Artificial Intelligence in Medicine (2005) **35**, 107—119

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Computational modeling of oligonucleotide positional densities for human promoter prediction☆

Vipin Narang*, **Wing-Kin Sung, Ankush Mittal**

Department of Computer Science, S16 #06-02, 3 Science Drive 2, National University of Singapore, Singapore 117543, Singapore

Received 30 October 2004; received in revised form 31 January 2005; accepted 22 February 2005

KEYWORDS
Promoter modeling;
Bayesian networks;
Regulatory region prediction

Summary

Objective: The gene promoter region controls transcriptional initiation of a gene, which is the most important step in gene regulation. *In-silico* detection of promoter region in genomic sequences has a number of applications in gene discovery and understanding gene expression regulation. However, computational prediction of eukaryotic poly-II promoters has remained a difficult task. This paper introduces a novel statistical technique for detecting promoter regions in long genomic sequences.
Method: A number of existing techniques analyze the occurrence frequencies of oligonucleotides in promoter sequences as compared to other genomic regions. In contrast, the present work studies the *positional densities* of oligonucleotides in promoter sequences. The analysis does not require any non-promoter sequence dataset or any model of the background oligonucleotide content of the genome. The statistical model learnt from a dataset of promoter sequences automatically recognizes a number of transcription factor binding sites simultaneously with their occurrence positions relative to the transcription start site. Based on this model, a continuous naïve Bayes classifier is developed for the detection of human promoters and transcription start sites in genomic sequences.
Results: The present study extends the scope of statistical models in general promoter modeling and prediction. Promoter sequence features learnt by the model correlate well with known biological facts. Results of human transcription start site prediction compare favorably with existing 2nd generation promoter prediction tools.

☆ Availability: Binary executable of the promoter prediction model, named BayesProm, is available at: http://www.comp.nus.edu.sg/~bioinfo/BayesProm (accessed: 1 May 2005).

* Corresponding author. Tel.: +65 687 410 17; fax: +65 677 945 80.
E-mail address: vipinnar@comp.nus.edu.sg (V. Narang).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2005.02.005

1. Introduction

In eukaryotic cells, protein coding genes are transcribed by the RNA polymerase II enzyme. However, for initiating transcription, assistance from a number of DNA-binding proteins called transcription factors is required. Promoter is defined as the regulatory DNA sequence region containing all binding sites for the transcription factors involved in transcriptional initiation with RNA polymerase II. It surrounds the transcription start site, and is present mostly in the upstream region. As the post-sequencing era of genomics is focusing on analysis of the wealth of genome sequence data in hand, a reliable computational tool for the detection of eukaryotic poly-II promoters in genomic sequences has several potential uses, such as in determining the 5′ end of genes, locating the first exon within genomic sequences, study of gene expression regulation, and so forth. Although a number of computational promoter prediction tools have been developed, the performance is still inadequate and the problem remains open for research [1,2].

Computational promoter prediction involves differentiating promoter versus non-promoter regions in a given genomic sequence, and predicting the locations of transcription start sites (TSS). The computational algorithm has two aspects: (i) recognition of transcription factor binding sites/motifs, and (ii) modeling the combinations and context of these binding sites within promoter sequences. While a lot is known about various transcription factor binding motifs that play an active role in eukaryotic poly-II promoters [3], and a number of methods are available to recognize these motifs, a great amount of diversity has been observed in the organization of promoters. The diversity and complexity of promoter sequences makes general computational promoter prediction a challenging task.

Several existing tools [4—6] have utilized positional weight matrices (PWM) derived from experimental data [7] for detecting putative transcription factor binding sites within a given genomic sequence. Earlier tools such as Autogene [4], Promoter Scan [5] etc. identified sequence regions with a high density of binding sites (detected using PWM) as possible promoters. However, research has shown that the location and combination of the binding sites is also important in a promoter [8,9]. A recent tool, Eponine [6], significantly improved the quality of promoter predictions by associating with each PWM the probability distribution of its position relative to the TSS.

Another category of tools [10—12] recognize promoters based on their sequence composition. Characteristic features of promoter sequences are learnt automatically from a set of training examples using machine learning or statistical techniques. An unknown sequence is then classified as promoter or non-promoter based on its feature content. For example, PromFD and PromFind algorithms [10,11] analyze occurrence frequencies of oligonucleotides in promoter versus non-promoter training datasets. Oligonucleotides with a significantly higher rate of occurrence in promoter (non-promoter) sequences are identified as promoter (non-promoter) specific features. An unknown sequence region is classified as promoter or non-promoter based on a scoring scheme that determines the quantitative differential between its promoter versus non-promoter specific oligonucleotide content. PromoterInspector [12] adopts the same method, but using IUPAC groups, which are oligonucleotides permuted with wildcards.

Most existing computational promoter prediction tools are based either on positional weight matrices or on oligonucleotide occurrence frequency analysis. About 6—10 years ago, the first generation of tools could predict 30—40% of the actual TSS, while reporting one false positive every 1000 bp [1]. Recent research has focused on achieving improved TSS prediction performance through better tuning and increased modeling complexity. This has resulted in some 2nd generation tools [2], such as PromoterInspector, Eponine, Dragon Promoter Finder [13], etc. whose accuracy is suitable for whole genome scale prediction. In addition, biologically motivated approaches such as CpG+ [14] and gene start finding tools such as First Exon Finder [15] and Dragon Gene Start Finder [16] have exploited features such as CpG islands and first splice donor sites to further improve the accuracy of TSS prediction.

An important but relatively less explored approach for promoter prediction is using purely statistical models. An early attempt in this direction by Audic and Claverie [17] used simple Markov chains of order four to six to model promoter sequences. However, the authors reported low performance of the model due to its simplicity and its overfitting of training promoter sequences. Ohler et al. [18] used interpolated Markov chains, which is a generalization that combines several simple Markov chains of different orders. It takes into account statistics of higher orders without overfitting the model to training data. Ohler et al. [18] initially reported performance equivalent to first generation promoter prediction tools. However, improved results have been reported recently upon retraining the model on a larger dataset of *Drosophila* core promoters [19]. In a slightly different context of locating regulatory regions in genomic sequences