MISCORE: Mismatch-Based Matrix Similarity Scores for DNA Motif Detection

Dianhui Wang and Nung Kion Lee

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

Abstract. To detect or discover motifs in DNA sequences, two important concepts related to existing computational approaches are motif model and similarity score. One of motif models, represented by a position frequency matrix (PFM), has been widely employed to search for putative motifs. Detection and discovery of motifs can be done by comparing kmers with a motif model, or clustering kmers according to some criteria. In the past, information content based similarity scores have been widely used in searching tools. In this paper, we present a mismatch-based matrix similarity score (namely, MISCORE) for motif searching and discovering purpose. The proposed MISCORE can be biologically interpreted as an evolutionary metric for predicting a kmer as a motif member or not. Weighting factors, which are meaningful for biological data mining practice, are introduced in the MISCORE. The effectiveness of the MISCORE is investigated through exploring its separability, recognizability and robustness. Three well-known information content-based matrix similarity scores are compared, and results show that our MISCORE works well.

1 Introduction

Motif refers to a collection of transcription factor binding sites (or simply, binding sites) located in promoter regions of genes. Detection of binding sites is crucial in deciphering gene regulatory networks. In the past years, computational tools have been developed to discover putative binding sites and achieved some promising results. The rapid sequencing of genome data with related species for footprinting resulted in further discovery of many unknown binding sites. In addition, collections of true binding sites obtained using SELEX, chromatin immunoprecipitation (ChIP) or other wet lab techniques have become more readily accessible from several public databases. Examples of these databases are JASPAR [11] and TRANSFAC [14]. Motif models can be constructed from these collections, and it can be used to find potential novel binding sites in DNA sequences using a model specific scoring or similarity function.

Motif detection is a unique classification task because only a small number of positive examples are usually available and a large number of negative examples are ill-defined. This problem is usually referred as one-class instead of multiple

M. Köppen et al. (Eds.): ICONIP 2008, Part I, LNCS 5506, pp. 478–485, 2009.