2012 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.31(2012) © (2012)IACSIT Press, Singapoore

Advanced Numerical Representation of DNA Sequences

Swarna bai Arniker [1,2], Hon Keung Kwan [2]

[1] Directorate of Laser Systems, Research Centre Imarat, Hyderabad 500 069, Andhra Pradesh state, India,

[2] Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Ontario,

Abstract. DNA sequence analysis using digital signal processing methods requires mapping of base sequence to numerical sequence. The choice of the numerical representation of a DNA sequence affects how well its biological properties can be reflected in the numerical domain for the detection and identification of the characteristics of special regions of interest. This paper presents various advanced methods of DNA numerical representation for DNA sequence analysis, their relative merits and demerits, and includes some concluding remarks.

Keywords: DNA numerical representation, fixed mapping, physic chemical property, statistical property

1. Introduction

Since the completion of the Human Genome Program (HGP) [1], there is a need to analyze information contained in a growing volume of deoxyribonucleic acid (DNA) sequence database of the human and model organisms. Digital signal processing (DSP) approach has become increasingly important in genomic DNA research to reveal genome structures to identify hidden periodicities and features which cannot be revealed by conventional DNA symbolic and graphical representation techniques [2]. In genomic signal processing (GSP), the mapping of the discrete bases of a DNA sequence to a discrete numerical sequence is required for DSP-based analysis [3]-[5]. A simple and commonly used mapping scheme is the Voss representation [6]. However, many other advanced methods have also been introduced such as the 2-bit binary [10], the 4-bit binary [11], the paired nucleotide [12]-[13], the 12-letter alphabet [14]-[15], the digital Z-signals [16], the phase specific Z-curve [17], the genetic code context [18], the inter-nucleotide distance [19], the correlation function [20], the Galois field [21], and the frequency of nucleotide occurrence [5,8]. Reference [7] describes about the binary; the integer, the real, the complex; and the DNA walk representations for autoregressive modeling and feature analysis of DNA sequences but the reasons for selecting each representation have not been discussed. Reference [8] discusses about the Voss, the tetrahedron, the real number and its variants, the complex, the quaternion, the EIIP, the paired numeric, and the Z-curve representations for period-3 based exon (coding region) prediction, the merits and demerits of each representation were not discussed extensively and the classification of representation was not carried out. To fill this gap, we made an attempt [9] to provide a survey on various numerical representation methods. We classified various numerical representations to two main groups; discussed each representation with a numerical example, its merits and demerits, and also some concluding remarks. Later, It was found from literature that there several other numerical representations which we could not discuss in [9] and this motivated us to write this paper to discuss about other advanced numerical representations in a similar fashion. In this paper we shall present, classify, compare, and discuss various advanced DNA numerical representation methods for digital signal processing and analysis.

2. Advanced Numerical Representation

DNA consists of double stranded anti-parallel helix built by concatenating nucleotides consisting of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Complementary property exists between DNA double-strands, as an A on one strand always binds with a T on the other strand, and similarly, a C always binds with a G. In order to apply digital signal processing, the bases of a DNA sequence are to be mapped onto their corresponding numerical values. Advanced numerical representation methods of DNA sequences can be broadly classified into three major groups as described in the following sub-sections.

2.1. Fixed Mapping Methods

In fixed mapping techniques, the nucleotides of DNA data are transformed into a series of arbitrarily numerical sequences. Fixed mapping include the 2-bit binary, the 4-bit binary, the paired nucleotide, and the 12-letter alphabet representations.

The 2-bit binary representation [10] maps the nucleotides A, C, G, and T into two-bit binary namely, 00, 11, 10, 01 respectively resulting into a 1-dimensional indicator sequence. Similarly, in the 4-bit binary encoding [11] the nucleotides A, C, G, and T are mapped as 1000, 0010, 0001 and 0100 respectively resulting into a 1-dimensional indicator sequence. The paired nucleotide representation [12]-[13] assigns binary values to a set of two letter DNA alphabets. In the first convention all A, T bases are assigned 0 and all C, G bases are assigned 1, In the second convention all C, T bases are assigned 0 and all A, G bases are assigned 1, and in the third convention all G, T bases are assigned 0 and all A, C bases are assigned 1 thus leading to a 1-dimensional indicator sequence of three different conventions.

The binary mapping of 12-letter alphabet representation [14]-[15] reflects the nucleotide composition within codons by capturing the differential base composition at each codon position, as $A_{12} = \{A_0, A_1, A_2, C_0, C_1, C_2, G_0, G_1, G_2, T_0, T_1, T_2\}$ where, for example, T_2 means that there is a nucleotide T in the third position of a codon in the given DNA sequence which is assigned a value 1 and the absence of a nucleotide is represented as 0. This mapping results in a 12-dimensional binary indicator sequence.

2.2. Physico Chemical Property based Mapping

In this type of mapping, biophysical and biochemical properties of DNA biomolecules are used for DNA sequence mapping, which is robust and is used to search for biological principles and structures in biomolecules. This mapping includes the digital Z-signals, the phase-specific Z-curve, and the genetic code context representations.

Digital Z-signals [16] decomposes the DNA sequence into three series of digital signals, based on Z-curves. These three series of digital signals, Δx_n, Δy_n and Δz_n can only have the values of 1 or -1. Δx_n is equal to 1 when the nth base is A or G (purine), or -1 when the nth base is C or T (pyrimidine); Δy_n is equal to 1 when the nth base is A or C (amino-type) or -1 when the nth base is G or T (keto-type). Similarly, Δz_n is equal to 1 when the nth base is A or T (weak hydrogen bond), or -1 when the nth base is G or C (strong hydrogen bond).

Phase-specific Z-curves [17] describe the distribution of bases at first, second and third codon positions, respectively resulting in 9-dimensional feature representation. The curve for the DNA sequence with bases at positions 0, 3, 6, ……, forms a phase-specific curve called phase-1 Z curve. Similarly, the Z curves with bases at positions 1, 4, 7, …., and 2, 5, 8, …., are called the phase-2 and phase-3 Z curves, respectively. Thus, the phase-1, phase-2, and phase-3 Z curves describe the distributions of bases at first, second and third codon positions respectively. For each phase-specific Z curve there are three components, as for the ordinary Z curve. The three components of the phase-1 Z curve are denoted by x_0, y_0, z_0, respectively, and $x_1, y_1, z_1, x_2, y_2, z_2$ are defined similarly.

Genetic code context (GCC) representation [18] incorporates the composition and distribution of the amino acid information in three reading frames. In this method, each consecutive codon from the three reading frames in the DNA sequence is converted to an amino acid and each amino acid in turn is represented by a unique complex number, of which the real parts and imaginary parts are from the hydrophobicity properties and residue volumes of the amino acids, respectively. This results in a single dimension indicator sequence in amino acid domain.

2.3. Statistical Property based Mapping

In this mapping the DNA alphabets are mapped in terms of different properties like the nucleotide distance, the correlation function, the Galois field, and the frequency of nucleotide occurrences. In the inter-nucleotide distance representation [19] each base symbol is replaced by a number k which is the base distance between the next similar base in the DNA sequence. In case a similar base is not found then the sequence value of that base is the length of the remaining base in the DNA sequence. It can be represented as one dimensional indicator sequence. The Correlation function [20] compares each base in a DNA sequence to its various neighbours along the sequence, scoring 1 when the two bases are identical and 0 otherwise and then summing them along the complete DNA sequence and this process is repeated from the first base to the last base in the DNA sequence.

A single Galois indicator sequence GF (4) [21] is formed by assigning the numerical values to the nucleotides A=0, C=1, G=3, and T=2 in a DNA sequence. In the frequency of nucleotide occurrence representation [5,8], the nucleotides are represented by their frequency of occurrence of A, C, G, and T in exons of GENSCAN data set, allowing either single-or four-sequence representation of DNA.

3. Merits and Demerits

Numerical representation of a DNA sequence when it is being used in conjunction with DSP techniques can identify hidden periodicities, nucleotide distributions, and features which cannot be revealed easily by conventional methods such as DNA symbolic and graphical representations. Each of the DNA numerical representations in fixed mapping offers different properties, and maps a DNA sequence into one to twelve numerical sequences. The 2-bit, and the 4-bit binary representation are applied mostly for neural network based systems for gene identification [10], and promoter prediction [11]. With paired nucleotide representation wavelets are used to smooth G+C profiles to locate characteristic patterns in genome sequences and secondly, a wavelet scalogram is used as a measure for sequence profile comparison in bacterial genomes [13]. DFT power spectrum using 12-letter alphabet representation produces a stronger spectral component for bactereophage phi-X174 [14] when compared with Voss representation for the same DNA sequence. The 12-letter alphabet is applied to find borders between coding and noncoding DNA regions by an entropic segmentation method for R. prowazekii, E.coli, M.jannaschii DNA sequences, whose results are more accurate than those obtained with moving window technique [15]. Digital Z-signal representation in conjunction with lengthen-shuffle Fourier transform algorithm has been successful in detecting period-3 property in short length coding regions in Fickett and Tang benchmark data sets [16]. Phase-specific Z-curves forms a 9-feature vector which helps to classify the coding from non-coding regions in the yeast genome at better than 95% accuracy by Fisher discriminate analysis [17].

Genetic code context generate different Fourier spectrum for different sequences, unlike Voss representation that produce same Fourier spectrum for two different sequences. From this point of view, GCC method will have more potential in gene finding and DNA sequence classification and function prediction. The period-3 property was successfully investigated in DNA sequences of myeloid zinc finger protein 1 splice variants ZNF42 gene of Human genome using the GCC mapping method [18].

The inter-nucleotide signal is a novel way of digital signal representation of genomic data that reveals the existence of discriminatory spectral envelope in the coding region and for some in promoter regions of coding sequences in the Burset and Guigo datasets [19]. These methods require more fine tuning for which more works need to be done, before it is ready for application [19]. Correlation function readily and evocatively displays regular patterns in DNA sequences with Fourier and wavelet transforms. This procedure has been applied to sequences from the human chromosome 22, to *nef* genes from various HIV clones and to myosin heavy chain DNA [20].

Galois field allows complex operations on a finite symbolic set and enables powerful tools for DNA analysis that can be explored further by genome researchers [21]. Frequency of nucleotide occurrence in exons is a key parameter for any DNA representation to be used for the detection of these regions. This representation validates that exons are rich in nucleotides ‘C’ and ‘G’ and provide marked improvement over other representations for exon detection in the GENSCAN data set of human genomic sequences using the DFT-based spectral content measure [5,8].

4. Conclusions

Which one of the numerical representation techniques is to be used in association with DSP depends on a particular application. Primarily, under fixed mapping representation methods, the 4-bit binary is found to be more suitable for neural network based classifiers for DNA sequence analysis. The paired nucleotide representation reflects the distribution of the nucleotides and has been successful in detecting structures with Fourier transform and wavelet analysis. The Fourier transform of 12-letter alphabet can be used as a preliminary spectral indicator of period-3 in a DNA sequence.

The digital Z-signals, and the GCC reflects the DNA physico chemical property very well, it could be explored further with wavelet transform or time-frequency analysis for identification of protein coding regions. These methods are robust, independent, less redundant, and have biological interpretation which can be regarded as useful representations for DNA sequence analysis.

Under statistical property based mapping the correlation function is a useful tool to visualize various different periodicities in a DNA sequence, the frequency of nucleotide occurrence is limited to standard data sets thus tending to be a model dependent method. Various other representations like the inter-nucleotide, and the Galois field, require more fine tuning for which more works need to be done, before it is ready to apply for various organisms.

The Genetic code context representation incorporates the amino acid information such as composition and its distribution in three reading frames, provides unique spectral signature for each DNA sequence compared to Voss representation and has high potential in gene finding and DNA sequence classification and function prediction. Genome researchers are encouraged to explore GCC representation widely.

5. References

[1] R. J. Robbins, B. David, and S. Jay. Informatics and the Human Genome Project. *IEEE Engineering in Medicine and Biology Magazine*, 1995, 14(6): 694-701.

[2] A. Roy, C. Raychaudhury, and A. Nandy. Novel techniques of graphical representation and analysis of DNA sequences- A review. *Journal of Biosciences*, 1998, 23(1):55-71.

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

[4] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton. Using signal processing techniques for DNA sequence comparison. *in Proc. of the 1989 Fifteenth Annual Northeast Bioengineering Conference*, 1989, 73-174.

[5] M. Akhtar, J. Epps, E. Ambikairajah. Signal processing in sequence analysis: advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*. 2008, 2(30):310-321.

[6] 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.

[7] N. Chakravarthy, A. Spanias, L. D. Lasemidis, and K. Tsakalis. Autoregressive modeling and feature analysis of DNA sequences. *EURASIP Journal of Genomic Signal Processing*, 2004, 2004(1):13-28.

[8] 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),* 2007, 1-4.

[9] H. K. Kwan and S.B. Arniker. Numerical representation of DNA sequences. *IEEE International Conference on Electro/Information Technology (EIT),* 2009. 307:310.

[10] R. Ranawana and V. Palade. A Neural network based multi-classifier system for gene identification in DNA sequence. *Neural Computing and Applications*, 2005, 14(2):122-131.

[11] B. Demeler, G. W. Zhou. Neural network optimization for E.coli promoter prediction. *Nucleic Acids Res.*, 1991, 19(7):1539-1599.

[12] P. Bernaola-Galvan, P. Carpena, R. Roman-Roldanet, J. L. Oliver. Study of statistical correlations in DNA sequences. *Gene*, 2002, 300(1-2):105-115.

[13] P. Lio, and M. Vannucci. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics*, 2000, 16(10):932-940.

[14] J. A. Berger, S. K. Mitra, and J. Astola. Power spectrum analysis for DNA sequences. *in Proc. of Seventh International Symposium on Signal Processing and its Applications*,2003, 2:29-32.

[15] P. B.-Galvan, I. Grosse. P. Carpena, J. L. Oliver, R. R.-Roldan, H. E. Stanley. Finding borders between coding and noncoding DNA regions by an enthropic segmentation method. *Physical Review Letters*, 2000, 85(6):1342-1345.

[16] M. Yan, Z.-S. Lin, C.-T. Zhang. A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, 1998,14(8):685-690.

[17] C.-T. Zhang and J. Wang. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nuc. Acids Res.*, 2000, 28(14):2804-2814.

[18] C. Yin, S. Yau. Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2008, 223-227.

[19] A. S. Nair, and T. Mahalakshmi. Visualization of genomic data using inter-nucleotide distance signals. *IEEE International Conference on Genomic Signal Processing GSP*, 2005.

[20] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, L. Marcourt. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *Journal of Theoretical Biology*, 2000, 206(3):323-326.

[21] G. L. Rosen. Signal processing for biologically-inspired gradient source localization and DNA sequence analysis. *Ph.D dissertation*, Georgia Institute of Technology, Atlanta, August 2006.