2-Simplex Mapping for Identifying the Protein Coding Regions in DNA

Durga Ganesh Grandhi and C. Vijay Kumar
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India

***Abstract*— Genomic Signal Processing is an emerging interdisciplinary area. The problem of Identifying Protein Coding Regions in DNA is addressed using signal processing techniques in this paper. DNA can be thought as a string formed from the alphabet set $\mathcal{A} = \{A, C, G, T\}$. It is found that, in protein coding regions the symbols have periodicity of 3 [1], which can be used as a cue to identify the protein coding regions using signal processing techniques. This is possible only if the symbol sequences are mapped to numbers. In this paper a new lower dimensional mapping is proposed which reduces the computational complexity by half, producing results nearly equal to those produced by a higher dimensional mapping.**

I. Introduction

Deoxyribo Nucleic Acid (DNA) is a double stranded structure made up of four nucleotides (bases) - Adenine, Cytosine, Guanine and Thymine, which are denoted with the letters A,C,G and T respectively, which is shown in Fig-1 [2].

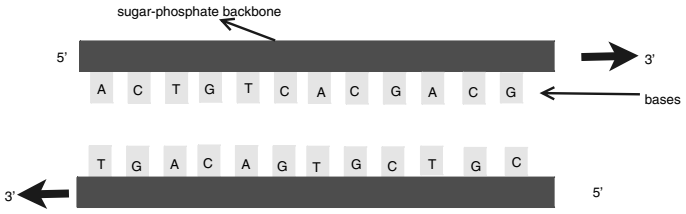

Fig. 1. The two strands of the DNA

As shown in Fig-2[2], the eukaryotic DNA is divided into genes and inter-genic spaces. Genes are further divided into exons and introns. The bases in the exons are assumed to be divided in groups of three consecutive bases, called codon. Each codon will code an amino acid. The sequence of these amino acids is called protein. So exons are called protein coding regions. It is found that the bases in the protein coding regions have 3 periodicity. In genomic signal processing area this property is exploited to identify the exons using signal processing methods.

The identification of protein coding regions involves two steps:

- Mapping symbols to numbers
- Identification of period-3 regions

A set of symbols is not a *field* [3] because algebraic operations on symbols are usually not meaningful. For example, addition, multiplication and numeric ordering cannot be performed on symbols. In order to apply signal processing methods to this problem, symbol sequence (DNA sequence) need to be mapped to number sequence. The numbers need to have the properties of a field. Mapping should be chosen properly because it inherits mathematical structure of a field, which is originally not present in the DNA sequence. For example, consider a mapping: $A \leftrightarrow 1$, $C \leftrightarrow 2$, $G \leftrightarrow 3$ and $T \leftrightarrow 4$. This mapping would suggest that one nucleotide (base) is some how greater than another, a property that DNA sequence does not possess.

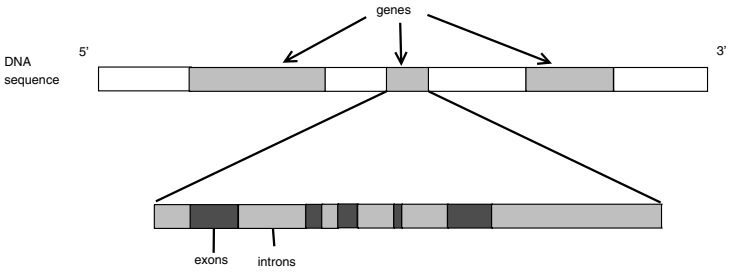

Fig. 2. Structure of Eukaryotic DNA

An *ideal mapping* should be such that the period-3 component of the DNA sequence should be independent of the mapping of the nucleotides, which is possible only through symmetric mapping [4]. Once the mapping is done, signal processing techniques can used to identify period-3 regions in the DNA sequence.

The average length of a chromosome is of the order of millions of bases. So it needs vast number of computations for identifying the protein coding regions. The computational complexity can be reduced either at mapping time or at implementation time. This paper proposes a new mapping technique called 2-simplex mapping, which reduces the computational complexity for identifying protein coding regions in the DNA.

Existing Mappings :

The mapping from symbol sequence to number sequence can be done in many ways. One of the most popularly used mapping is *Voss Mapping* [5]. In voss mapping, for each symbol α (A, C, G, T) an indicator sequence $x_\alpha(n)$ is defined as given below.

DNA Sequence:	T	T	G	T	C	A	C	T	C	G	G
$x_A(n)$:	0	0	0	0	0	1	0	0	0	0	0
$x_C(n)$:	0	0	0	0	1	0	1	0	1	0	0
$x_G(n)$:	0	0	1	0	0	0	0	0	0	1	1
$x_T(n)$:	1	1	0	1	0	0	0	1	0	0	0

The indicator sequences are made of two numbers $0, 1$. In the indicator sequence $x_\alpha(n)$, 1 indicates the presence of

1-4244-1272-2/07/$25.00 ©2007 IEEE

base α and zero indicates its absence. Voss mapping is a four dimensional mapping, because each base in the DNA sequence is represented by a four dimensional vector composed of either '0' or '1'. The number of 1's in any vector is exactly one. In [6], the authors used a complex number mapping as given below.

$$A = 1 + j,\ C = -1 - j,\ G = -1 + j,\ T = 1 - j. \quad (1)$$

In [7], the authors used a slightly variant mapping of Eq.(1) as given in equation Eq.(2).

$$A = +1,\ C = -j,\ G = -1,\ T = +j \quad (2)$$

In [4], tetrahedron mapping has been used. In which each nucleotide is assigned to one of the four corners of a regular tetrahedron. The vectors drawn from the center (which is also the origin of R^3) of the tetrahedron to the corners represents each of the four nucleotides as shown in Fig-3.

In [8], the author proposed an optimized 2-dimensional mapping technique. Even though it gives better results, it is dependent on the DNA sequence to be processed. One has to compute this mapping scheme before processing the DNA sequence for finding the protein coding regions.

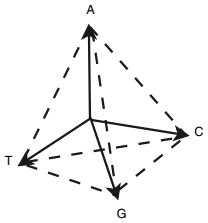

Fig. 3. Tetrahedron Mapping

The remaining paper is organized as follows. A new mapping function is presented in section II. Section III discusses, the filtering setup and results. Section IV provides concluding remarks.

II. 2-Simplex Mapping

Simplex is defined as a Euclidean geometric spatial element having the minimum number of boundary points, such as a line segment in one-dimensional space, a triangle in two-dimensional space, or a tetrahedron in three-dimensional space. 2-simplex mapping means triangle based mapping. The 2-simplex mapping is as shown in the Fig-4. Any of the three bases of A, C, G, T are assigned to the three vertices of an equilateral triangle (whose center at the origin) and the remaining one is assigned to the origin. The vectors from the origin to each vertex represent three nucleotides and the fourth one is '0'. For example one such 2-simplex mapping is $A = 0.25\hat{x} + 0.96\hat{y},\ C = 0,\ G = 0.7\hat{x} - 0.7\hat{y},\ T = -0.97\hat{x} - 0.26\hat{y}$ as shown in Fig-4.

Given the DNA sequence, each symbol is replaced with its associated vector. Now the DNA sequence is a sequence of 2-dimensional vectors. X_x, X_y are the two indicator sequences for the 2-simplex mapping, Where X_x and X_y represents the

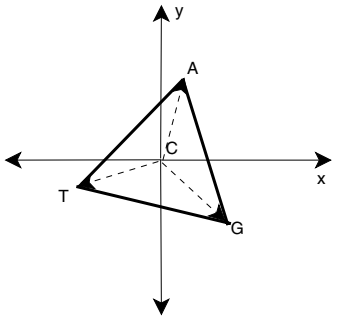

Fig. 4. 2-simplex Mapping

collection of x and y components in the above mentioned vector sequence. Both these indicator sequences are passed through an IIR antinotch filter [9], whose passband is located at $2\pi/3$ (period-3 corresponds to $2\pi/3$ component in frequency domain). The IIR antinotch filter used is derived from the allpass filter whose transfer function is

$$A(z) = \frac{R^2 - 2R\cos\theta z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \quad (3)$$

which has poles at $Re^{\pm j\theta}$ and zeros at $1/Re^{\pm j\theta}$. The allpass filter can be divided into two power complementary filters, one is notch filter and the other is antinotch filter. Out of which the antinotch filter can be obtained by

$$H(z) = \frac{1 - A(z)}{2} \quad (4)$$

After solving, the transfer function of Antinotch filter is

$$H(z) = \frac{(1 - R^2)}{2}\left[\frac{1 - z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}}\right] \quad (5)$$

where R value can be chosen near to 1, but should be less than 1 to make filter stable. The filter response for different values of R is shown in Fig-5.

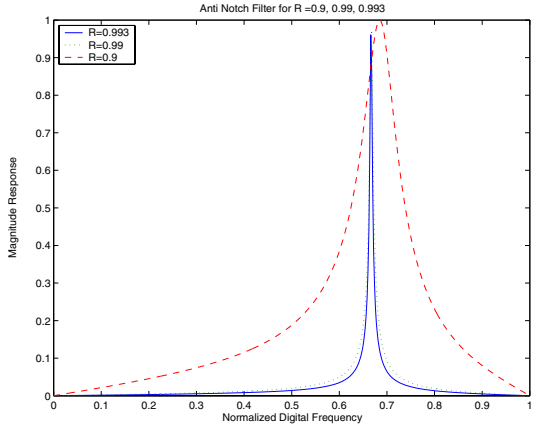

Fig. 5. The frequency response of IIR Antinotch filter for different locations of the poles is shown.

III. Filtering Setup

The block diagrams of filtering used for identifying the protein coding regions is as shown in Fig-7 and Fig-8 for voss and 2-simplex mappings respectively. It is better to realize the antinotch filters using lattice structure than direct forms, because direct form structures are extremely sensitive to parameter quantization. The lattice structure implementation using one multiplier sections is as shown in Fig-6.

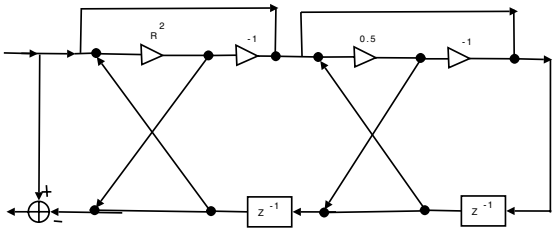

Fig. 6. Lattice Structure realization

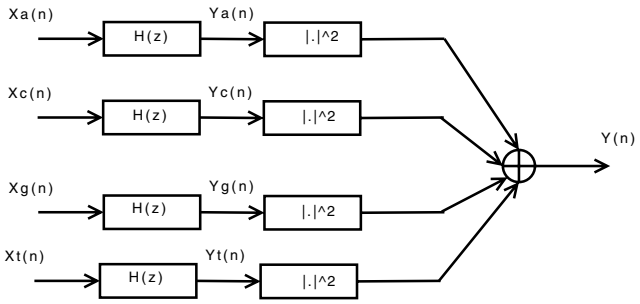

Fig. 7. Filtering setup for voss mapping

The filtering is done for the gene F56F11.4 in the C-elegans chromosome III. This gene is having five exons. The filtered outputs for Voss mapping and 2-simplex mapping are shown in Fig-9. The computational requirement of proposed mapping requires almost half of the requirement of Voss mapping. It is evident from Fig-7,8. There is no degradation in the performance as evident from Fig-9. The 2-simplex mapping is better symmetric than any other two dimensional mapping. This mapping scheme is independent of the query DNA sequence. The advantage with this mapping scheme is that it does not require any training and any computations before processing the DNA sequence.

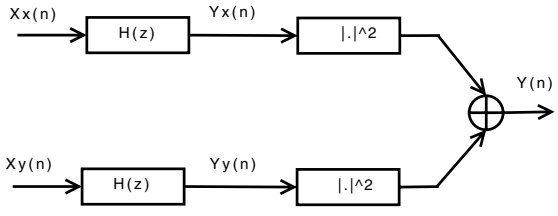

Fig. 8. Filtering setup for 2-simplex mapping

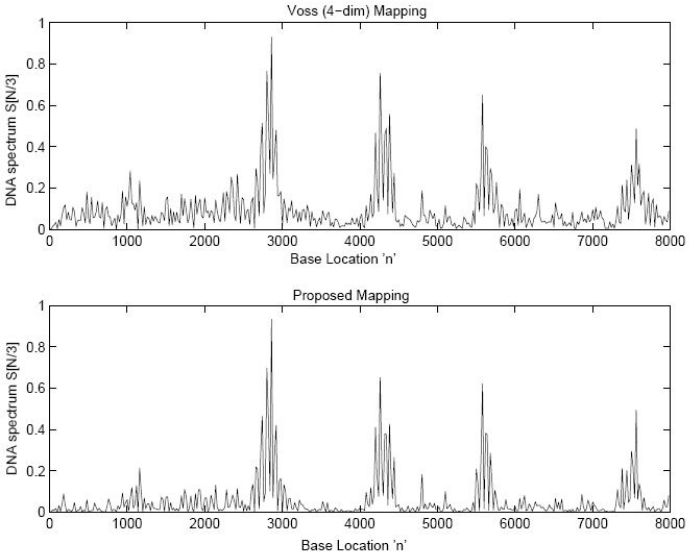

Fig. 9. Comparison of Voss mapping and proposed mapping

IV. Conclusions

A 2-simplex mapping can be used as a trade off between symmetry and computational complexity. The 2-simplex mapping reduces the computations by half compared to voss mapping which is a significant reduction. Since we are dealing with huge amount of data, every computation is of our concern.

References

[1] E. N. Trifonov and J. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence," *Proc. Natl. Acad. Sci. USA*, vol. 77, pp. 3816–3820, 1980.

[2] P. P. Vaidyanathan, "Genomics and Proteomics: A signal processor's tour," *IEEE circuits and systems magazine*, vol. 4, pp. 6–29, 2004.

[3] K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed. Printice-Hall of India Pvt Ltd, 2001.

[4] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *Journal of Theoretical Biology*, vol. 118, no. 3, pp. 295–300, Feb 1986.

[5] R. F. Voss, "Evolution of Long-Range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters, The American Physical Society*, vol. 68, no. 25, pp. 3805–3808, 1992.

[6] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 8–20, 2001.

[7] N. Rao and S. J. Shepherd, "Detection of 3-periodicity for small genomic sequences based on AR technique," *International Conference on Communications, Circuits and Systems,ICCCAS*, vol. 2, pp. 1032– 1036, June 2004.

[8] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Oxyford University Press, Bioinformatics*, vol. 16, pp. 1073–1081, 2000.

[9] P. P. Vaidyanathan and B. J. Yoon, "Digital filters for gene prediction applications," *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 306–310, Nov 2002.