Chapter 41
De Novo Motif Prediction Using the Fireworks Algorithm

Andrei Lihu
Politehnica University of Timişoara, Romania

Ştefan Holban
Politehnica University of Timişoara, Romania

ABSTRACT

De novo motif discovery is essential in understanding the cis-regulatory processes that play a role in gene expression. Finding unknown patterns of unknown lengths in massive amounts of data has long been a major challenge in computational biology. Because algorithms for motif prediction have always suffered of low performance issues, there is a constant effort to find better techniques. Evolutionary methods, including swarm intelligence algorithms, have been applied with limited success for motif prediction. However, recently developed methods, such as the Fireworks Algorithm (FWA) which simulates the explosion process of fireworks, may show better prospects. This paper describes a motif finding algorithm based on FWA that maximizes the Kullback-Leibler divergence between candidate solutions and the background noise. Following the terminology of FWA's framework, the candidate motifs are fireworks that generate additional sparks (i.e. derived motifs) in their neighborhood. During the iterations, better sparks can replace the fireworks, as the Fireworks Motif Finder (FW-MF) assumes a one occurrence per sequence mode. The results obtained on a standard benchmark for promoter analysis show that our proof of concept is promising.

INTRODUCTION

De novo motif finding is crucial to understanding and controlling the regulatory cell processes. Motifs are short (between 6-22 base pairs) putative nucleotide sequences presumed to represent the sites where *transcription factors* (TF) and other regulatory proteins bind to DNA (D'haeseleer, 2006).

The problem of finding DNA motifs is NP-complete (Pevzner & Sze, 2000). The input is represented by a collection of nucleotide sequences in which one must find overrepresented unknown patterns of unknown lengths. The motifs are complex entities because they can be found on both strands of the

DOI: 10.4018/978-1-5225-8903-7.ch041

DNA (i.e. the patterns can be complementary) and can also present mismatches (mutations). If a motif of length w is sought (i.e. a *w*-mer) in n sequences, each of size m, then, without considering possible mutations, there are $(m-w+1)^n$ candidate solutions.

Motifs are represented using profile matrices or the IUPAC consensus notation. Profile matrices, also known as *positional weight matrices* (PWM), are the most commonly used approach and they contain the log-likelihoods of the site-specific frequency of nucleotides against a baseline pre-set background model. These matrices are used to calculate several metrics that can show how far a candidate motif is from a random sequence (Xia, 2012). Moreover, they may be utilized to generate visual representations of motifs as sequence logos, as described by Schneider and Stephens (1990).

Algorithms for motif finding that represent motifs with PWMs are called *profile-based* methods, while those that represent motifs using consensuses are known as *word-based* algorithms (Das & Dai, 2007; Zambelli, Pesole, & Pavesi, 2013).

The profile-based algorithms are search heuristics that iteratively improve an initial PWM. Examples include, but are not limited to: *multiple expectation maximization for motif elicitation* (MEME) by Bailey, Williams, Misleh, & Li (2006); the *Gibbs Sampler* by Hon and Jain (2006), and Lawrence *et al.* (1993) and also *AlignACE* by Hughes, Estep, Tavazoie, and Church (2000).

The word-based algorithms are enumeration methods that search through the space of 4^w candidate solutions. They scan the input sequences for motifs that match up to e mutations. Examples include the algorithm *Weeder* (Pavesi, Mauri, & Pesole, 2001), the *discriminative regular expression motif elicitation* algorithm (DREME) (Bailey, 2011) and the *oligo-analysis* (van Helden, André, & Collado-Vides, 1998).

Typically, the output of *de novo* motif finding methods consists of a list of motifs found (presented in the PWM or the consensus format) ranked according to their significance value.

Evolutionary computation approaches, like *genetic algorithms* (GAs) and *particle swarm optimization* (PSO) have been only utilized meagerly in *de novo* motif prediction. The *genetic algorithm for motif elicitation* (GAME) by Wei and Jensen (2006) and the *genetic algorithm with local filtering and adaptive post-processing* (GALF-P) by Chan, Leung, and Lee (2008) are two examples of algorithms that use standard genetic operators to evolve PWMs. A word-based method that relies on PSO is described in a study by Lei and Ruan (2009). Although such methods claimed good results, they never gained widespread usage among the *next generation sequencing* (NGS) practitioners.

The performance of most *de novo* motif finding algorithms has been historically far from satisfactory. An early study by Tompa *et al.* (2005) that assessed 13 algorithms on 56 eukaryotic datasets has shown that no algorithm surpassed a level of 22% in sensitivity in identifying the binding sites of several known motifs, while another work, a study by Hu, Li, and Kihara (2005), reconfirmed the general low performance. A particularly important but negative aspect is the prevalence of false positives, as discussed in an article by Zia and Moses (2012).

Algorithms for motif discovery can handle promoter and/or ChIP-Seq sequences. Initially, the methods were used for promoter analysis, i.e. analyzing long nucleotide sequences from co-expressed genes. However, the advent of chromatin immunoprecipitation followed by next generation sequencing (ChIP-Seq) posed a new challenge in *de novo* motif finding: although the sequences were shorter, there were massive amounts of data to be processed. Even if several novel algorithms were designed to work with ChIP-Seq data, older algorithms were also adapted to handle more input sequences.