

Application of Random Forest Classifier to DNA Promoter Site Data

N. Butuk

J. Wetiba

Department of Mathematics

H. Howard-Lee

Department of Biology

Prairie View A & M University

ABSTRACT

In this paper we address the problem of promoter site recognition in eukaryotes DNA sequence. We apply a novel approach of the recently introduced random forest classifier that has been shown to outperform neural networks. Preliminary results are presented of our long term effort of developing efficient promoter site recognition software module. Our approach involves combination of several advanced mathematical techniques to develop an efficient module to be incorporated into gene recognition software. Gene recognition is essential to understanding existing and future DNA sequence data. During transcriptional initiation, there is a large variety of transcription factors interacting and cooperating in promoter regions in complex ways. To answer the question of how genetic information is processed, promoter identification becomes a necessary step, especially in eukaryotes in which the promoters are involved in various biochemical processes. Developing computational methods to find promoter sequence patterns is therefore vital for achieving the goals of for example the Human Genome Project.

1. INTRODUCTION

The recent availability of long genomic sequences, opens a new field of research (genome informatics) devoted to the analysis of their structure and function. This has led to rapid evolution of this new field of genome informatics. The major research topics in this field include gene recognition, functional analysis, structural determination, and family classification for the identification of genes. More specifically, the field involves identifying protein/RNA-encoding genes, recognizing functional elements on nucleotide sequences, understanding biochemical processes and gene regulations, determining protein structures from amino acid sequences, and modeling RNA structures, as well as performing comparative analysis of genomes and gene families. This paper is concerned only with a subset of the topic gene recognition.

Gene recognition involves the identification of DNA functional elements as well as signals recognized by the transcriptions, splicing and translation machinery. The functional elements include promoters, exons (initial, internal, and terminal exons), introns, 5' and 3' untranslated regions, and intergenic regions.

Gene identification typically has two phases; coding region prediction and gene structure determination. Separate modules (i.e. content and signal sensors) are built to capture features of the individual DNA functional elements and signals, and then combined into a gene model. The major modules are those designed for promoters, exons, introns, and intergenic regions, as well as for splice sites and start sites. The main mathematical models used in most gene software include hidden Markov models, neural networks, linear discriminate functions, and decision

trees. Dynamic programming is usually used for optimal gene structure construction. Currently a number of gene prediction programs are available such as GeneID, GeneMark, FGENEH, GRAIL, Gene Parser, Genie, GENSCAN, GeneDecoder and MORGAN [Wu and Mclarty, 2000].

Computational methods for automated DNA sequencing (genome annotation) are critical to understanding and interpreting the bewildering mass of genomic sequence data presently being generated and released. In a few years, sequencing new genomes and individuals will become routine practice. This raw data is not immediately useful and interpreting it places major demands on the development of efficient computational algorithms.

In this paper, we shall address the development of a new algorithm to be used as a module in gene identification software for the recognition of promoter sites.

There are two approaches to developing modules for promoter site recognition; *ab initio* methods and homology based methods. *Ab initio* involves application of statistical modeling of genomic sequence alone i.e. it relies on properties intrinsic to nucleotide sequences. In homology-based gene finding, alignment methods are used. Programs capable of aligning nucleotide-sequences to DNA databases (such as FASTX) are used. This approach however, relies on homologous genes being known. Hence, novel genes cannot be found using this approach, [Zien, et al., 2000].

In this paper we will address the *ab initio* approach for the recognition of promoter regulatory sites in higher eukaryotes. This is a challenging task in DNA sequence analysis. Promoter sites typically have a complex structure consisting of multiple functional binding sites for proteins involved in the transcription initiation process. This region spans up to 300 basepairs upstream and 50 base pairs downstream of the transcription start site (TSS). Successful promoter recognition will enable major advances in multi-gene recognition. Also knowledge of TSS facilitates locating the correct initiation codon resulting in better coding gene (exons) prediction. Better TSS recognition will also lead to a better understanding of the structure and mechanism of regulatory elements and the entire gene regulatory process.

Earlier work on promoter site recognition, involved statistical studies of polymerase II promoter regions in eukaryotes, where statistical weight matrices based on counts of specific nucleotide at a fixed position were constructed. Matrices for each individual elements such as the TATA box were constructed. The use of Neural Networks (NN) have played a significant part [Larsen, et al., 1995]. Fickett and Hatzigeorgions, (1997) have presented an overview of these early approaches as well as a state of the art review. They concluded that the problem of eukaryotic promoter prediction is far from being solved. Recently Reese, et al., (2000) have developed a time-delay neural network (TDNN) for the recognition of promoter sites in the fruit fly genome that was recently completed in 1999. Appropriate representative data sets was obtained from the Eukaryotic Promoter Database (EPD) and the Genebank.

Our approach to developing an efficient module for promoter site recognition combines several mathematical advanced techniques to develop a novel module. The techniques to be used include; fractal image generation, evolutionary computation, image processing techniques and probabilistic methods as well as state of the art classification algorithms. In section 2 we give a brief description of fractal image generation using DNA sequence data, as well as an image processing technique for representing generated fractal images. In section 3, we shall give a brief description of the Hidden Markov Model for pre-processing promoter site data. In section 4 we shall describe the Random Forest Classifier and in section 5 we give the biological significance

of the promoter site. In section 6 we shall present the results and finally give conclusions in section 7. The data sets used in this work appear in the appendix.

2. FRACTAL DNA IMAGE GENERATION

In this paper, we have used the *similitude* type of iterated functional system, IFS to generate DNA fractal images. It is a type of linear mapping which is actually a composition of three simpler mappings. It includes a scaling factor, a rotation about the origin through an angle, and a translation. In standard matrix form it is written as

$$T \begin{pmatrix} x \\ y \end{pmatrix} = s \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

where s , θ , e , and f are fixed scalars for a particular similitude.

To generate a fractal image, we need a certain number of these similitude's (i.e. T_1, T_2, \dots, T_n). Ashlock and Golden, (2003) were the first to use similitude IFS to visualize DNA sequence data. Our approach follows essentially their method of evolving optimum similitudes. However, we use a novel method of evaluating the fitness function of our evolutionary algorithm. Also, to our knowledge we are the first to apply this evolutionary approach to promoter sequence recognition.

We used 8 similitudes, T_1, \dots, T_8 each of which represented one of the 8 RGB colors; Black, Red, Green, Blue, Cyan, Magenta, Yellow, and White. The idea is to use a DNA sequence to drive a fractal image-generating algorithm consisting of eight similitude IFSs. Obviously, the question then becomes that of choosing the best set of similitudes that will generate distinct images to be able to distinguish promoter sequence images from other DNA functional elements. With four real parameter s, θ, e, f that describe a given similitude, there are an infinite number of similitudes that can be specified. Our purpose is to search this huge space of similitudes for the optimum set of similitudes that can be used for promoter region recognition. To search this space we used the evolutionary computational approach of Genetic Algorithms (GAs). The search parameters for the GA are s, θ, e, f with ranges.....