Visualization Of Genomic Data Using Inter-Nucleotide Distance Signals

Achuth Sankar S. Nair[1], T.Mahalakshmi[2]
Department of Computer Science, University of Kerala, India
Department of Computer Science, University of Kerala, India

I. ABSTRACT

Recent interest in Genomic Signal Processing has centered around studies on digital indicator sequences corresponding to the alphabet set {A, G, C, T}. Applications of Fourier transforms to such signals have yielded exon characterization peaks. Spectral analysis has also been used in finding genes, determining correct reading frames and in studying active regions in protein sequences. The indicator sequences are information conserving signal representations of genomic data. However it cannot be assumed that they are the best forms of digital signal representations of genomic data, as far as visualizing hidden characteristics of the genes are concerned. This paper presents visualization of genomic data using *the inter-nucleotide distance sequence*. It is shown that these signals can highlight hotspots in genomic sequences through the application of Fourier transforms.

KEYWORDS: Genomic Signal Processing, Genomic data visualization, inter-nucleotide distance signals.

II. INTRODUCTION

Genomic signal processing (GSP) is a relatively new area in bio-informatics, which deals with digital signal representations of genomic data and analysis of the same using conventional digital signal processing (DSP) techniques. Major works reported in genomic signal processing [1, 9, 10, 11] have generated considerable interest in this area. Most of the work in GSP is centered on indicator sequence of digital representation of genomic data.

Given a genomic 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 follows:

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, and
i_T(n) =1 if s(n)= 'T' else 0, n=1 to N.

The indicator sequences are binary sequences. The indicator sequences for the short DNA fragment AGTTCTACCGAGC is given below.

i_A = 1000001000100
i_G = 0100000001010
i_C = 0000100110001
i_T = 0011010000000

An example of indicator sequences of gene S67057.1 of Cricetulus migratorius (Armenian hamster) of length 608 are shown in Fig.1.

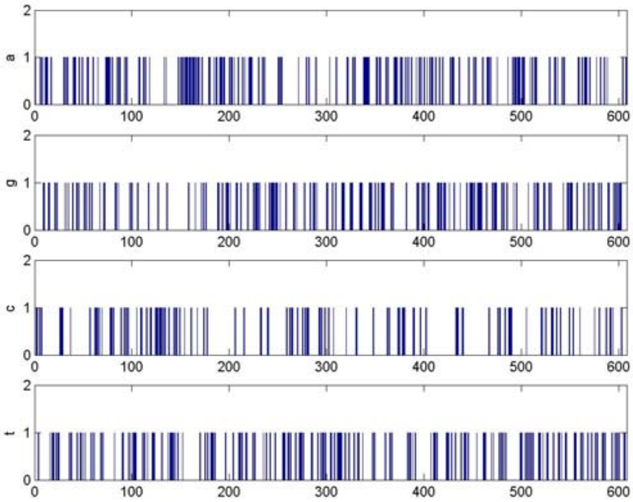

Fig. 1 Indicator sequences of gene S67057 of Cricetulus migratorius (Armenian hamster) of length 608

These representations are obviously information conserving, as it is a trivial process to reconstruct the genomic sequence exactly from the four indicator sequences. It is also obvious that one of the sequences is redundant as the alphabets are constrained to the set {A, G, C, T} and hence it is possible to reconstruct the genomic sequence exactly from any three-indicator sequences. This dimensionality reduction can be effected by redefining the indicator sequences as follows [1]:

i_X(n)=(√2/3){2i_T(n)–i_C(n)– i_G(n)}

i_Y(n) = (√6/3){i_C(n) – i_G(n)}

i_Z(n)=(1/3){3i_A(n)–i_T(n)–i_C(n)– i_G(n)}

It has been shown that a simple Discrete Fourier Transform (DFT) of the DNA sequence represented as indicator sequences can show peaks at N/3, if the sequence codes for proteins [6, 12, 13]. For example DFT of the coding region of gene f56f11.4a of length 1236 base pairs (Caenorhabditis elegans) demonstrated a peak at frequency k =408 (Fig. 2).

[1] Honorary Director, Centre for Bioinformatics, University of Kerala, Thiruvananthapuram, India 695581,Achuthnair@hotmail.com
[2] Asst. Prof., Sree Narayana Institute of Technology, Kollam, India 691010 and
Systems Manager (Honorary), National Institute of Computer Technology, Kollam, India 691001, mlakshmi@sancaharnet.in

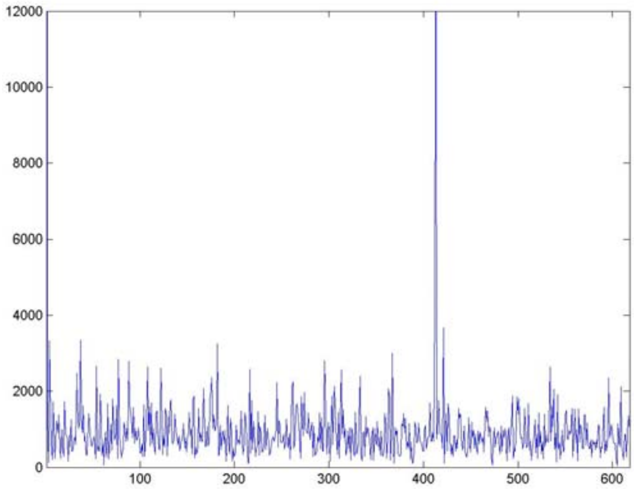

Fig. 2 DFT of the coding region of f56f11.4a (Caenorhabditis elegans) demonstrated a peak at frequency k =408

Other techniques used in gene prediction include Hidden Markov Modules (HMM) [17, 18] which are probabilistic models. Just like in gene prediction DSP plays an important role in protein sequence analysis. A protein molecule is a long sequence of amino acids connected together by a covalent peptide bond [15]. Most of the biological process in living organisms is driven by proteins. A protein molecule folds into a 3D shape determined entirely by the amino acid contained in it which enables it to interact selectively with a small number of other molecules. The term hotspot is used to indicate the sites at which other molecules can attach to a protein molecule and there are many such hot spots. Spectral analysis has also been used in studying these hotspots [1, 16] as well as in determining correct reading frames [1, 19, 20].

III. VARIATIONS IN GENOMIC SIGNAL REPRESENTATIONS

Literature survey reveals that even though the indicator sequence representation of genomic data is by far the most popular digital signal representation, there have been variations reported [1, 2, 3]. One prominent example is that of complex representation [5]. This representation is arguably biologically significant as complex conjugates are used to represent the pairing bases (A with T and G with C). Such representations translate some of the attributes of the bases into mathematical properties. Fig. 3 indicates the result of absolute DFT of gene f56f11.4a (Caenorhabditis elegans) with base pairs 1236 and using complex sequence representation with A = 1 + i, C = -1 - i, G = -1 + i, T= 1 - i.

Another variation consists of mapping nucleotide symbols to digits 0, 1, 2 and 3. According to Cristea [4] the selection of these particular digits are claimed as optimal choice resulting from the condition of minimally non-monotonous correspondence between the codons and the amino acids that lead to best auto-correlated extra-genic gentic signals. Criesta also observes that genetic signals built from genes show low auto-correlation, even for neighboring samples, a feature usually associated with noise. This is considered as consistent with the fact that the functionality of a protein is not fully determined by its primary structure. He further observes that the extra-genic genetic signals obtained from non-coding DNA sequences have good correlation with close neighbors, that decreases abruptly with distance, features that are typical for piecewise smooth natural signals.

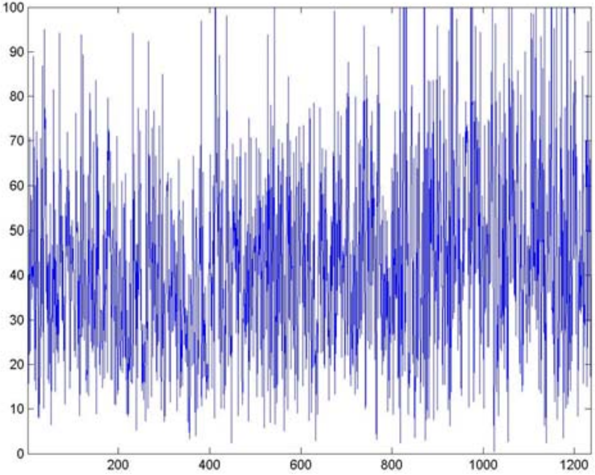

Fig. 3. Absolute DFT of f56f11.4a (Caenorhabditis elegans) *with base pairs 1236 and using complex* sequence representation with A = 1 + i, C = -1 - i, G = -1 + i, T = 1 - i.

IV. INTER-NUCLEOTIDE DISTANCE REPRESENTATION

It was mentioned earlier that indicator sequences are information conserving. With every representation of genomic signals, this property is crucial, but cannot be taken as the sole criterion for selection of representation. The motivation of the current work is that as genomic signal processing is a nascent field, exploratory studies are required to temper its foundations. Hence this paper explores one of the alternate signal representations of genomic data. The authors have employed *inter-nucleotide distance* sequences for further applications of digital signal processing.

The inter-nucleotide distance sequence, herein after referred to as *in,* is defined as follows:

Given a DNA sequence S = s(1), s(2), s(3), ……s(N),
in(n) = k, *where k* = min value of *i* such that
s(i)=s(n+i), n+i<=N else k = N-n.

An example of inter-nucleotide distance sequences for the short DNA fragment AGTTCTACCGAGC is given in the Fig. 4.

DNA sequence	A	G	T	T	C	T	A	C	C	A	G	C
In sequence	6	9	1	2	3	6	3	1	3	2	1	0

Fig. 4 Inter-nucleotide distance sequence for the short DNA fragment AGTTCTACCGAGC

This is essentially replacing every base symbol by a number k which is the base distance between the next similar base. In case such a similar base is not found then the *in* sequence value of that base is the length of the remaining sequence. This is obviously not an information conserving representation, as it would be impossible to reproduce the DNA sequence from the representation. However, this representation and variations of the same can be argued to have biological significance in certain cases. If the *in* sequence is used to represent peptide chain, these distance measures would have direct implications on the folding points of the protein. In case of RNA sequences, if distance measures are taken between bonding bases, it would again be strong indicators of intraRNA structures like loops and bulges.

Fourier and wavelet transforms are two important tools of Digital Signal Processing and both have been applied effectively in Genomic Signal Processing [6, 7, 8]. An impulse in the Fourier transform at zero frequency represents a constant component in the time domain. The value of a Fourier transform at a given frequency depends on the time domain signal at all values of time. In the present work, DFT was chosen to analyze the *in* sequences. DFT was applied to *in* sequence representation of a set of 17 genes. The method used to identify discriminatory signals was to over-plot the frequency spectrum using a window of certain size. A window of size 18 was selected after experimenting with different sizes, since it was found to be most appropriate for the present case. Fig. 5 and Fig. 6 given below indicates such discriminatory signals corresponding to two genes.

Promoter regions are genuine sequence sites which are key to expression of genes. They serve the biological purpose of enabling binding prior to transcription. They are usually identified by detecting certain conserved patterns in the sequence upstream of gene coding regions [14]. For example E.Coli promoters are known to contain Pribnow box, a region which has the consensus sequence TATAAT at position –10 and another region with consensus TTGACA at –35.

Analysis of the experiment with 17 genes revealed the existence of discriminatory spectral envelope. It was observed in the case of 10 genes (Table I) in and around the promoter region. In the remaining cases (Table II) such envelope were found in the coding region. These two sets are given in Table I and II respectively. Appendix A gives description of genes used for experiment. Genes were chosen from the dataset constructed by Burset and Guigo and the discriminatory spectrum was classified based on the predicted coding regions given in the set.

V. Conclusion

It is concluded that the inter-nucleotide signals are a novel way of digital signal representation of genomic data which is seen to have a discriminatory capability in highlighting the promoter region of gene sequences. It is seen from the results of applying DFT to *in* sequence signals that they can discriminate promoter regions in a majority of cases of genes experimented with. While more broadband trials are required to confirm the applicability of this method and to fine-tune them, initial explorations are encouraging. The pattern match [20, 21, 22] for identifying promoter regions seems to be superior to the *in* sequence spectral analysis at the moment. The authors are investigating improvements over the method and also alternate signal representations.

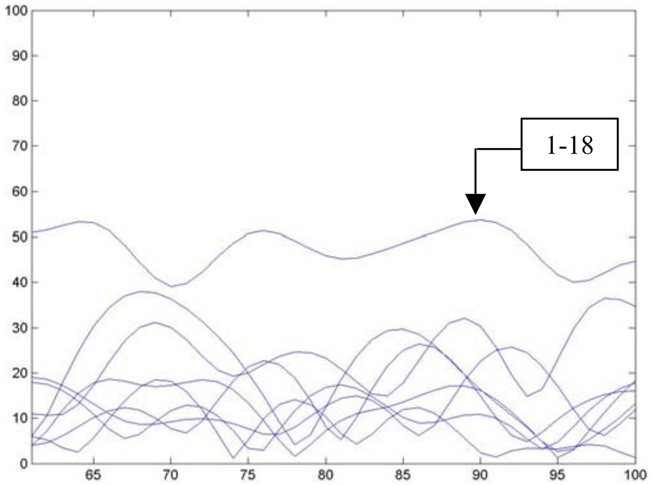

Fig 5 Over plot of *DFT of U00432 (Cyprinus carpio – common crap) with base pairs 1583 and using "in-sequence" representation. A discriminatory spectral envelope was found in the region 1-18*

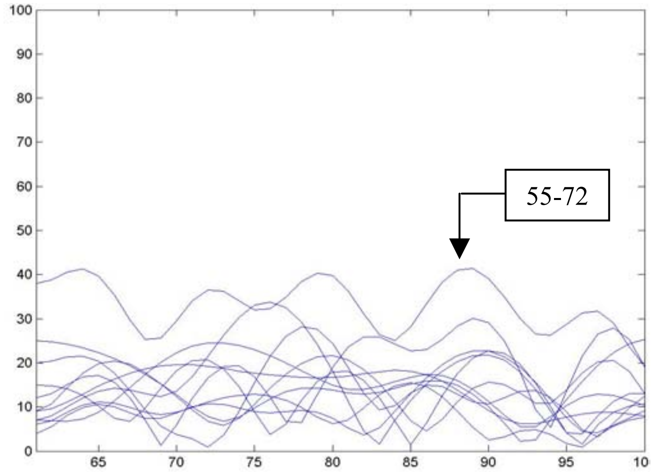

Fig. 6 Over plot of DFT of S67057 (Cricetulus migratorius - Armenian hamsters) with base pairs 1754 and using "in -sequence" representation. A discriminatory spectral envelope was found in the region 55-72.

Name	**BP**	**Coding Area**	**Discriminatory Spectral Envelope**
ECGLOAP1	1141	268-877	37-54, 199-216
U00432	1583	1-1523	1-18
S67057	1754	776-1546	55-72
BOVANPA	1769	313-1394	1-18, 199-216
MACHBGA1	2105	403-1850	37-54
PIGAPA1	3333	751-2370	181-198
ACU08131	5392	521-4247	379-396
AGGGLINE	7360	3066-4521	1099-1116
CALEGLOBIN	1698	144-1527	55-72, 1153-1180
TARHBG	2147	173-1887	1-18, 2035-2042

Table I. Discriminatory spectral envelope observed in the promoter area when DFT was applied to in sequence representation of set of genes

Name	BP	Coding Area	Discriminatory Spectral Envelope
DMPROTP1	624	122-425	181-198
GGPROP2	648	36-505	91-108, 325- 342
BOVGAS	1066	540-999	757-774
RABCRP	1438	333-1262	487-504,811-828
FDTNFA	1722	1-1722	37-54, 433-450
LAYEGLOBIN	1702	146-1531	937-954, 1171-1189
OAMT11	2055	995-1788	559-576, 1495-1512

Table II. Discriminatory spectral envelope observed in the coding area when DFT was applied to in sequence representation of set of genes.

REFERENCES

[1] D. Anastassiou, "Genomic Signal Processing", IEEE Signal Processing Magazine, Vol. 18, no 4, pp. 8-20, July 2001.

[2] Elena Pirogova P., Fang Q., Akay M., Irena Cosic Z., "Investegations of the structural and functional relationships of Oncogene Proteins", Proceedings of IEEE, Vol. 90, no 12, Dec 2002.

[3] P.P. Vaidyanathan and B. J. Yoon, "Gene and exon prediction using allpass-based filters", in Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC, Oct 2002.

[4] Cristea P.D., "Genetic Signals: An emerging concept", Proceedings of IWSSIP 2001.

[5] Cristea P.D., "Conversion of nucleotides sequences into genomic signals", J. Cell. Mol. Moed., Vol. 6, no 2, pp. 279-303, 2002.

[6] S. Tiwari, S. Ramachandran, A. Bhattacharya., S. Bhattacharya and R. Ramaswamy, " Prediction of probable genes by Fourier analysis of genomic sequences", CABIOS, Vol. 113, pp. 263-270, 1997.

[7] E. Coward, "Equivalence of two Fourier methods for biological sequences", Jour. of Math. Bio., Vol. 36, pp. 64-70, 1997.

[8] Issac B., Singh H., Kaur H., Raghava G.P.S., "Locating probable genes using Fourier Transform approach", Bioinformatics, Vol. 18, no 1, pp. 196-197, 2002.

[9] D. Anastassiou " Digital Signal Processing", Technical report, Dept of EE, Columbia University, 2002.

[10] Mitra S.K., "Digital Signal Processing: A Computer Based Approach", McGraw-Hill, 1998.

[11] J.W. Fickett and C.S. Tung , "Assessment of Protein coding measures", Nucleic Acids Res, 20, 6441-6450, 1992.

[12] J.W. Fickett, "Recognition of Protein Coding Regions in DNA sequences", Nucleic Acids Res, vol 10, pp. 5303-5318, 1982.

[13] V.R. Chechetkin and A.Y. Turygin, "Size-dependence of three-periodicity and long-range correlations in DNA sequences," Phys. Lett. A, vol. 199, pp. 75-80, 1995.

[14] Mount D.W., "Bioinformatics: Sequence and Genome Analysis", Chapter 8, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2001.

[15] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, "Essential Cell Biology", Garland Publishing Inc., New York, 1998.

[16] P.P. Vaidyanathan and B.J. Yoon, "The role of signal processing concepts genomics and proteomics", Journal of the Franklin Institute, Special issue on Genomics, 2004.

[17] A. Korosh, I. Saira Mian and D. Haussler, "A hidden Morkov Model that finds genes in E.Coli DNA", Nucleic Acids Research, Vol. 22 pp. 4768-4778, 1994.

[18] S.L. Salzberb, A.L. Dekher, S. Kasif and O. White "Microbial gene identification using interpolated Markov models", Nucleic Acids Research, Vol. 26, no. 2, pp. 544-548, 1998.

[19] D. Anastassiov, "Frequency-domain analysis of bimolecular sequences", Bioinformatics, vol. 16, no. 12, pp. 1073-1082, Dec. 2000.

[20] Acuts http://pbil.univ-lyon1.fr

[21] Alibaba2 http://www.gene-regulation.de

[22] Epd http://www.epd.isb-sib.ch

Appendix A – Description of the Genes used for Samples

Name	Organism
ACU08131	Anolis carolinensis (green anole)
AGGGLINE	Ateles geoffroyi (black-handed spider monkey)
BOVANPA	Bos taurus (cow)
BOVGAS	Bos taurus (cow)
CALEGLOBIN	Callithrix jacchus (white-tufted-ear marmoset)
DMPROTP1	Didelphis marsupialis (southern opossum)
ECGLOAP1	Equus caballus (horse)
FDTNFA	Felis catus (cat)
GGPROP2	Gorilla gorilla (gorilla)
LAYEGLOBIN	Lagothrix lagotricha (common woolly monkey)
MACHBGA1	Macaca mulatta (rhesus monkey)
OAMTII	Ovis aries (sheep)
PIGAPA1	Paroxysmal nocturnal hemoglobinuria, included; pnh, included phosphatidylinositol glycan, class a, pseudogene 1
RABCRP	Oryctolagus cuniculus (rabbit)
S67057	Cricetulus migratorius (Armenian hamsters)
TARHBG	Tarsius syrichta (tarsier)
U00432	Cyprinus carpio (common carp)