CITY UNIVERSITY OF HONG KONG

香港城市大學

Detection of Periodic Signals in DNA and Protein Sequences

在脫氧核糖核酸和蛋白質序列中檢測週期信號

Submitted to

Department of Electronic Engineering

電子工程學系

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

哲學博士學位

by

Song Yu

宋愚

September 2013

二〇一三年九月

ABSTRACT

Proteins play critical roles in life. Protein coding genes in eukaryotes can be long and are usually "split" into discrete DNA segments called introns and exons. Exons will be spliced together, while introns are discarded, to make mature messenger RNA that can be translated to form proteins. Exon detection in DNA sequences is an important task. Laboratory experiments to detect exons are laborious and expensive to conduct. Therefore, there exists a great demand for computational exon detection methods. Most exons in the human genome are short, while introns, and therefore genes, are normally much longer. Short exons are hard to detect by computational methods because their signals are weak.

In this thesis, a new approach is proposed for short exon detection. Exons usually contain a periodicity of 3, as 3 consecutive nucleotides are biologically translated to an amino acid, which makes exon detection possible by applying signal processing methods. In our method, symbolic DNA sequences are first converted into numerical signals based on four DNA structural properties, which are obtained from physical models or biological experiments. These conversions are biologically more meaningful than those based on subjectively assigned numbers. Then an autoregressive (AR) model is applied to find genome regions which contain the periodicity of 3. This AR model-based analysis is able to produce stronger power spectral density (PSD) peaks and weaker artifacts than the discrete Fourier transform (DFT). To overcome the non-stationarity of DNA sequences, we use moving

windows of different sizes in the AR model. Experiments on the human genome show that our multi-feature based method works well for short exon detection.

Despite the good performance,the above AR model based approach has a high computational load. To overcome this, the DNA structural features are mapped to a new set of values. The three signals generated by the mapped feature values are normalized and averaged before their power spectral density is estimated. This substantially reduces the computational load while still maintains a good accuracy for exon detection.

The AR model is also utilized to distinguish solenoid and non-solenoid proteins (proteins that do or do not have specific repeated structural domains). Traditional signal processing methods cannot detect interspersed repeats. Therefore, AR model is firstly applied to find possible repeating subsequences within a protein sequence. Then an iterative hidden Markov model (HMM) is utilized to identify interspersed repeats in the protein sequence. Experimental results show that this method substantially improves the performance of solenoid protein recognition.

CONTENTS