Se2Si: A Software Tool for Biological Sequence Analysis

S. S. Vinod Chandra[1] & S. Acuthsankar Nair[2]

[1]Department of Computer Science and Engineering, College of Engineering, Thriuvananthapuram, India, *Email: vinodchandrass@gmail.com*

Centre for Bioinformatics, University of Kerala, India, Thriruvananthapuram, *Email: sankar.achth@gmail.com*

ABSTRACT

This paper introduces ideas for converting biological sequences like DNA, RNA or Protiens to numerical sequences. No single method may be adequate enough to predict all the genes from any genomes. So the recent trends in gene finding are to try out various types of combinational methods for gene finding. This paper combines several methods for converting biological sequences to numerical signals and by there bringing out various hidden information in them. Several mathematical operations to explore more information on the biological sequences are proposed. These numerical signals can be analyzed using various DSP (Digital Signal Processing) techniques. A software tool named Se2Si is the outcome of our work and is available at http://www.sooryakiran.com/products_ se2si.html

INTRODUCTION

Human Genome Project – was initiated as joint effort of U.S. Department of Energy and the National Institute of Health in 1990 to get complete, error-free and fully annotated sequence of entire human genome. According to the draft of human genome, the total number of genes is estimated at around *25,000* to *30,000* (1).The functions are unknown for over 50% of discovered genes. This refers to the problem of *finding exons* which are the templates of making proteins. This problem is a pivotal issue in gene finding. The problem is complex as no hard and fast rule has been has been discovered so far. If the functionality of the entire genome can be found out, it will be one of the milestones in the medical field.

Genomic Signal Processing is defined as the application of the DSP on the genomic data. It is a discipline that studies the processing and analysis of genomic signals (2, 3, 4). Its application includes intron - exon identification, searching for similarities between two sequences. Two important phases of GSP are the mapping from nucleotide symbols to numbers and the processing of mapped numerical sequences such as similarity testing, feature extraction. There are several mapping techniques for the conversion of DNA sequences to signals. But a single method may not bring out the entire hidden information. Our work tries to create a platform to combine several mapping techniques and explore the information thereof. Mapping techniques can extended to RNA and Protien sequences. Analysis of these sequences can be done for obtaining statistical information on a sequence like amino acid composition, searching similarity between two sequences, predicting and analyzing the secondary structure based on the sequence, predicting and analyzing tertiary structure and folding for RNA and protein sequences. Two significant tools in the gene finding area are GeneMark (5) and GeneScan(6). GeneScan uses Indicator function to convert sequences to signals and fourier anlaysis is done to distinguish between coding and non-coding regions. GeneMark uses specific parameters of the Markov models to determine the protein coding regions of a DNA.

Biological sequences are usually represented as character data. In order to study more about them, analysis of character data is very difficult. So it is better to convert these character data to signals, so that analysis becomes quite easier. At present, there is no single tool to generate signals as part of GSP. We develop a tool named Se2Si (Sequence to Signal Converter) in Java, which generate signals, by incorporating various methods discussed in the coming section.

MATERIALS AND METHODS

Mapping can be a crucial choice since it can hide or reveal the information of the given sequence. Two types of mapping are discussed here. They are Distance based mapping and Parameter based mapping techniques.

Distance based Mapping

Distance mapping involves mapping techniques based on the distance or position of the nucleotides in the biological sequence. Some of the distance mapping techniques are Indicator Sequence, Inter-Nucleotide Distance, Cumulative Categorical Periodogram, Position Count Function, BiNucleotide Distance etc.

Binary Indicator Sequence

Binary Indicator sequence (7) is a method of mapping. We will get four numerical sequences which are also called indicator sequences for a given DNA sequence. For a protein sequence there will be twenty such indicator sequences. Indicator Sequences contains numbers 0 or 1 to indicate the absence or presence of nucleotide in the original sequence. The four sequences represent the frequency content of each nucleotide.

Consider a DNA sequence *S* of length N, S = S(1), S(2), S(3), …S(*n*), indicator sequences i_A, i_G, i_C, i_T are defined as

i_A(*n*)=1 if S(*n*) ='A' else 0, *n* = 1 to N

i_G(*n*)=1 if S(*n*) ='G' else 0, *n* = 1 to N

i_C(*n*)=1 if S(*n*) ='C' else 0, *n* = 1 to N

i_T(*n*)=1 if S(*n*) ='T' else 0, *n* = 1 to N

Consider a sequence AGCTAGTTTC. The indicator sequences for each nucleotide can be given as

i_A=1000100000

i_G=0100010000

i_C=0010000001

i_T=0001001110

Inter-Nucleotide Distance

This mapping technique is based on the distance between the identical nucleotide symbols. Here a numerical sequence for a particular biological sequence is obtained by replacing the nucleotide by the distance between the next similar nucleotide. Consider a DNA sequence *S* of length *N*, *S* = *s*(1), *s*(2), *s*(3), …., *s*(*n*), Inter-Nucleotide distance sequence function *inter* is defined as

inter(*n*) = *k*, where *k* = min value of i such that *S*(*n*)=*S*(*n*+*i*), *n*+*i*<=N

else *k* = *N*–*n*.

The numerical sequence obtained by applying Inter-Nucleotide function on a sequence A G C T A T is

inter = 5 4 3 2 1 0

Binucleotide Distance

This mapping technique is based on the distance between the complementary nucleotide distance, That is distance between AT, CG. Consider a DNA sequence *S* of length *N*, *S* = *S*(1), *S*(2), *S*(3), …*S*(*n*). The Binuleotide distance function *bin* can be defined as

bin(*n*) = *k*, where *k* = min value of *i* such that

if *s*(*n*) = *A*, then *s* (*n* + *i*) = *T*,

if *s*(*n*) = *T*, then *s*(*n* + *i*) = *A*,

if *s*(*n*) = *C*, then *s*(*n* + *i*) = *G*,

if *s*(*n*) = *G*, then *s*(*n* + *i*) = *C*, *n* + *i* < = N,

else *k* = *N* – *n*

For the sequence AGCTA, the binucleotide sequence obtained is

bin = 3 1 2 1 0

Categorical Cumulative Periodogram

CCP (8) is computed by determining the number of cycles with each of the possible periods (1 to *n*-1). CCP calculates *number of cycles in a period*. A cycle occurs in a sequence when the same nucleotide turns up again. A period of the cycle is taken as the number of nucleotides in between + 1. Consider sequence ACAAACC

CCP(1) = No. of occurrences of cycle with period 1

= No. of occurrences of cycle with zero intervening event

= 3

CCP( ) measures the existence of pairs of identical elements at a distance *i* base pairs.

Parameter Mapping

Parameter based mapping involves mapping techniques based on the biochemical values of the nucleotide symbols present in the given sequence. The UMBC AAIndex Database (http://www.evolvingcod e.net:8080/aaindex) is database of aminoacid which contains biochemical parameters. These parameters have some predefined values. Some of the parameters are EIIP, Hydrophobicity etc. Electron Ion Interaction Pseudo Potential (EIIP) is the estimation of the Energy of delocalized electrons in nucleotides. When we substitute the EIIP values for the biological sequence, it represents the distribution of free electron energies along the sequence (9). Hydrophobicity is the property of nucleotide that is repelled from the water. Further operations on the parameter mapped sequence can be performed in order to bring out the hidden properties that may not be exhibited during the first mapping.

We can perform any wild ***mathematical operations*** on the parameter mapped sequence (9). If *x* is the mapped sequence then the new sequence *y* can be obtained by performing the operations like Pow(*x*,***n***), Sqrt(*x*,*n*), Exp(*nx*), Sin(*nx*), Log(*nx*) …or by combination of these methods. That is we can form an expression in *x* which can be linear or quadratic etc. If the parameter values are entered as complex number *Z* (*a* + *ib*), then the various ***complex operations*** that can be performed on it are,

Magnitude	$\sqrt{a^2+b^2}$
Angle	$\tan^{-1} x/y$
Norm	$a^2 + b^2$
$Z.Z^*$	$\|Z\|^2$
$Z+Z^*$	2Real (Z)

The other mathematical techniques that is applied on the sequence are Clustering, Filtering and Autocorrelation

Clustering

Clustering algorithms (10) are used to find groups of "similar" data points among the input patterns. *K* means clustering is an effective algorithm to extract a given number of clusters of patterns from a set. *K* can be defined as the cluster constant, which should be less than the length of the nucleotides in a particular sequence.

Consider a sequence having parameter values as random numbers.

A	R	N	D	E
1	20	10	5	3

Let $k = 3 \rightarrow$ there will be 3 clusters and number of centroids will be 3

Let the centroids	=	1, 20, 10
Cluster 1, centroid	=	1
Elements of cluster1	→	A, D, E
New centroid	=	(1 + 5 + 3)/3
	=	4.5
Cluster 2, centroid	=	20
Elements	→	R
Cluster 3, centroid	=	10
Elements	→	N

Final parameter values of the sequence after clustering will be as follows.

A	R	N	D	E
4.5	20	10	4.5	4.5

Filtering

Filtering techniques (11) are applied on the mapped sequences inorder to remove unwanted parts of a signal. That is to extract the useful parts. Some of the filtering techniques used are

1. Simple Gain Filter

   $Yn = k*Xn$ for $n = 0, 1, 2\ldots$

where Yn = Filter Output

Xn = Filter Input

k = Filtering Constant

$k > 1 \rightarrow$ Filter acts as amplifier

$0 < k < 1 \rightarrow$ Filter acts as attenuator

$k < 0 \rightarrow$ Filter acts as inverting amplifier

2. Two-term Difference Filter

   $Yn = Xn - Xn-1$, $n = 0, 1, 2\ldots$

3. Central Difference Filter

   $Yn = (Xn - Xn-2)/2$, $n = 0, 1, 2\ldots$

Table 1 shows the filtering values for the above filtering types for a protein sequence after mapping with EIIP parameter.

Autocorrelation

The correlation function shows how similar two signals are, and for how long they remain similar when one is shifted with respect to the other. Correlating a signal with itself is called autocorrelation. For performing autocorrelation as a mapping technique, the shift constant has to be calculated that can be a value between $-(N-1)$ to $+(N-1)$, where n is the length of the entered sequence. If the shifting constant is a positive value, we have to shift the values to right else the appropriate values to the left.

Consider A, G, C, T has been mapped to some parameter values 1, 2, 3, 4. Then the autocorrelation function with positive shifts is as follows:

1 2 3 4

1 2 3 4

A(0) = 1*1 + 2*2 + 3*3 + 4*4 = 30

1 2 3 4

1 2 3 4

A(1) = 1*0 + 2*1 + 3*2 + 4*3 = 20

1 2 3 4

1 2 3 4

A(2) = 1*0 + 2*0 + 3*1 + 4*2= 11

1 2 3 4

1 2 3 4

A(3) = 1*0 + 2*0 + 3*0 + 4*1= 4

The final correlated signal obtained for the sequence is as follows.

Sequence	A	G	C	T
Signal	30	20	11	4

Table 1
Filtering

EIIP	*Simple gain filter (k = 10)*	*Two term Difference filter*	*Central Difference filter*
0.2953	2.953	0.2953	0.2953
0.7593	7.593	0.4640	0.4640
0.0285	0.285	–0.7308	–0.7308
1.0000	10.00	0.9715	0.9715
0.6563	6.563	–0.3437	–0.3437
0.6025	6.025	–0.0538	–0.0530
0.0459	0.459	–0.5564	0.5566
0.0396	0.396	–0.0163	–0.0063

RESULTS

The ideas proposed in this paper are implemented in a tool named Se2Si. With this tool the user can perform any wild operations on the signals to bring out the hidden information. With the help of this tool user can add new parameters and there is also provision for viewing and editing these parameter values. User can try various mapping techniques on any biological sequences. For every distance functions and mapping techniques, the result obtained is numerical sequences. Using these results, a researcher can apply Digital Signal Processing technique like *Fourier transform* to extract the information contents. The fourier spectra of a DNA sequence reveal a peak where the composition of nucleotides represents a particular biological function. Let $x(n)$ be the numerical sequence that is obtained as the output for a particular biological sequence of length N. $X(k)$ be the fourier transform and it is evaluated as

$$X_e[k] = \sum_{n=0}^{N-1} x_e[n] e^{(-j2\pi kn/N)}, k = 0,1,2,-N-1$$

Corresponding value of power spectrum is, $S(k)=|X(k)|^2$

When $S(k)$ is plotted against k, it reveals a peak at $N/3$ for a coding region and no such peak is observable for a non coding region (Fig 1). It has been experimentally proven that the 3-base periodicity which leads to $N/3$ peaks in coding regions are due to non-uniform distributions of nucleotides in the three positions of codons and there is a definite correlation between these nucleotide distributions and the $N/3$ peak (12). This technique holds for all parameter mapped sequences like EIIP indicator sequences and for distance functions like Inter-Nucleotide sequences and Binucleotide sequences. For Binary Indicator sequences $S(k)$ is calculated as $|X_A(k)|^2 + |X_G(k)|^2 + |X_C(k)|^2 + |X_T(k)|^2$, since the output will be four numerical sequences one for each nucleotide. In the case of CCP indicator sequences, when $S(k)$ is plotted against k,

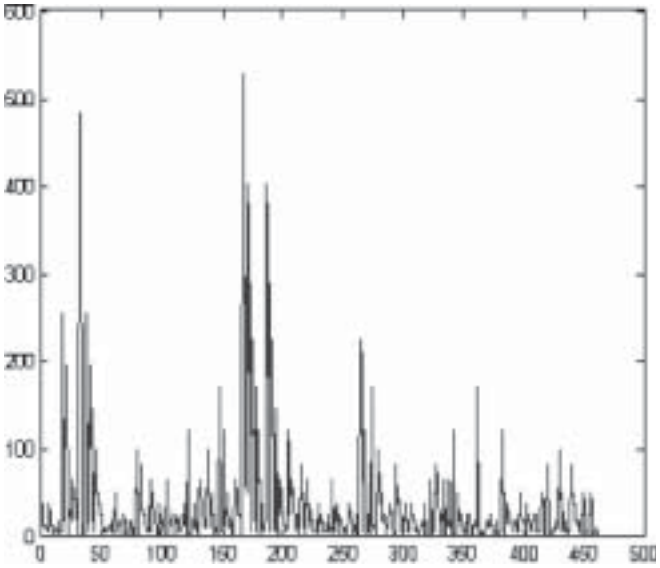

Figure 1: FFT of Binucleotide Mapping of a Human Gene with Base Pair N = 465, Demonstrated Peak at k = 155(N/3)

troughs are obtained at $N/3$ positions in the case of coding regions and such troughs are absent in non coding regions (Fig. 2). CCP method and Parameter mapping methods, reduces the computational overhead by 75%, when compared to the Binary Indicator method.

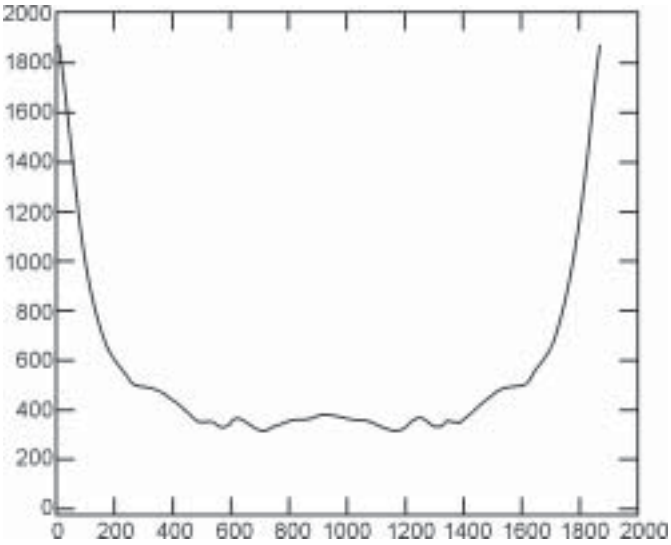

Figure 2: FFT of CCP Mapping of Humelafin with Base Pair 1878, Demostrated Trough at 626

ACKNOWLEDGEMENT

We thank Vandana V. Rajan, Department of Computer Applications, College of Engineering Trivandrum for the help and support. We also thank SooryaKiran Bioinformatics (P) Ltd., Industry Incubation Centre, University of Kerala, Thiruvananthapuram – 695 581 for the support provided.

REFERENCES

[1] http://www.ornl.gov/sci/techresources/Human_Genome/pu blicat/publications.shtml#gtl

[2 P. P. Vaidyanathan and B. J. Yoon (2004) The Role of Signal Processing Concepts Genomics and Proteomics. *Journal of the Franklin Institute*, Special issue on Genomics.

[3 P. D. Cristea (2003) Large Scale Features in DNA Genomic Signals. *Signal Processing*. **83(4)**. 871–888.

[4] D. Anastassiou (2000) Frequency Domain Analysis of Biomolecular Sequences. *Bioinformatics* 1073–81.

[5] Borodovsky M. andMcIninch J. (1993) GeneMark: Parallel Gene Recognition for Both DNA Strands. *Computers & Chemistry*, **17(19)**, 123–133.

[6] Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S. and Ramaswami (1997) Prediction of Probable Genes by Fourier Analysis of Genomic Sequences. *Comput Appl Biosc.* **13**, 263–270.

[7] D. Anastassiou (2001) Genomic Signal Processing. *IEEE Signal Processing Magazine.* **18(4)**, 8–20.

[8] Achuthsankar S. Nair & T. Mahalakshmi (2006) Are Cateogorical Periodograms and Indicator Sequences of Genomes Spectrally Equivalent? *In-Silico Biology*. **6(3)**, 215–222.

[9] Achuthsankar S. Nair and Sreenadhan. S, (2006) A Coding Measure Based on Electron-Ion Interaction Pseudopotential (EIIP) for Locating Exons Through Digital Filtering of DNA Sequences. *Bioinformation*, **1(6)**, 197–202.

[10] http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html

[11] Achuthsankar S. Nair and Sreenadhan.S, (2006) An Improved Digital Filtering Technique using Nucleotide Frequency Indicators for Locating Exons, *Journal of the CSI*, **36(1)**, 60–66.

[12] http://www.cse.yorku.ca/~datta/bioinformatics.html