Pattern Recognition Letters 31 (2010) 2147–2160

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Comparison of optimization techniques for sequence pattern discovery by maximum-likelihood

Chengpeng Bi *

Bioinformatics and Intelligent Computing Lab, Division of Clinical Pharmacology, Children's Mercy Hospitals, Schools of Medicine, Computing and Engineering, University of Missouri, Kansas City, MO 64108, USA

ARTICLE INFO

Article history:
Available online 8 September 2009

Keywords:
Maximum-likelihood
Expectation maximization (EM)
Markov chain Monte Carlo
Motif discovery
Multiple local alignment
Gene regulation

ABSTRACT

Among a set of observed relevant DNA sequences coming from a set of co-regulated genes, there exist some short, functional yet hidden sub-sequence patterns which recurrently appear across genomic sequences. The task of sequence pattern discovery, also known as motif discovery, is to uncover these unseen subsequences ab initio and then build a motif model for them. A plethora of motif algorithms has been designed to tackle this problem. This paper aims to compare a set of optimization techniques by consolidating them under the same maximum-likelihood (ML) framework. The framework unifies a suite of motif-finding algorithms by maximizing the same function, that enables a systematic comparison of different optimization schemes as well as provision of practical guidance on using these techniques. As a foundation, the ML framework is built for two categories of iterative optimization techniques (i.e. deterministic and stochastic) capable of exploring the sequence alignment space. The deterministic algorithms are to maximize the likelihood function by performing iteratively greedy local search. The stochastic algorithms are to iteratively draw motif location samples using Monte Carlo simulation and simultaneously keep track of solutions with local maximum-likelihoods. A total of five ML-based sequence pattern-finding algorithms are developed, evaluated and compared using simulated and real biological sequences. Results show that deterministic algorithms are more time-efficient than its stochastic counterparts, but their performance is not as good as the stochastic algorithms.

1. Introduction

Deoxyribonucleic acid or DNA for short is a large biological molecule on which the genetic information is encoded, and it establishes the characteristics of living cells within an organism. DNA can be defined by a linear sequence of simply repeating units (i.e. nucleotides) consisting of three components: a sugar, phosphate, and one of four heterocyclic bases, that is, Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). For example, ATGTTTTCAA is a 10-nucleotide DNA sequence (a single strand). DNA molecule indeed exists as double strands of two nucleotide chains. Two chains are held together by the A–T and C–G base-pairings (i.e. complementarity rule) that are connected by hydrogen bonds. For example, the human genome consists of about 3 billions of these nucleotides. A specific spelling word or subsequence on the genome codes a gene that can be transcribed into messenger ribonucleic acid (mRNA), and then mRNA is translated into a protein molecule. The principle of biological information flow from a piece of genomic DNA (a gene) to mRNA to a protein is called central dogma of molecular biology (Alberts et al., 2002). Notice that not every gene produces mRNA and then codes a protein, instead some genes may produce ribosomal RNA or other small non-coding RNAs or microRNAs. Quite recently, microRNAs (miRNAs) have emerged as a major class of regulatory genes, present in most metazoans and important for a diverse range of biological functions, see a recent review for more details (Rajewsky, 2006). At every moment, a cell has to determine where, when and how much of each gene is to be expressed. To accommodate this need, a genomic DNA contains *cis*-regulatory sequences or *cis*-elements in promoter regions to which regulator proteins, so called transcription factors (TF), can bind, see the illustration in Fig. 1. These proteins either activate or suppress the assembly of the transcription machinery and thereby regulate the expression of genes (Alberts et al., 2002). Different *cis*-elements associated with different TFs can be logically combined by arranging their binding sites (i.e. a *cis*-regulatory module) on the DNA such that the TFs bind cooperatively or exclude each other. Although a functional gene product may be a miRNA or a protein, the majority of known mechanisms regulate protein coding genes. Any step of the gene's expression may be modulated, from DNA–RNA transcription to the post-translational modification of a protein. Gene regulation gives the cell control over its

* Fax: +1 816 983 6515.
E-mail addresses: bi.chengpeng@gmail.com, cbi@cmh.edu

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.09.005

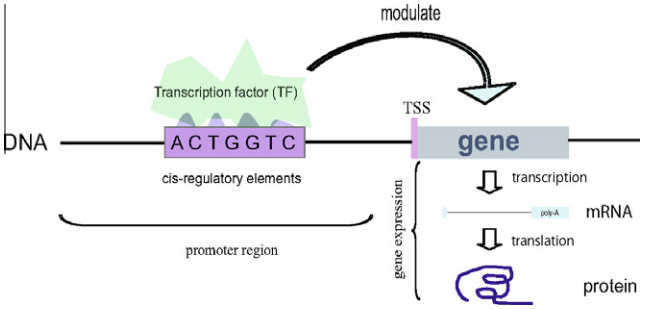

Fig. 1. Components of gene regulation. A transcription factor (i.e. TF) binds to a specific site (transcription-factor binding site (TFBS) or *cis*-element: ACTGGTC) that is either proximal or distal to a transcription start site (TSS) located right upstream of a gene (i.e. promoter region). Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression of which product is a protein. The regulation could be either positive (activation) or negative (repression). A gene is transcribed to messenger RNA (mRNA), and then mRNA is translated into a protein.

structure and function, and it is the basis for an organism to grow and survive.

Gene expression is under strict modulation, that is, it determines how much (expression amount), when (timing) and which tissue (cell differentiation) it will express. Although gene regulation is critically important to understand biological and pathological processes, the detailed mechanism is still not very clear. Therefore, one of the greatest challenges facing molecular biology is the understanding of the complex mechanisms regulating gene expression (Michelson, 2002; Harbison et al., 2004; ENCODE, 2007). Fig. 1 illustrates a simplified gene regulation. A genomic sequence can be roughly divided into coding regions (genes) and promoter/regulatory regions (often upstream of genes). To trigger a gene expression, one or more transcription factors (TFs) first bind to their cognate DNA segments in the promoter region and form the DNA-TF complex, and then this interaction will influence the transcriptional machinery and the gene starts expression. Obviously, identification of *cis*-regulatory sequences (i.e. *cis*-element) is the first step toward unraveling the complex gene expression, which is composed of multiple interacting gene regulation, a genetic network. The current understanding is that much of the specificity of gene expression can depend on how proteins bind to *cis*-regulatory DNA sequences and facilitate or repress the assembly of the transcriptional machinery at the promoter (Fig. 1). Historically, *cis*-regulatory elements have been determined through labor-intensive molecular biology reporter assays and DNaseI footprinting, whereas the potential binding sites of known transcription factors have been determined through in vitro DNA selection assay experiments and gel shift assays (Ji and Wong, 2006; MacIsaac and Fraenkel, 2006). Recently, genome-wide assay of in vivo binding sites is available through the ChIP-chip experiments (ChIP combined with microarray technology) (Ren et al., 2000) or ChIP-seq (ChIP coupled with massively parallel sequencing) (Johnson et al., 2007).

Identification of regulatory regions and binding sites is a prerequisite for elucidating gene regulation. Experimental identification of these sites is expensive, time-consuming and labor-intensive (Stormo and Hartzell, 1989), and therefore developing computational approaches to the challenging motif problem become essential to address these issues. A binding motif is often 4–25 base pairs long and it specifies the binding affinity of DNA–TF interaction. A motif represents a set of related binding sites that can be recognized by the same TF, see an example in Fig. 2A. These binding sites are hidden in a extremely long genomic sequence. If these sites have been experimentally verified, then a model can be built on these sites, either in a consensus or position weight matrix (PWM) format, and it is then used to scan a genomic sequence and predict more potential binding sites. This is known as motif scanning method. However, more challenging motif discovery requires an algorithm that can find binding sites de novo. In other words, given a set of unaligned sequences that potentially contain related motifs, a de novo motif discovery algorithm is designed to locate these related or statistically over-represented binding sites and build a model for them. Despite numerous algorithmic endeavors in an attempt to tackle the motif-finding problem, it still remains extremely challenging, because these biological motifs are short, degenerate and hard to enumerate (MacIsaac and Fraenkel, 2006; Tompa et al., 2005).

There are two categories of algorithms developed to efficiently tackle the de novo motif discovery problem: (1) the word enumerative method, and (2) the PWM updating method including Expectation Maximation (EM) (Lawrence and Reilly, 1990; Bailey and Elkan, 1994) and Gibbs sampling (Geman and Geman, 1984; Lawrence et al., 1993; Liu, 1994). In addition, a greedy algorithm called CONSENSUS was early developed (Stormo and Hartzell, 1989). The word enumerative method has been well developed to grapple with the motif problem. It enumerates all possible spelling words in a set of related DNA sequences and creates a frequency table for all words detected. Potential motifs are those words that would not possibly occur by chance. A suffix tree is often used to accelerate word search (Pavesi et al., 2001). In the word enumerative approach, the exhaustive search strategy seems impractical for long motif and large alphabet size, because the number of all possible words grow exponentially with word length and the alphabet size. Some practical enumeration algorithms often use tradeoffs on the alphabet size and the number of letter mutations allowed, see recent reviews and references therein (Ji and Wong, 2006).

The PWM updating method has been widely adopted and efficiently implemented to solve the de novo DNA motif-finding problem. This method often initializes a PWM motif model by randomly aligning the sequences, and then iteratively refines the model until a convergence. The EM and Gibbs sampling are two major computing techniques used in the PWM updating method. The EM motif algorithms are commonly built on the maximum-likelihood models (Lawrence and Reilly, 1990; Bailey and Elkan, 1994; Bailey and Elkan, 1995), whereas Gibbs motif samplers are rooted on Bayesian computation (Liu et al., 1995; Liu, 2002; Liu, 2002; Jensen et al., 2004). It is thus hard to compare these algorithms and give direct technique guide. Although EM and Gibbs methods can be laid on a common statistical foundation to facilitate their comparison, they were less explicitly investigated in motif discovery problem. This paper builds such foundation on the maximum likelihood framework, which unifies the existing motif-finding algorithms and facilitates a systematic comparison as well as provides practical guidance on using these techniques.

This paper aims to compare a set of motif discovery algorithms by consolidating them under the maximum-likelihood (ML) framework. Several motif discovery algorithms based on the common statistical foundation are then developed to maximize the log-likelihood function by imputing the unobserved data. This would enable the evaluation and comparison of different algorithms and their performance under the same objective function. For *cis*-regulatory decoding or motif discovery problem, the binding sites or motif locations are such unobserved or missing data. Two different optimization schemes are investigated to explore the missing data space. The first is the deterministic algorithms including EM and other greedy heuristic methods that are to optimize a specified objective function by performing iteratively optimal local search in the alignment space. The second is the stochastic algorithms such as Gibbs sampling (Geman and Geman, 1984; Lawrence et al., 1993; Liu et al., 1995; Liu, 2002) and Metropolis–Hastings