Digital Signal Processing for Human Genome Sequencing

Author: Alpha Gould  
Supervisor: A/Prof E. Ambikairajah  
Assessor: Dr. E. Aboutanios

Motivation for Genomic Signal Processing

If gene detection can be automated, new genes can be found more easily, leading to a more complete biological picture. It is hoped that once we have a greater understanding of genes we will be better able to fight disease.

This thesis uses signal processing techniques to propose a new method of gene detection (based on wavelet packet decomposition) and several numerical representations.

Biological Background

DNA

DNA is a helical structure made up of two sugar-phosphate backbones and four nucleotides: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T).

Genes

Genes encode the information for building proteins and are ultimately made up of amino acids. Amino acids are made up of triplets of nucleotides, known as codons. It is this fact that is the basis for genomic signal processing as it shows that genes have basic periodicity of 3.

Genomic Signal Processing

Methodology

The generalised methodology for genomic signal processing involves three steps as shown below.

Proposed DNA Representations

DNA sequences are represented numerically so that they can be analysed. The author has proposed five numerical representations for DNA, three using real numbers and two using complex numbers.

An example of the proposed numerical representations is $A = 2$, $C = 3$, $G = 3$, $T = 2$. For the sequence $x = AGTTCCGA...$, the four indicator sequences are:

- $X_A = 2,0,0,0,0,0,0,2...$
- $X_C = 0,0,0,0,0,3,3,0...$
- $X_G = 0,3,0,0,0,0,3,0...$
- $X_T = 0,0,2,2,2,0,0,0...$

These sequences could be combined to produce:

- $X_{C,G} = 0,3,0,0,0,3,3,0...$
- $X_{A,T} = 2,0,2,2,2,0,0,2...$

Conclusions

The wavelet packet method was the third most successful feature extraction method and with proper tuning and perhaps an alternative kind of wavelet significant improvements could be made.

Of the proposed representations, one of them is an improvement on a well established representation. This could be due to its representing physical features of DNA.