Algorithms for Molecular Biology BioMed Central

Research

Open Access

EXMOTIF: efficient structured motif extraction

Yongqiang Zhang and Mohammed J Zaki*

Address: Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180, USA

Email: Yongqiang Zhang - zhangy0@cs.rpi.edu; Mohammed J Zaki* - zaki@cs.rpi.edu
* Corresponding author

Published: 16 November 2006

Algorithms for Molecular Biology 2006, **1**:21 doi:10.1186/1748-7188-1-21

Received: 23 July 2006
Accepted: 16 November 2006

This article is available from: http://www.almob.org/content/1/1/21

Abstract

Background: Extracting motifs from sequences is a mainstay of bioinformatics. We look at the problem of mining structured motifs, which allow variable length gaps between simple motif components. We propose an efficient algorithm, called EXMOTIF, that given some sequence(s), and a structured motif template, extracts all *frequent* structured motifs that have quorum *q*. Potential applications of our method include the extraction of single/composite regulatory binding sites in DNA sequences.

Results: EXMOTIF is efficient in terms of both time and space and is shown empirically to outperform RISO, a state-of-the-art algorithm. It is also successful in finding potential single/composite transcription factor binding sites.

Conclusion: EXMOTIF is a useful and efficient tool in discovering structured motifs, especially in DNA sequences. The algorithm is available as open-source at: http://www.cs.rpi.edu/~zaki/software/exMotif/.

Introduction

Analyzing and interpreting sequence data is an important task in bioinformatics. One critical aspect of such interpretation is to extract important motifs (patterns) from sequences. The challenges for motif extraction problem are two-fold: one is to design an efficient algorithm to enumerate the frequent motifs; the other is to statistically validate the extracted motifs and report the significant ones.

Motifs can be classified into two main types. If no variable gaps are allowed in the motif, it is called a *simple motif*. For example, in the genome of *Saccharomyces cerevisiae*, the binding sites of transcription factor, GAL4, have as consensus [1], the simple motif, CGG[11,11]CCG. Here [11,11] means that there is a fixed "gap" (or don't care characters), 11 positions long. If variable gaps are allowed in a motif, it is called a *structured motif*. A structured motif can be regarded as an ordered collection of simple motifs with gap constraints between each pair of adjacent simple motifs. For example, many *retrotransposons* in the *Ty1-copia* group [2] have as consensus the structured motif: MT[115,136]MTNTAYGG[121,151]GTNGAYGAY. Here MT, MTNTAYGG and GTNGAYGAY are three simple motifs; [115,136] and [121,151] are variable gap constraints ([minimum gap, maximum gap]) allowed between the adjacent simple motifs. More formally, a structured motif, $\mathcal{M}$, is specified in the form:

$$M_1[l_1, u_1]M_2[l_2, u_2]M_3 \dots M_{k-1}[l_{k-1}, u_{k-1}]M_k$$