Asian Journal of Convergence in Technology Volume IV Issue I
ISSN NO: 2350-1146 I.F-5.11

A Survey On Numerical Representation Of DNA Sequences

P.Kamala Kumari [1], Dr.J.Beatrice Seventline[2]

[1]Assistant Professor, Dept of ECE
Muffakham Jah College of Engineering and Technology
Hyderabad, India
kamalakumari16@gmail.com
[2]Professor, Dept of ECE
GITAM University
Visakhapatnam, India
seventline.joseph@gitam.edu

Abstract— **Genomic Signal Processing (GSP) has been a challenging area of research for the past two decades and has increasingly attracted the attention of researchers in the Digital Signal Processing field. The prerequisite for genomic sequence analysis is the conversion of Deoxyribo-nucleic acid (DNA) sequence (ATTCGA….) to a numerical sequence. Further, DSP algorithms can be implemented on these representations for genomic analysis like prediction of coding regions, protein coding regions, cancer cells. In this paper, the numerical representation of DNA sequences are presented and compared in terms of methodology, dimensionality, merits and demerits.**

Keywords— Bioinformatics,DNA,Genomic signal processing, DNA mapping.

I. INTRODUCTION

Bioinformatics is the rapidly developing area of computer science devoted to collecting, organizing, and analyzing DNA and protein sequences [4]. It is the science of refining biological information into biological knowledge using computers. The analysis, processing, and use of genomic signals for gaining biological knowledge constitute the domain of GSP. The aim of GSP is to integrate the theory and methods of signal processing with the global understanding of functional genomics, with special emphasis on genomic regulation. The prerequisite for applying the signal processing algorithms on DNA is the numerical representation or mapping of DNA. The conversion of genomic sequences from the symbolic form given in the public genomic databases into digital genomic signals allows using signal processing procedures for processing and analyzing genomic data.

This paper is organized as follows: Section II presents the biological concepts of DNA, Section III elaborates different mapping techniques Section IV presents comparison between mapping techniques in terms of dimension, merits, demerits and Section V gives the conclusion.

II. BASIC CONCEPTS OF MOLECULAR BIOLOGY

The main nucleic genetic material of cells is represented by DNA molecules and carries the hereditary information of organisms. The DNA double helix molecules comprise of two anti-parallel intertwined complementary strands, each is a helicoidally coiled heteropolymer as depicted in Fig.1.[4]

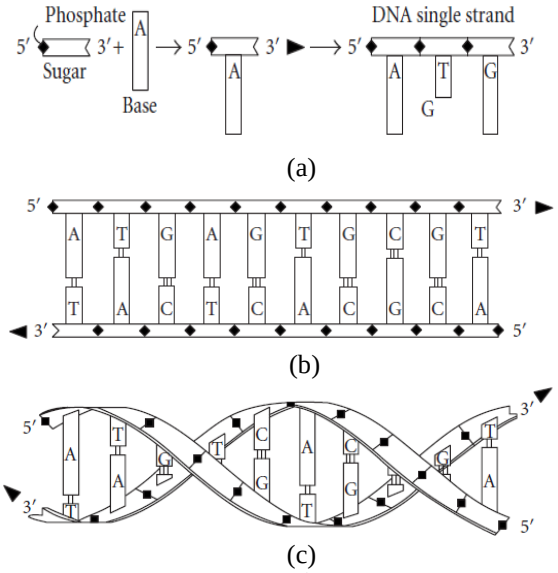

Fig 1: Schematic model of the DNA molecule.

The repetitive units are the nucleotide bases, each consisting of three components linked by strong covalent bounds: a monophosphate group linked to a sugar that has lost a specific oxygen atom, the deoxyribose group linked to a nitrogenous base.

The genetic information is encoded in the DNA sequence in the form of four important nucleotide bases called as

Adenine(A), Thymine(T), Guanine(G) and Cytosine(C). . Thymine (T) and cytosine (C) are called pyrimidines, adenine (A) and guanine(G) are called purines. The base "A" always pairs with base "T" and base "G" always pairs with base "C". As a consequence, the two strands of a DNA helix are complementary, store the same information, and contain exactly the same number of A and T bases and the same number of C and G bases so it is enough to take one strand into account for sequence analysis. The hydrogen bonds within the complementary base pairs keep the strands together. DNA sequence would be considered as a long stretch of nucleotide bases (e.g."CGATCGGCAATG.....") arranged in specific order.

A DNA strand can be divided into genes and intergenic spaces. Genes are responsible for protein synthesis. A gene can be further subdivided into exons and introns for cell with nucleus (eukaryotes). Cells without a nucleus are called prokaryotes and do not contain introns. Coding regions are called exons and non-coding regions are called introns. According to the current GenBank statistics, exons in the human genome account for about 3% of the total DNA sequence, introns for about 30%, and intergenic regions for the remaining 67%.The relationship between DNA sequence, genes, intergenic spaces, exons and introns is illustrated in Fig. 2[8].

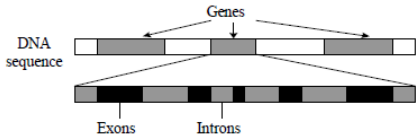

Fig 2: Organisation of genes depicting introns and exons

The coding regions within genes are denoted by start and stop codons. Codons are a subsequence of three letters within the DNA sequence. There are 64 possible codons (triplet). Each codon instructs the cell machinery to synthesize an amino acid. There are one start codon and three stop codons. A start codon signifies the beginning while a stop codon signifies the end of the protein-coding region. The rest of the codons correspond to one of the twenty possible amino acids of a protein.

III. DNA MAPPING

In order to apply digital signal processing, the nucleotides of a DNA sequence should be mapped to their corresponding numerical values. Numerical representation methods of DNA sequences can be broadly classified into three groups i)Fixed Mapping ii)Variable mapping and iii) Physico-chemical property based mapping. In the following sub sections we present the different mapping techniques in detail.

A. Fixed Mapping

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

One of the most oldest and popularly used mapping technique is the Voss mapping.It maps the four nucleotide bases A,T,G, and C into four binary indicator sequences $s_A(n)$, $s_T(n)$, $s_G(n)$ and $s_C(n)$ respectively[1]. In each indicator sequence, the numerical '1' indicates the presence of corresponding base and the numerical '0' indicates its absence. For example if the sequence s(n)=[…A T T G C C A T G….],then

$s_A(n)$=[…1 0 0 0 0 0 1 0 0….] (1)
$s_T(n)$=[…0 1 1 0 0 0 0 1 0….] (2)
$s_G(n)$=[…0 0 0 1 0 0 0 0 1….] (3)
$s_C(n)$=[…0 0 0 0 1 1 0 0 0….] (4)

This mapping is considered as a four dimensional mapping, since each base in genomic sequence is represented by four dimensional vector.

In Tetrahedron representation, the four sequences $[s_A(n),s_T(n),s_G(n),s_C(n)]$ are mapped to the four vertices of a regular tetrahedron. The four 3-D vectors drawn from the center to the corners of tetrahedron represented each of the four bases [2]. This representation reduces the number of indicator sequences from four to three but in a manner symmetric to all the four sequences. Each DNA character can be resolved into a four 3-D vectors (RGB) as

$[a_r, a_g, a_b]=[0, 0, 1]$ (5)
$[t_r, t_g, t_b]=[\ \ , 0, \ ]$ (6)
$[g_r, g_g, g_b]=[-\ \ , -\ \ , -\ ]$ (7)
$[c_r, c_g, c_b]=[-\ \ , \ \ , -\ ]$ (8)

The above coefficients can be used to define the following three numerical sequences

$X_r(n)= (2s_T(n) – s_C(n) – s_G(n))$ (9)
$X_g(n) = (s_C(n) – s_G(n))$ (10)
$X_b(n)= (3s_A(n) – s_T(n) – s_C(n) –s_G(n))$ (11)

Where $s_A(n),s_T(n),s_G(n)$ and $s_C(n)$ are four Voss indicators

The dimensionality of the tetrahedral representation can be reduced to two, by projecting the tetrahedron on a complex plane [3]. The complex representation of DNA sequence is given by

$$x(n)= a\, s_A(n) +c\, s_C(n)+g\, s_G(n) +t\, s_T(n). \quad (12)$$

Where a= 1+j, t=1-j, c=-1-j, and g=-1+j. In this representation, the complementary pairs of bases A-T and C-G are expressed by the symmentry with respect to the real axis. The purine/pyrimidine pairs have the same imaginary parts.

The nucleotide representation is further reduced to one-dimensional resulting in integer representation [4]. The technique is obtained by mapping numericals {0,1,2,3} to the four nucleotides as : T=0,C=1,A=2 and G=3.However,this mapping technique implies a structure on the nucleotides such as purine (A,G) > Pyrimdine(C,T).

In Real method [5], the four nucleotides are assigned with the numerical A=-1.5, T=1.5,C=0.5 and G=-0.5, which bears complementary property, is efficient in finding the complementary of a DNA sequence. For example in the sequence TGCAG represented by 1.5,-0.5, 0.5,-1.5,-0.5, the change of sign and reverse the representation -1.5,0.5,-0.5,1.5,0.5 yields the sequence ACGTC.

In Quaterion representation method [6], the nucleotides are mapped to four vectors symmetrically placed in the 3-D space that is oriented towards the vertices of a regular tetrahedron. The representation is three dimensional and the axes express the differences “weak minus strong bonds”, “amino minus Keto”, and “ purines minus pyrimidines”. By choosing { ±1} coordinates for the vertices of the embedding cube, the vectors that represent the four nucleotides take the simple form:

$$a =i+j+k\ ,\ c=i-j-k,\ g=-1-j+k,\ t=-1+j-k \quad (13)$$

In QPSK based method [7], the four nucleotides are assigned with the values a=j, c=1, t=-1, g=-j which is analogy to the QPSK modulation. For example if the sequence is CTAGT, the corresponding QPSK mapping representation is given by 1, -1, j, -j,-1.

B. Varibale Mapping

Variable mapping technique is based on twiddle factor[8], the four nucleotides A,G,T and C are represented by complex numerical value which varies with position in codon along the DNA sequence based on the 16th different twiddle factor W_N^K that occupy 16 different locations on the circle. The circle is divided into 4 quadrants for 4 nucleotides where A,C,G, and T are represented by the values of twiddle factors of 1st ,2nd, 3rd and 4th quadrant respectively. Unlike fixed mapping technique the values of each nucleotide will vary through the entire sequence depending on the location of codon and the position of nucleotide within the codon. The 1st nucleotide of a codon will be considered in the place of ‘1’ position. Similarly 2nd and 3rd nucleotides of the codon will be placed in 2nd and 3rd position as depicted in Fig.3.

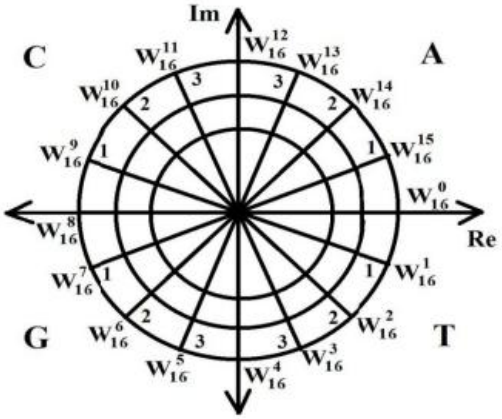

Fig.3.Variable mapping technique

C. Physico Chemical Property Based Mapping

In this type of mapping techniques, biophysical and biochemical properties of DNA molecules are used for DNA sequence mapping.This method is robust and used to search biological principles and structures in biomolecules. These techniques include the atomic number, the paired numeric, the DNA walk, the z-curve, the EIIP and the Pseudo EIIP representation.

A single atomic number indicator sequenced is formed by assigning each nucleotide with the atomic number as C=58, T=66, A=70 and G=78 in the DNA sequence [14]. For example if the sequence is ATTGC, the corresponding atomic number representation is as x(n)={70,66,66,78,56}.

In Paired Numeric representation [11], nucleotides (A-T,C-G) are to be paired in a complementary manner and values of +1 and -1 are used respectively to denote A-T and C-G nucleotide pairs.. According to this representation, for a DNA sequence ‘TAGCCAT..’, corresponds to the sequence x(n)= {+1,+1,-1,-1,-1,+1,+1…}.

For the study of scale-invariant long range correlations of DNA sequence , a graphical representation of DNA sequences called DNA walk has been adopted [12].This representation shows a graph of DNA sequence in which the walker steps up(+1) if the nucleotide is pyrimidine (T or G) , while the walker steps down(-1) if the nucleotide is purine (A or C).The DNA walk allows to visualize the concentration of purines and pyramidines in the DNA sequence. The graph continues to move upwards and downwards as the sequence progresses in a cumulative manner with its base along x-axis.

The Z-curve [13] is a three –dimensional zigzag curve that provides a unique representation for analysis and visualization of DNA sequence. The three components of the Z-curve $\{x_n\ y_n\ z_n\}$ represent three main dichotomies of the nitrogenous bases biochemical properties. They can arranged in classes

1) Molecule structure: A and G are Purine (R) while C and T are Pyrimidine(Y).

2) Strength of links: bases A and T are linked by two hydrogen bonds (W-weak) while C and G are linked by three hydrogen bonds(S-strong).

3) Radical content: A and contain the amino NH_3 group large groove(M-class) while T and G contain the keta group(K class).

The Z-curve is composed of a series of nodes P,P, P,……P with coordinates x_n y_n z_n ,where n=0,1,2,…L and L is the length of the DNA sequence.

$x_n = (A_n + G_n)-(C_n+T_n)$ R-Y (14)

$y_n= (A_n+C_n) –(G_n+T_n)$ M-K (15)

$z_n= (A_n+T_n)-(C_n+G_n)$ W-S (16)

The Electron-ion interaction pseudopotential (EIIP) representation gives the distributing of the free electrons energies along a DNA sequence and also the quasi-valence number associated with each nucleotide [10]. The energy of delocalized electrons in amino acids and nucleotides has been calculated as the EIIP. If these values are used in the string x[n], the sequence is named as 'EIIP indicator sequence',$x_e[n]$. For example if x[n]= A T T G C A A , then $x_e[n]$=[0.1260 0.1335 0.1335 0.0806 0.1340 0.1260 0.1260].

The optimized version of traditional EIIP representation, called Pseudo-EIIP representation [9]. The optimization is obtained using quasi-Newton algorithm based on the Broyden-Fletcher-Goldfarb-Shanno updating formula. In order to achieve consistency between the optimized characteristic values and the EIIP values, we need to ensure that the four variables are always positive and their numerical values are normalized at the end of each iteration such that their sum is always equal to the sum of the EIIP values. The Pseudo –EIIP representation of nucleotides are A=0.1994, T= 0.1933, G=0.0123 and C=0.0692

IV. Comparision between mapping techniques

In this section, all numerical representation methods and their merits and demerits are summarized and shown in Table 1.The dimension of each mapping technique is given in second column. It also summarizes the merits and demerits in third and fourth column respectively.

TABLE I:Dimension,Merits and Demerits of DNA Numerical Representation (1:Voss; 2:Terahedron; 3:Complex; 4:Integer; 5:Real; 6:Quaterion; 7:QPSK; 8:Variable; 9:Atomic Number; 10:Paired Numeric; 11:DNA walk;12: Z-Curve; 13:EIIP; 14:Pseudo EIIP)

	Dimension	Merits	Demerits
1	4	Efficient spectral detector of base distribution offering numerical and graphical visualization	Redundancy,linearly dependent set of representation
2	3	Periodicity detection	Reduced redundancy
3	1,4	A-T and C-G are reflecting complementary feature of nucleotides	Introduces base bias in time domain analysis
4	1	Simple technique	Mathematical properties not present in DNA sequence
5	1	A-T and C-G are complement	Mathematical properties not present in DNA sequence
6	1,4	Overcoming base bias	Working on DQFT only
7.	1	Follows the molecular structure	Randomness
8	3,4	Highly accurate for gene prediction	Computational complexity
9	1,4	Reflecting DNA physic chemical property	Requiring further exploration
10	2	Reflecting DNA structural property ; reducing complexity;	Requiring further exploration
11	1	Providing long range correlation information; offering graphical visualization.	Not suitable for lengthy (>1000 bases) sequences
12	3	Clear biological interpretation; reduced computation, offering numerical and graphical visualization.	Not suitable for lengthy (>1000 bases) sequences
13	1,4	Reducing computational overhead; improving gene discrimination capability	Failing to detect coding regions in some genomes.
14	1,4	Improved accuracy along with high efficiency	Requiring further exploration

V. Conclusion

The paper presents a review of various methods for numerical representation of DNA. Each of the DNA numerical representation offers different properties. In Fixed mapping representation methods, DNA nucleotides do not necessary reflect the original structure of DNA. The Variable mapping

rule is based on codon position and is more appropriate in gene prediction. The Physico Chemical Property based mapping exploits the structural differences of introns and exons. Fixed mapping techniques contain more redundant data than Physico Chemical Property based mapping. Hence we summarise that choice of appropriate representation depends on a particular application.

References

[1] R.F Voss,"Evolution of long range fractal correlations and 1/f noise in DNA base sequences".Physical review letters,1992,68(25):3805-3808.

[2] B.D Silverman and R.Linker, " A measure of DNA periodicity",[J] Theor.Biol. vol 118,pp.295-300,Febraury 1986.

[3] D.Anastassiou, Genomic Signal processing [M],IEEE signal processing magazine ,2001,18(4):8-20

[4] P.D.Cristea, Genetic signal representation and analysis [c]. in Proc.SPIE Inter. Conf. on Biomedical Optics , 2002,4623:77-84.

[5] N.Chakravarthy,A.Spanias,L.D Lasemidis and K.Tsakalis, Autoregressive modeling and feature analysis of DNA sequences[J]. EURASIP Journal of genomic signal processing ,January 2004,1:13-28

[6] M.Akhtar,J.Epps and E. Ambikairajah, "on DNA numerical representation of period-3 based exon prediction", in Proc of IEEE workshop on Genomic signal processing and statistics(GENSIPS) June 2007,pp 1-4.

[7] Singha Roy S, Barman S, " Identification of protein coding region of DNA sequencing using multirate filter",Computational Avan Commun Circuits Syst 2014. Doi: 10.1007/978-81-322-2274-3_16

[8] Singhan Roy S,Barman.S (2016) , " Polyphase filtering with variable mapping rule in protein coding region prediction",Microsytem Technologies,Vol-22,issue 4,doi: 10.1007/s00542-016-2884-5.

[9] P. Ramachandran, W. Lu, and A. Antoniou, "Filter-based methodology for the location of hot spots in proteins and exons in DNA," IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1598-1609, June 2012.

[10] Achuthsankar S. Nair and Sreenadhan S. Pillai, "A coding measure scheme employing electron-ion interaction pseudo potential (EIIP)," Bioinformation, vol. 1, pp. 197-202, October 2006.

[11] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), June 2007, pp. 1-4.

[12] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "Visualization and analysis of DNA sequences using DNA walks," Journal of the Franklin Institute, vol. 341, pp. 37-53, January-March 2004.

[13] R.Zhang and C.T.Zhang,"Z curves ,An Intutive Tool, for Visualizing and Analyzing the DNA sequences," J.BioMol. Struct. Dyn., vol. 11,pp. 767-782, 1994.

[14] Todd Holden, R. Subramaniam, R. Sullivan, E. Cheng, C. Sneider, G.Tremberger, Jr. A. Flamholz, D. H. Leiberman, and T. D. Cheung,"ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes," in Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE),vol. 6694, August 2007,pp. 669417-1 to 669417-10.