I.J. Image, Graphics and Signal Processing, **2012, 7, 47-53**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2012.07.05

Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions

J. K. Meher
Computer Science and Engineering, Vikash College of Engineering for Women, Bargarh, Odisha, India.
e-mail: jk_meher@yahoo.co.in

M. R. Panigrahi
Chemical Engineering, Vikash College of Engineering for Women, Bargarh, Odisha, India.
e-mail: madhaba_r@yahoo.com

G. N. Dash
School of Physics, Sambalpur University, Odisha, India
e-mail: gndash@ieee.org

P. K. Meher
Institute for Infocomm Research, Singapore
e-mail: pkmeher@i2r.star.edu.sg

***Abstract*—** Discrimination of protein coding regions called exons from noncoding regions called introns or junk DNA in eukaryotic cell is a computationally intensive task. But the dimension of the DNA string is huge; hence it requires large computation time. Further the DNA sequences are inherently random and have vast redundancy, hidden regularities, long repeats and complementary palindromes and therefore cannot be compressed efficiently. The objective of this study is to present an integrated signal processing algorithm that considerably reduces the computational load by compressing the DNA sequence effectively and aids the problem of searching for coding regions in DNA sequences. The presented algorithm is based on the Discrete Wavelet Transform (DWT), a very fast and effective method used for data compression and followed by comb filter for effective prediction of protein coding period-3 regions in DNA sequences. This algorithm is validated using standard dataset such as HMR195, Burset and Guigo and KEGG.

***Index Terms*—** Discrete Wavelet Transform, Comb filter, Indicator sequence, Protein coding regions

I. INTRODUCTION

DNA sequences can be considered as strings made of symbols drawn from the alphabet {*A*, *C*, *G*, *T*}.It is made of coding and non-coding regions. Coding regions are also called exons, code for proteins where as non-coding regions called introns or "junk" DNA [1]. In eukaryotes, the exons are found to be separated by introns, whereas in prokaryotes they are placed continuously without any introns in between. Computational gene prediction is based on sequence similarity searches and signal-based searches [2]. Exon detection deals with content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms. Many algorithms are applied for modeling gene structure, such as dynamic programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and neural network. Based on these models, a great number of gene prediction programs have been developed [3]. The signal processing approach has played a major role in gene prediction using period-3 property.

It is known that the protein coding regions of DNA sequences exhibit a period-3 behavior which results specifically from the existence of the codon sequences. Identification of period-3 regions therefore helps in predicting the gene locations; and allows the prediction of specific exons within the genes of eukaryotic cells [4]. In order to predict the location of protein coding region, a sliding data frame (sliding window) with a small step size is employed. The existence of three-base periodicity exhibited by the genomic sequence as a sharp peak at frequency *f*=1/3 in the power spectrum in the protein coding regions helps in the prediction of exons [5].

The first step in signal processing based gene prediction involves numerical representation of DNA string. The most fundamental representation involves the substitution of binary numbers to get four indicator sequences. Other methods have been adopted using *z*-curve [6], quaternion [7], complex numbers [8], EIIP [9], Gailos field assignment [10], frequency of nucleotide occurrence [11], paired numeric [12] to make indicator sequence in DSP methods to improve the sensitivity and selectivity.

The existing DSP tools for the identification of protein coding regions of DNA sequences based on the period-3 behavior are Discrete Fourier transform (DFT) and digital filter. DFT is used to detect period-3 property in DNA sequences [13] in which the DFT of length *N* for input indicator sequence is defined for four bases and the absolute value of power of DFT coefficients is plotted to get period-3 peak at coding regions. The digital filtering techniques such as the antinotch filter and multistage filter have been used to identify period-3 property in DNA sequences [14]. In digital filtering method for each indicator sequence of the respective base corresponding filter output is computed and the sum of the square of filter outputs is plotted to extract the period-3 region of the DNA sequence effectively. In order to reduce the computational complexity, comb filter has been used by J. K. Meher et al. [15] and better selectivity and sensitivity have been obtained.

Gene prediction in eukaryotes based on the DFT by spectral rotation measure is presented by Koltar and Lavner [16]. The 3-periodicity is explained in more detail by Tuqan and Rushdi [17] as related to the codon bias using two stage digital filter and multirate DSP model. Modified Gabor-Wavelet transform is used by Jesus et al. [18] for the identification of protein coding regions having advantage of being independent of the window length. The spectrum for DNA sequences is discussed based on an entropy minimization criterion by Galleani and Garello [19].

The dimension of the DNA string is huge; hence it requires large computation time. Newly determined genomes have to be stored and compressed in an efficient manner. For related species, they have to be organized in such a way that simple cross-referencing is possible. Efficient compression may also reveal some biological functions, aid in phylogenic tree reconstruction.

Xin Chen et al. have presented a lossless compression algorithm, GenCompress, for genetic sequences, based on searching for approximate repeats [20]. In 2004 Neva Cherniavsky and Richard Ladner have explored the utility of grammar-based compression of DNA sequences [21]. Grammar-based compression algorithms infer context-free grammars to represent the input data. The grammar is then transformed into a symbol stream and finally encoded in binary. There have been developed several special-purpose compression algorithms for DNA sequences such as Grumbach and Tahi [22], Lanctot, Li and Yang [23]. These algorithms use the structures and can achieve high compression ratio. Two characteristic structures of DNA sequences are known. One is called palindromes or reverse complements and the other structure is approximate repeats. Several specific algorithms for DNA sequences that use these structures can compress them less than two bits per symbol. Toshiko Matsumoto, Kunihiko Sadakane and Hiroshi Imai [24] have improved the Context Tree Weighting Method (CTW) so that characteristic structures of DNA sequences are available. The DNA compression has been performed by Don Adjeroh et al. in 2002 based on Burrows-Wheeler Transform (BWT) [25]. Repetition analysis is performed based on the relationship between the BWT and important pattern matching data structures, such as the suffix tree and suffix array.

DNA sequences are inherently random and therefore cannot be compressed efficiently. The methods used so far for compression do not capture the intricate structure of DNA sequences because of its vast redundancy, hidden regularities, long repeats and complementary palindromes. The exon identification task carried out by existing methods has its own limitations as it is observed in signal processing tools. Due to this gene prediction problem still remains a challenging task in terms of better sensitivity, selectivity and speed using existing tools. In such situations shortcomings of the previous approaches motivate to develop new approaches to have improved accuracy, speed and less computational complexity.

In this paper an efficient method using wavelet transform has been used for compression of DNA string without loss of any information that reduces the computational load effectively and the resulting reduced sequence is subjected to signal processing tool namely the comb filter that effectively use the period-3 property in a genomic sequence for the prediction of protein coding regions and has lower computational complexity In order to validate the results of the proposed predictor, prediction measures such as discriminating factor, sensitivity and specificity are evaluated with HMR195, Burset and Guigo and KEGG standard data sets.

The rest of the paper is organized as follows. Section-2 presents the proposed computationally method for compression of DNA string and method for the identification of protein coding regions. Section-3 presents the simulation and result analysis of the proposed methods and Section-4 presents the conclusions of this paper.

II. PROPOSED ALGORITHM

The objective of this paper is to compress the DNA sequence without loss of information of the coding regions to reduce the computational load. We are only interested in lossless compression algorithms. The motivation is using less memory to store reduced DNA sequence and reduce the computational time.

The overall process is represented in the form of a flow chart as shown in Figure.1. DNA sequence consisting of four bases is encoded with numerical values of respective dipole moments. Now the DNA sequence is converted to numerical sequence. The resulting sequence is compressed by using wavelet transform. Now in one level decomposition using db7 the sequence of detailed coefficients is subjected to a digital filter that can effectively detect the period-3 regions in the DNA sequence. Comb filter is chosen for its low computational complexity.

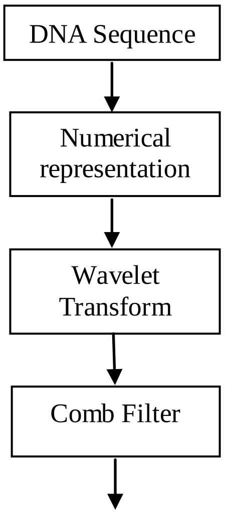

Figure.1 Overall process for detection of protein coding regions with lossless compression.

A. N umerical Presentation

The DNA sequence is a string of consisting of four nucleotides such as *A*, *T*, *C* and *G*. But genomic signal processing deals with numerical sequence. There are various numerical representations, but the methods having high sensitivity and specificity is obtained by using dipole moment of nucleotides. It is known that dipole moment of nucleotides helps in the detection of exon in a DNA sequence effectively [26]. If we substitute the dipole moments for *A*=0:4629, *G*=6:488, *C*=3.943 and *T*=1.052, we get a numerical sequence which represents the distribution of polarity of a chemical bond within a molecule along the DNA sequence.

For a DNA sequence of an organism, *x* = TATGAATAC, then substituting the values dipole moment of the corresponding nucleotide, we get *y* = [1.052 0.4629 1.052 6.488 0.4629 0.4629 1.052 0.4629 3.943]. Now the resulting numerical representation is subjected for compression.

B. Wavelet transform for compression of DNA string

Wavelets are a family of basis functions that can be used to approximate other functions by expansion in orthonormal series. They combine such powerful properties as orthonormality, compact support, varying degrees of smoothness, localization both in time or space and scale (frequency), and fast implementation. One of the key advantages of wavelets is their ability to spatially adapt to features of a function such as discontinuities and varying frequency behavior [27]. The compact support means that each wavelet basis function is supported on a finite interval and it guarantees the localization of wavelets. That is, a region of the data can be processed without affecting the data outside this region.

A wavelet transform is a lossless linear transformation of a signal or data into coefficients on a basis of wavelet functions. In signal processing, a transformation technique is used to project a data in one domain into another where hidden information can be extracted. A wavelet transform decomposes a signal into several groups of coefficients. These coefficient vectors contain information about characteristics of the data at different scales. Fine scales capture local details of coefficients and coarse scales capture global features of a signal. Performing the discrete wavelet transform (DWT) of a signal *x* is passing it through low pass filters (scaling functions) and high pass filters simultaneously. The result at each pass of the filtering of the signal is a convolution of the impulse response *g* of the filter and the signal. Mathematically, this result can be represented as

$$y(n)=\sum_{k=-\infty}^{\infty} x[k].g[n-k] \quad (1)$$

The frequency of the signal is halved after passing the signal through a filter. So, by Nyquist's rule, half of the samples can be discarded. This is achieved by down-sampling or decimation by a factor 2, that is, removing every alternative coefficient in *y*(*n*). Hence, after simultaneously passing a signal through high pass and low pass filters and the subsequent down-sampling, the number of coefficients will be equal to half the length of the original input for each filter. Therefore, the wavelet transform of a signal for both high pass filters and low pass filters can be represented by the following two equations

$$y_{low}(n)=\sum_{k=-\infty}^{\infty} x[k].g[2.n-k] \quad (2)$$

$$y_{high}(n)=\sum_{k=-\infty}^{\infty} x[k].h[2.n-k] \quad (3)$$

In matrix form, wt =WX where W =[L;H] where L and H are impulse responses of low pass and high pass filters and wt is wavelet transform of the input signal *X*. The two filters used at each stage of decomposition must be related to each other by *g*[*L* -1- *n*] = (1)*n*.*h*[*n*] where *g* and *h* are the impulse responses of the two filters and *L* is such that $0 \leq n < L$. These filters are known as quadrature mirror filters. The wavelet coefficients vector resulted from applying wavelet transform to a signal consists of both $Y_{\mathrm{high}}(n)$ (also called detailed coefficients) and $Y_{\mathrm{low}}(n)$ (also called approximation coefficients) coefficients in order. DWT proceeds further by recursively applying two convolution functions each producing an output stream that is half of the length of the original input, until the resolution (number of approximation coefficients) becomes one or resolution level zero. Number of detailed coefficients at each level *j* is equal to *n*/2*j*. The term 'scale' used in the context of wavelet transform at a level *j* is given by =2/*j*-1. The maximum level of decomposition depends on the wavelet function used for transformation. For example, the maximum level of decomposition of a signal *x* for Haar wavelet is given by $\log_2(x)$. Figure.2 depicts the entire process of DWT.

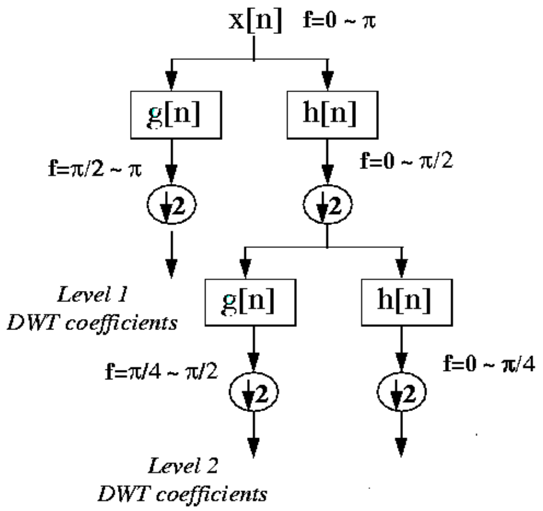

Figure.2 A two level DWT for *N* data.

The number of data is halved after every filtering and down sampling operation. A wavelet transform is applied on output of high pass filter recursively keeping the output coefficients of each low pass filtering operation at each stage. The wavelet transform of a data at any level *n* of decomposition consists of approximation coefficients only at *n*th level and all detailed coefficients up to *n*th level.

A number of wavelet families like symlet, coiflet, daubechies and biorthogonal wavelets are already in use. They vary in various basic properties of wavelets like compactness. Among them, Haar wavelets belonging to daubechies wavelet family are most commonly used wavelets in database literature because they are easy to comprehend and fast to compute. The db7 is used for one level decomposition of input query signal sequence. In this process the sequence is compressed to 50%. The resulting sequence of detailed coefficients is subjected to comb filter that exhibit period-3 property.

C. *Prediction of protein coding regions*

It is known that the protein coding region of DNA sequence exhibit period-3property. The existing signal processing methods such as discrete Fourier transform, digital filter such as notch filter and comb filter can effectively predict these regions. The comb filter has been choosen in this paper for its low computational complexity and better sensitivity and specificity [15].

Amplitude response of comb filter is comprised of a series of regularly spaced spikes of interleaved passbands and stopbands which looks like a hair comb. A comb filter can also be viewed as a notch filter in which the notches or the nulls occur periodically across the frequency band [28].

The difference equation of a comb filter can be written in a general form:

$$y(n) = [b_0 x(n) - b_1 x(n-n_1)] + a y(n-n_2) \quad (4)$$

where b_1 and *a*, respectively, denote the feed-forward and feedback gain coefficients, n_1 and n_2 are fixed delays, $x(n)$ denotes the *n*th sample of the input signal, $y(n)$ is the output at time instant *n*. Taking the *z*-transform we can get the transfer function of comb filter to be

$$H(z) = Y(z)/X(z) = b_0 \cdot z^{(n_2-n_1)} \left[\frac{z^{n_1} - b}{z^{n_2} - a} \right] \quad (5)$$

where b = b_1/b_0.

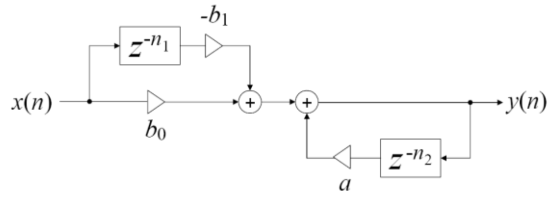

Figure.3 The signal flow graph for a comb filter defined by the difference equation of (4).

A generalized comb filter with both feedforward and feedback coefficients as shown in Figure.3 can effectively recognize protein coding regions.

III. SIMULATION AND RESULT ANALYSIS

The standard sequences have been extracted from a variety of sources that include the important genomes of Homo sapiens. Mainly, three data sets are used as bench mark for this purpose such as KEGG gene sequence database prepared by M. Kanehisa and S. Goto [29], the dataset prepared by Burset and Guigo [30] and HMR195 prepared by Sanja Rogic [31]. The single indicator sequence using dipole moment property of nucleotides is used as numerical representation [26]. These sequences are subjected to the proposed wavelet-based lossless compression scheme using one level of decomposition. In this work, we focus on lossless compression of DNA sequence. The resulting compressed sequence is subjected to comb filter which exhibit period-3 property and also it is computationally efficient.

In order to validate the results of the proposed predictor, prediction measures such as discrimination factor *D* [9], sensitivity (S_N), specificity (S_P) [32] which are defined as follows. In a good number of cases all the proposed methods performed well.

$$D = \frac{\text{Lowest of exon peaks}}{\text{Highest peak in noncoding regions}} \quad (6)$$

$$S_P = \frac{T_P}{T_P + F_P} \quad (7)$$

$$S_N = \frac{T_P}{T_P + F_N} \tag{8}$$

where T_P=true positive, F_P=false positive and F_N=false negative. T_P corresponds to those genes that are correctly predicted by the algorithm and also exist in the GenBank annotation. F_P corresponds to the coding regions identified by a given algorithm which are not present in the standard annotation. F_N is coding region that is present in the GenBank annotation but not predicted to be coding by the algorithm being used. Higher the value of *D* better is the discrimination. If *D* is more than one (*D*>1), all exons are identified without ambiguity. High sensitivity and specificity are desirable for higher accuracy.

The list of genes under study of different datasets and the performance analysis of various DSP approaches are shown in Table.1. It summarizes the simulation results of genes from different datasets. In all the examples cited the proposed encoding methods show better discrimination compared to the existing methods. The simulation result shows high discriminating factor, sensitivity and specificity for the proposed methods. The proposed method shows high peak at exon locations in compared to existing methods as shown in figures. Figure.4 shows the exon prediction results for gene F56F11.4a with accession no: AF099922 in the C. elegans chromosome-III without compression and with compression using wavelet transform. In both the cases the five peaks corresponding to the exons can be seen at the respective locations.

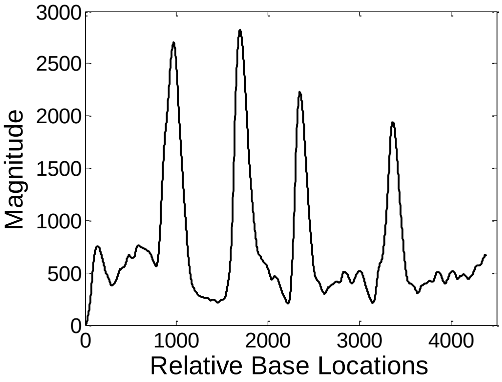

Figure.4 Gene F56F11.4a of C.Elegans chromosome III showing 5 exons using wavelet based compressed DNA.

The integrated approach using wavelet transform for reduction of DNA string and comb filter for effective detection of protein coding region can sense the exons effectively by showing high peak at gene locations with lower computation. Thus there is no loss of information.

TABLE.1 PREDICTION MEASURES OF PROTEIN CODING USING COMB FILTER OF RAW DNA SEQUENCE AND WAVELET BASED COMPRESSED DNA SEQUENCE

Gene Name Accession No.	Data	Length	Quality Measures		
			D	S_N	S_P
PP32R1, AF00A216, Homo Sapiens	Raw sequence	5761	7.5	1	0.75
	Compressed Sequence	2892	1	1	1
U17081 Human fatty acid binding protein (FABP3) gene	Raw sequence	9001	1	1	0.5
	Compressed Sequence	4600	1.5	1	0.66
AF092047 Homo sapiens homeobox protein Six3 (SIX3	Raw sequence	4441	1.1	1	0.6
	Compressed Sequence	2220	1.2	1	0.75
F56F11.4a C.Elegans Chromosomes III	Raw sequence	8000	1.2	1	1
	Compressed Sequence	4500	1.2	1	1
AB016625 Homo sapiens OCTN2 gene	Raw sequence	25861	0.82	1	0.75
	Compressed Sequence	12925	1.2	1	1

IV. CONCLUSION

Given the amount of data typically generated by DNA based experiments and gene prediction problem in particular, there is need to find methods to compress the data efficiently. In this work, starting with the nature of DNA sequence, we have proposed a simple model that captures both the Structure and the general statistics in DNA sequence. This work presents a computationally efficient method based on the wavelet transform designed for reduction of DNA string without loss of information that can be used for prediction of protein coding region efficiently using comb filter. The result shows that the proposed method can effectively detect the protein coding region. Ultimately the lossless compressed DNA sequence reduces the computational time effectively.

ACKNOWLEDGMENT

This research was supported by the management of the college. The authors thank Sri D. Murali kishna, the Chairman of Vikash College of Engineering for Women for his encouragement.

REFERENCES

[1] Z. Wang, Y. Z. Chen and Y. X. Li, “A Brief Review of Computational Gene Prediction Methods,” *Genomics Pro- teomics Bioinformatics*, Vol. 2, No. 4, 2004, pp. 216-221.

[2] J. W. Fickett, “The Gene Identification Problem: Over-view for Developers,” *Computers &*

Chemistry, Vol. 20, No. 1, 1996, pp. 103-118.

[3] Catherine Mathe, Marie-France Sagot, Thomas Schiex and Pierre Rouze, Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Research*, 2002, Vol. 30, No. 19, pp-4103-4117

[4] P. D. Cristea, "Genetic signal Representation and Analy-sis," Proceedings of SPIE Conference, *International Biomedical Optics Symposium (BIOS'02)*, Vol. 4623, 2002, pp. 77-84.

[5] B. D. Silverman and R. Linsker, "A Measure of DNA Periodicity," *Journal of Theoretical Biology*, Vol. 118, No. 3, 1986, pp. 295-300.

[6] R. Zhang and C. T. Zhang, "Z Curves, an Intuitive Tool for Visualizing and Analyzing the DNA Sequences," *Journal of Biomolecular Structure & Dynamics*, Vol. 11, No. 4, 1994, pp. 767-782.

[7] A. K. Brodzik and O. Peters, "Symbol-Balanced Quater-nionic Periodicity Transform for Latent Pattern Detection in DNA Sequences," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP* '05*)*, Vol. 5, 2005, pp. 373-376.

[8] P. D. Cristea, "Genetic signal Representation and Analy-sis," *Proceedings of SPIE Conference, International Biomedical Optics Symposium* (*BIOS*'02*)*, Vol. 4623, 2002, pp. 77-84.

[9] A. S. Nair and S. P. Sreenathan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopo-tential (EIIP)," *Bioinformation*, Vol. 1, No. 6, 2006, pp. 197-202.

[10]G. L. Rosen, "Signal Processing for Biologically-Inspired Gradient Source Localization and DNA Sequence Analy-sis," Ph.D. Thesis, Georgia Institute of Technology, At-lanta, 2006.

[11]A. S. Nair and S. P. Sreenathan, "An Improved Digital Filtering Technique Using Frequency Indicators for Lo-cating Exons," *Journal of the Computer Society of India*, Vol. 36, No. 1, 2006.

[12]M. Akhtar, J. Epps and E. Ambikairajah, "On DNA Nu-merical Representations for Period-3 Based Exon Predic-tion," *IEEE International Workshop on Genomic Signal Processing and Statistics*, Tuusula, 2007.

[13]D. Anastassiou, "Frequency-Domain Analysis of Bio-molecular Sequences," *Bioinformatics*, Vol. 16, No. 12, 2000, pp. 1073-1082.

[14]P. P. Vaidyanathan and B. J. Yoon, "The Role of Signal Processing Concepts in Genomics and Proteomics," *Journal of the Franklin Institute*, Vol. 341, No. 1-2, 2004, pp. 111-135.

[15]J. K. Meher, P. K. Meher and G. N. Dash "Improved Comb Filter based Approach for Effective Prediction of Protein Coding Regions in DNA Sequences", *International Journal of signal and information processing* (*JSIP*)*, Scientific Research Publishing*, Vol.2, N0.2, pp. 88-99, May-2011.

[16]D. Koltar and Y. Lavner, "Gene Prediction by Spectral Rotation (SR) Measure: A New Method for Identifying Protein-Coding Regions," *Genome Research*, Vol. 13, No. 8, 2003, pp. 1930-1937.

[17]J. Tuqan and A. Rushdi, "A DSP Approach for Finding the Codon Bias in DNA Sequences," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 2, No. 3, 2008, pp. 343-356.

[18]P. Jesus, M. Chalco and H. Carrer, "Identification of Pro-tein Coding Regions Using the Modified Gabor-Wavelet Tranform," *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, Vol. 5, No. 2, 2008, pp. 198- 207.

[19]L. Galleani and R. Garello, "The Minimum Entropy Mapping Spectrum of a DNA Sequence," *IEEE Transac-tion on Information Theory*, Vol. 56, No. 2, 2010, pp. 771-783.

[20]Chen, X., Kwong, S., and Li, M. A compression algorithm for DNA sequences and its applications in genome comparison. In Proceedings of the 10th Workshop on Genome Informatics (GIW-99) (Dec. 1999), pp. 52–61.

[21]Neva Cherniavsky and Richard Ladner, Grammar-based Compression of DNA Sequences, UW CSE Technical Report 2007-05-02, May 2004.

[22]Grumbach, S. and Tahi, F., A new challenge for compression algorithms: genetic sequences, *Information Processing & Management*, 30:875–886, 1994.

[23]Lanctot, J.K., Li, M., and Yang, E., Estimating DNA sequence entropy. *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, 409–418, 2000.

[24]Toshiko Matsumoto1,3 Kunihiko Sadakane2 Hiroshi Imai, Biological Sequence Compression Algorithms, Genome Informatics 11: 43–52 (2000)

[25]Adjeroh, D., Zhang, Y., Mukherjee, A., Powell, M., and Bell, T., DNA sequence compression using the Burrows-Wheeler Transform, Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society, Dec 2004, 303 – 313.

[26]J. K. Meher, M. K. Raval, P. K. Meher and G. N. Dash. "New Encoded Single Indicator Sequences based on Physico-chemical Parameters for Efficient Exon Identification", *International Journal of Bioinformatics Research and Applications*

(*IJBRA*),*Inderscience Publishers*, Vol 8, Nos. 1/2, pp-126-140, 2012,

[27]S. K. Mitra, "Digital Signal Processing," Tata McGraw-Hill, New Delhi, 2006.

[28]A. V. Oppenheim and R. W. Schafer, "Discrete-Time Signal Processing," Prentice-Hall Inc., Upper Saddle River, 1999.

[29]M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acid Research,* Vol. 28, No. 1, 2000, pp. 27-30. doi:10.1093/nar/28.1.27

[30]M. Burset and A. R. Guigo, "Evaluation of Gene Struc-ture Prediction Programs," *Genomics*, Vol. 34, No. 3, 1996, pp. 353-367. doi:10.1006/geno.1996.0298

[31]S. Rogic, A. Mackworth and F. Ouellette, "Evaluation of Gene Finding Programs on Mammalian Sequences," *Ge-nome Research,* Vol. 11, No. 5, 2001, 817-832.

[32]G. Aggarwal and R. Ramaswamy, "Ab Initio Gene Iden-tification: Prokaryote Genome Annotation with GeneScan and GLIMMER," *Journal of Biosciences*, Vol. 27, No. 1, 2002, pp. 7-14.

Jayakishan Meher has received his PhD from Sambalpur University, M.Tech in Electronics and Telecommunication Engineering from VSSUT, Burla (formerly known as University College of Engineering, Burla, Odisha, India) and M.Tech in Computer Science & Engg from RV University, India in 2012, 2002 and 2007 respectively. Currently he is Associate Professor and Head of the department of Computer Science and Engg in Vikash College of Engg for Women, Bargarh, Odisha, India. His research interests include digital signal processing, genome analysis, microarray data analysis, Protein analysis, metal binding, drug design and disease classification and other bioinformatics applications. Recently, he has developed interest in VLSI design for implementation of signal-processing algorithms on bioinformatics applications.

Madhab Ranjan Panigrahi has received his Ph.D, M.Tech and B.Tech in Chemical Engg from IIT Kharagpur, IIT Madras and NIT Rourkela, India respectively. He has hand on experience of research as a senior scientist in Regional Research Laboratory, Bhubaneswar, Odisha, India. Currently he is the principal of Vikash College of Engg for Women, Bargarh, Odisha, India. His research area includes hydrodynamics**,** environmental science, energy management and herbal bioinformatics.

Gana Nath Dash received his PhD from the Sambalpur University, India in 1992. He is currently a Professor in the Department of Physics, Sambalpur University, India. He has published more than 135 papers in journals of repute and proceedings of conferences. He is a senior member of IEEE and a Fellow and Life member of IETE. His research interests include studies on microwave and other devices. Recently, he has developed interest in ANN and signal-processing applications.

Pramod Kumar Meher has received PhD in Science from Sambalpur University, India in 1996. Currently, he is Senior Scientist in the Department of Embedded Systems, Institute for Infocomm Research, Singapore. Previously, he has worked as a Visiting Faculty in the School of Computer Engineering, Nanyang Technological University, Singapore. The main area of his research interest is design of dedicated and reconfigurable architectures for computation-intensive algorithms pertaining to signal processing, image processing, communication, intelligent computing and bioinformatics. Recently, he is tending his research towards more fundamental aspects of hardware design including the quantum dot cellular automata, and nano-circuits and systems.