self-organizing Map Neural Network with mixed signals discrimination (SOMIX), to discover binding sites in a set of regulatory regions. Our framework integrates a novel intra-node soft competitive procedure in each node model to achieve maximum discrimination of motif from background signals. The intra-node competition is based on an adaptive weighting technique on two different signal models: position specific scoring matrix and markov chain. Simulations on real and artificial datasets showed that, SOMIX could achieve significant performance improvement in terms of sensitivity and specificity over SOMBRERO, which is a well-known SOM based motif discovery tool. SOMIX has also been found promising comparing against other popular motif discovery tools.

Keywords: self-organizing map, regulatory elements discovery, hybrid model.

1 Introduction

Identification of transcription factor binding sites (TFBS) is fundamental to understand gene regulation. The binding sites or motif instances are typically $10 \sim 15bp$ and degenerated in some positions. They are often buried in a large amount of non-functional background sequences which cause low motif signal-to-noise ratio. Computational discovery of the TFBS (that bind with common transcription protein) from the upstream DNA sequences of co-regulated genes, is regarded as computational motif discovery. Fundamental of these approaches is to search for motifs that are over-represented (over-abundance) in the input sequences compared to the background sequences. Algorithms based on various search strategies have been proposed to discover those over-represented motifs. They include MEME [1], ALIGNACE[2] and SOMBRERO[3]. In this paper, we aim to develop a self-organizing map (SOM) neural network with a customized hybrid node's model for motif discovery.

Standard SOM with weight vector as node model representation has been widely used in biological sequences clustering[4,5]. This representation is inappropriate for our purpose because it requires the input DNA sequences

* Corresponding author.

K.W. Wong et al. (Eds.): ICONIP 2010, Part II, LNCS 6444, pp. 242–249, 2010.