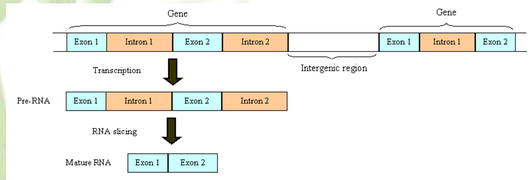
Digital Filters for Gene Prediction Applications

Ebrahim Abunasrah and Oleksandr Babenko
Advisor: Dr. Nidhal Bouaynaya
Department of Systems Engineering, University of Arkansas at Little Rock

Problem Overview

- The genes of Eukaryotic organisms are spliced into segments, called *exons* and segments called *introns*.
- During translation of the gene into protein, only the exons regions are decoded. The introns region is spliced away and its nucleotides might be used in other functions of the cell. Hence, the concatenation of the exons of a gene is called the *coding region* of the gene, whereas the concatenation of the introns is called the *non-coding region* of the gene.
- Biological determination of the exons and introns or coding and non-coding regions in a gene are extremely difficult, because there are no obvious biological markers.
- Digital signal processing provides powerful and efficient tools for the analysis of genomic data.

Gene Structure of Eukaryotic Organisms



Numerical Representation of DNA Sequences

- Genomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities.
- DNA molecules as well as well as **proteins** can be represented by numerical sequences.
- We create four binary sequences, one for each character (base), which specify whether a character is present (1) or absent (0) at a specific location. The resulted sequences are known as indicator sequences.

DNA Sequence	···	A	T	T	G	C	A	C	C	G	T	G	A	···
Indicator seq. for A	···	1	0	0	0	0	1	0	0	0	0	0	1	···
Indicator seq. for T	···	0	1	1	0	0	0	0	0	0	1	0	0	···
Indicator seq. for G	···	0	0	0	1	0	0	0	0	1	0	1	0	···
Indicator seq. for C	···	0	0	0	0	1	0	1	1	0	0	0	0	···

Spectrum Analysis of DNA Sequences

- The Discrete Fourier Transform (DFT) of a finite sequence *x[n], n= 0,…,N-1,* is defined as

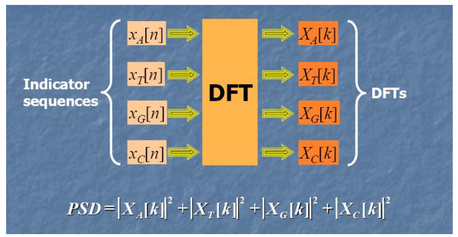
$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-i\frac{2\pi}{N}kn}$$

The DFT power spectrum at frequency k is:

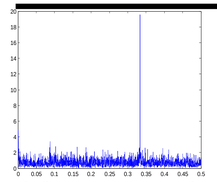
$$PS(k)=|X(k)|^2, k=0,...,N-1$$

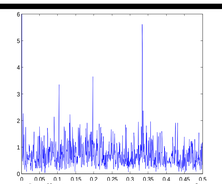
➡ The sequence *PS*(*k*) provides a measure of the frequency content at frequency *k*, which corresponds to an underlying period of *N/k* samples.

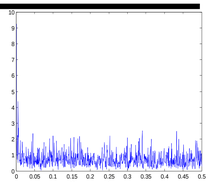
- The power spectrum of the genomic signal is computed as:

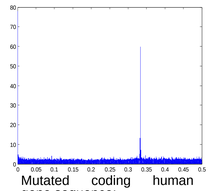


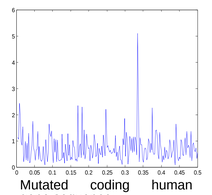
Simulation Results

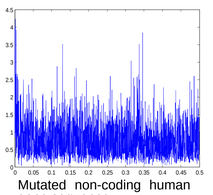

Coding gene sequence of C-elegans: *C_elegans_NM001026217_GI86561101_sequence*


Coding gene sequence of yeast: *Saccharomyces_cerevisiae_X57160_sequence*


Non-coding human gene sequence: *HS_AF527577_22094826_sequence*


Mutated coding human gene sequence: *HC_X17403_GI59591_sequence*


Mutated coding human gene sequence: *HS_NM_001123225_sequence*


Mutated non-coding human gene sequence: *HR_U49742_GI1236136_sequence*

Conclusions

- Protein-coding regions of DNA have been found to have a peak at frequency peak at frequency 1/3 in their Fourier spectra. This is called the period-3 property.
- The period-3 property might be related to the different statistical distributions of codons between protein coding and non-coding DNA sections.
- The period-3 property can be used as a basis for identifying the coding and non-coding regions in a DNA sequence.

References

[1] Bouaynaya N., Schonfeld D., *Non-Stationary Analysis of DNA Sequences*, *in IEEE Statistical Signal Processing Workshop*, pp. 200-204, August 2007.
[2] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, .Long-range correlations in nucleotide sequences,. *Nature, vol. 356, no. 6365, pp.* 168.170, March 1992.

Acknowledgment

We would like to acknowledge Ms. Nidhal Bouaynaya, Ph. D., an Assistant Professor at the University of Arkansas at Little Rock, for her guidance and support on this project. Also, we acknowledge Mr. Jerzy Zielinski for providing us with real genomic data. Mr. Zielinski is a Ph.D. candidate in the Department of Systems Engineering at UALR.

UALR University of Arkansas at Little Rock