Gene Prediction Algorithm

1. Introduction

Genes are regions in DNA, which can express themselves. These expressions are done by protein synthesis. This project is intended to develop an algorithm to identify regions in Genomic DNA that encode proteins. This is achieved by distinguishing genetic regions, which is otherwise known as ‘coding regions’ from non-genetic / dormant regions also known as ‘non– coding regions’ in a long of string of DNA using one of the well known Mathematical Models.

DNA of an organism contains sequences which are Genes and Gene related, as well as Extragenic ones. The former is further sub divided into coding-DNA, or Exons and Non-Coding DNA, or Introns. Gene prediction is a method of identifying probable positions of Exons in DNA. Exons are classified into four classes. They are: 5’ exons, internal exons, 3’ exons and intronless exons. The 5’ exons are characterized by a Transcriptional Start Site (TSS) on one end. The 3’ exon has Poly(A) sequence on one end, whereas Intronless exons have TSS on one end and Poly(A) on the other. The Internal exon (its exon) has neither a start or a stop sequence. Genes can be identified by the presence of a promoter (during transcription phase) or by the presence of Poly(A) (AAUAAA) sequence.

The complexity of identifying an Exon in DNA sequence can be further illustrated using human genome as an example. A human haploid Genome is 3000 Million Base Pairs long. Of this 3000 million Base Pairs only 90 Million Base pairs are identified to be coding DNA. ‘Finding these indistinct ‘needles’ in a vast genomic ‘haystack’ can be extremely challenging. In response to this challenge, computational prediction approaches have proliferated, that predict the location and structure of the genes .’ [Zhang Michael Q] Bioinformatics combines expertise and knowledge of people from biological and computational areas, and set a common stage for people from these background to work together to solve intellectual and practical challenges like this. There is a wide variety of Computational Mathematical model used for prediction of gene structure in organisms. The most commonly used ones are listed and explained below.

1. Linear Discriminant Analysis and Quadratic Discriminant Analysis:
   Statistical pattern-recognition methods that are used to categorize samples into two classes. Once samples have been represented as points in space, linear discriminant analysis (LDA) finds an optimal plane surface that best separates points that belong to two classes. Quadratic discriminant analysis (QDA) finds an optimal curved (quadratic) surface instead. Both methods seek to minimize some form of classification error.

2. Hidden Markov Models:
   Hidden Markov Models (HMMs) represent a system as a set of discrete states and as transitions between those states, each of the possible transitions having an associated probability.

3. Hexamer-Coding Measures
   Some methods interpret sequences as successions of words (so-called because nucleotides are not independent of each other, but tend to occur together as if in a word) of length k (k-tuples); 6-tuples are called hexamers. In-frame hexamer frequencies in a region of DNA have traditionally been used as a powerful way of discriminating coding regions from non-coding regions, as some words are more likely to be present in either type of DNA.

4. Weight Matrix Method and Weight Array Method
Semester Project Proposal

5. **Decision Tree**
A classification scheme, which can be used, for example, to split a sample into two sub-samples according to some rule (feature variable threshold). Each sub sample can be further split, and so on.

6. **Artificial Neural Networks**
The key element of the artificial neural network (ANN) model is the novel structure of the information processing system. It is composed of many highly interconnected processing elements that are analogous to neurons and are tied together with weighted connections that are analogous to synapses. Once it is trained on known exon or intron sample sequences, it will be able to predict exons or introns in a query sequence automatically.

[Samuel Costa, Ramiro. 2003]

2. **Problem Statement**
Develop an algorithm which will help to distinguish between coding regions and non-coding regions in a given string of DNA sequence, and thereby be able to predict the presence of a gene with fairly reliable accuracy.

3. **Related Researches**
Since the science of DNA and genes to this extent is fairly new to me I have spend much time understanding more about DNA and DNA sequencing before starting any analysis into deeper subjects like Gene prediction, promoters, gene expression etc. I started this out by browsing internet sites which explains in layman’s term about DNA and DNA sequencing and also the human genome project. Sites like www.ornl.gov/sci/techresources/Human_Genome gave me a starting point to my long and exciting journey into the much interesting world of DNA, Genes and Bio science in general. Each of my steps made me more humble before the infinite wisdom of creation, and made me realize that human kind has a very long voyage ahead, before we could claim to understand the complex rules (algorithms, if you may) that guide each function of a living organism. All we know is that, this is controlled by codes inside DNA sequence called Genes. Before this sudden realization draws me into the ever-pervasive world of spirituality, I immediately gained back my sense of duty and lead myself into reading and trying to understand more about Gene characteristics. This, I started with Protein synthesis of DNA. Genes are believed to be those areas of DNA which is capable of producing protein through a series of steps, of which processes like transcription and translation are the most important ones. One should understand these processes before trying to learn about Gene expression. The hand outs that were given in the class and the class lecture itself, helped me to have a base in understanding these processes. Later on when I went into exploring more on the new idea of Gene expression I came across something called promoters. A promoter is a DNA sequence to which RNA POL with its associated factors binds prior to the initiation of transcription. I procured some papers written on promoters with the help of my friend, one of which is written by himself. These papers gave me an idea about promoters and their importance in identifying Gene expressions. I also came across a power-point presentation called ‘Annotation of Genomic Sequence’ which also gave me useful information on Genes, Exon boundaries & splice sites, Beginning and end of translation, Regulatory elements etc. All these readings gave me convincing evidence that with the help of computational methods we can speed up the process of unfolding some of the mysteries that are hidden inside the gene sequences of living organisms. I then came across an article called ‘Computational Prediction Of Eukaryotic Protein-Coding Genes’ by Michael Q. Zhang. This article gave me authoritative information on structure and classifications of Exons and Introns. It also explained in detail different mathematical models used for gene prediction and also talked about some of the programs that are developed using these mathematical models. My further reading was a closer look on to these Mathematical models. An article by Ramiro Pablo Costa called ‘Gene Prediction Algorithm’ was my starting point. I then came across a paper on ‘Applications of hidden Markov models for comparative gene structure prediction’ by two students from University of Aarhus. Even though I did...
not understand much about the model, it gave me a pretty good idea on how to use mathematical models in biological research. By this time I made up my mind to use Linear Discriminant Analysis for developing my algorithm. This lead me to the article ‘Linear Discriminant Analysis - A Brief Tutorial’ by S. Balakrishnama, A. Ganapathiraju of Mississippi State University. This paper gave me an insight into the model and gave me a starting point for my Project.

4. Method

   Linear Discriminant Analysis (LDA)

5. Approach

   This gene prediction algorithm is going to be developed using the Mathematical Model, Linear Discriminant Analysis. Since this model is a statistical pattern-recognition method that is used to categorize samples into two classes, my algorithm will be developed to identify differences between two Training datasets, one of which will be a positive training set (like an identified Exon) and the other will be a negative training set (like psuedoxons). Once the samples have been represented as points in space and an optimal plane surface that best separates points that belong to two classes is identified, the sample DNA strand on which the Gene prediction to be done will be introduced. By identifying the characteristics closer to the positive training sets the gene prediction can be done fairly accurately.

6. Proposed Schedule

   02/12/04 Submit Proposal
   02/19/04 Receive final approval of the Proposal
   02/20/04 – 03/04/04 Project Startup and Planning
   03/05/04 – 03/22/04 Research and Analysis
   03/23/04 – 04/12/04 Design and Coding
   04/13/04 – 04/30/04 Unit and Integration Testing
   05/03/04 Implementation and Submission

7. References

   6. Introduction to DNA. [http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/pg1.html]
   7. Human Genome Project information. [www.ornl.gov/sci/techresources/Human_Genome]