Wavelet Transform for Detection of Conserved Motifs in Protein Sequences with Ten Bit Physico-Chemical Properties

J. K. Meher, M. K. Raval, P. K. Meher, and G. N. Dash

Abstract—**Detection of common motifs among proteins with low sequence identities provides important clues to the function of the proteins or to classify unknown proteins into proper families. Hence motif identification in protein sequences is essential for annotation of proteins from the sequence database among proteins with less than 30% homology. In the present work we have detected conserved regions in protein sequences using digital signal processing methods such as discrete Fourier transform (DFT) and wavelet transform with ten bit numerical representation of amino acids based on physico-chemical properties. The resulting ten bit numerical representation of each residue of the protein sequence has significant correlation with its biological activity. The conserved motifs are identified in peak regions from the DFT spectrum and wavelet spectrum. It is found that the new ten bit numerical representation using wavelet transform shows improved result than DFT. We have used wavelet transform to decompose protein sequences represented numerically by different indices such as positive charge, negative charge, polarity, charge, medium volume, small volume, aliphatic, aromatic chain and alicyclic character of the amino acids. The decomposed signals are then plotted to identify similar regions across all the proteins. Results indicate that wavelet transform using ten bit binary representation of physico-chemical properties is a promising approach for conserved motif detection. The proposed techniques are not only fast but also give the better interpretation of conserved motifs in protein sequences.**

Index Terms—**Conserved motif, discrete Fourier transform, physic-chemical properties, wavelet transform.**

I. Introduction

A sequence motif is a nucleotide or amino-acid sequence pattern that is biologically significant. Motif is a short sequence in a protein that regulates the function of the protein. There are nucleotide motif (in gene) and amino acid motif (in protein). Protein motif is used to assign the function of the protein. For proteins, a sequence motif is distinguished from a structural motif, a motif formed by the three dimensional arrangement of atoms in the molecule, which may not be adjacent. Motifs are small conserved regions within protein sequences. They usually carry specific structural or functional significance. Detection of common motifs among proteins with low sequence identities provides important clues to determine the function of proteins or to classify unknown proteins into proper families, since similarity search is often incapable of identifying proteins with less than 30% identity. Proteins having related functions may not show overall high homology yet may contain sequences of amino acid residues that are highly conserved.

Most classical methods for motif detection can be divided into two major categories. The first one involves crafting a consensus sequence or pattern to reflect conserved amino acids in the motif [1]-[3]. The second category of motif detecting algorithm involves using a scoring or weight matrix [4], [5]. Still other techniques used structural data [6] or statistics and data mining techniques such as MEME [7], [8]. There are software programs which, given multiple input sequences, attempt to identify one or more candidate motifs. One example is MEME, which generates statistical information for each candidate. Other algorithms include CisModule, AlignAce, PhyloGibbs, Weeder, Amadeus and FIRE. SCOPE is an ensemble motif finder that uses several algorithms simultaneously.

Recently signal processing tools have been used that play important role in protein sequence analysis. With the application of signal processing techniques like Discrete Fourier Transforms (DFT) to protein sequences, it is hardly surprising that Discrete Wavelet Transforms (DWT) have also been utilized for analyzing the interaction of protein sequences. It is used to study the similarity of two sequences across scales by taking the DWT of two protein sequences followed by cross correlation analysis at different scales [9]. DWT has been applied on hydrophobicity signals in order to predict hydrophobic cores in proteins [10], [11]. Protein sequence similarity has also been studied using DWT of a signal associated with the average energy states of all valence electrons of each amino acid [12]. Wavelet transform has been applied for transmembrane structure prediction [13]. CMDWave is a motif detection algorithm that predicts conserved motifs across multiple protein sequences by applying wavelet analysis and similarity detection techniques [14].

Sequence motifs are becoming increasingly important in the analysis of gene regulation. The abundance of both computationally and experimentally derived sequence motifs and their growing usefulness in defining genetic regulatory networks and deciphering the regulatory program of individual genes make them important tools for computational biology in the post-genomic era. There

Manuscript received January 11, 2012; revised February 27, 2012.

J. K. Meher is with the Department of Computer Science and Engg, Vikash College of Engg for Women, Bargarh, Odisha, India (e-mail: jk_meher@yahoo.co.in).

M. K. Raval is with the Department of Chemistry, Gangadhar Meher College, Sambalpur.

P. K. Meher is with the Department of Embedded Systems, Institute for Infocomm Research, Singapore.

G. N. Dash is with the the Department of Physics, Sambalpur University, India.

currently exist many publications with similar algorithms without a comprehensive benchmark so selecting one is not straightforward. Hence there is a hunt towards the development of new algorithm for prediction of conserved motifs in protein sequences which can have faster and accurate results. In the present work we propose a novel method to detect motifs using discrete Fourier transform (DFT) and wavelet transform. It uses ten bit numerical representation of the amino acids based on the physico-chemical properties. The proposed method is validated on light harvesting complexes protein sequences.

The rest of the paper is organized as follows. Section-2 presents digital signal processing approach for identification of conserved motif regions using discrete Fourier transform and wavelet transform. New numerical representation of ten digits binary is developed based on physico-chemical properties of amino acids. Section-3 deals with application of motif finding method on light harvesting complexes proteins that is taken as case study. It focuses on the results of the proposed methods. Section-4 presents the conclusions of this paper.

II. Proposed Methods

The signal processing methods play important role in predicting conserved regions in the protein sequences. It can detect the periodicity of amino acids in the residue locations. Due to this property conserved regions can be detected efficiently across multiple protein sequences. The technique is based on first carrying out a multiple sequence alignment of the input query sequences to make the sequences of the same size. This is followed by a mapping from string space to signal space using ten physico-chemical properties of amino acids for the query proteins. In this example, the protein sequence is converted to a ten bit binary representation using amino acid properties namely positive charge, negative charge, polarity, charge, medium volume, small volume, aliphatic, aromatic chain, alicyclic character. Bits are 1 for these characters and 0 for absence of the character [15]. The most significant bit (MSB) is 1 for all amino acids (Table 1).

A. DFT Method

The input protein sequences of Lhca1-4 (tomato), Lhcb1 (pea), Lhcb1, Lhcb2, Lhcb5, Lhcb6 (tomato) are aligned by Clustal-W [16]. The matrix is transposed to get the amino acid sequence of each column into row. The numerical values corresponding to the different amino acids are concatenated in the order of occurrence of the amino acids. The resulting numerical sequence then becomes the equivalent representation of the protein sequence. DFT is then applied to the numerical sequences. The Fourier transform of the numerical sequence $x(n)$ for every residue position is calculated and sum of the square is plotted against relative residue location.

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \quad 0 \le k \le N\text{-}1 \qquad (1)$$

The total power at frequency k then be expressed as

$$S[k] = |X[k]|^2 \qquad (2)$$

The highly conserved regions shows high peak that is used to obtain the conservation factor C. This is normalized to the range [0, 1]. A threshold ε is then applied to C for which $C > \varepsilon$ being dismissed as being dissimilar. The algorithm then outputs the conserved motifs across the sequences.

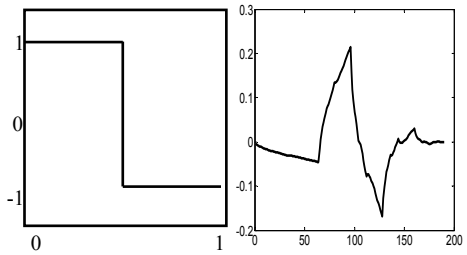
B. Wavelet Transform Method

Wavelet transform is a tool that can process both the stationary as well as non-stationary signal and has got multi-resolution capabilities. The main feature of wavelet transform is their ability to represent signals so as to obtain simultaneous time and scale localization of the signal. Due to these advantages it has been used effectively in many application areas of bioinformatics such as detecting patterns in DNA sequences, protein structure classification and microarray data analysis. A wavelet is a waveform that is localized in both time and frequency domains. Wavelets are generated from a single basic function called a mother wavelet, ($\Psi(t)$), by translation and dilation (scaling) operations.

$$\psi_{a,b} = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) \qquad (3)$$

where a is the scale parameter, b is the translation parameter, and the factor $1/\sqrt{|a|}$ is used to ensure that the energy of the scaled and translated versions are the same as that of the mother wavelet. Hence, a family of scaled and translated wavelets can be created based on the scaling and translation coefficients a and b [17].

The commonly used wavelets in practice are Haar, Daubechies, Gaussian wave, Mexican hat and Morlet wavelets. The selection of particular wavelet for any analysis depends on the kind of signal being studied and kind of signal variation to be captured. In case of analysis of protein sequence signal for detection of conserved regions, the Haar wavelet and Daubechies wavelet seemed to be choice. The mother wavelet of Haar wavelet is shown in Fig.1 (a). The Daubechies wavelet (db1) is the same as Haar wavelet. The wavelet function psi of the next member db2 of the family is shown in Fig.1(b).



(a) (b)

Fig. 1. (a) Haar wavelet, (b) Daubeties wavelet (db2).

The input protein sequences are aligned by Clustal-W [10]. The matrix is transposed to get the amino acid sequence of each column into row. The 10-bit numerical values corresponding to the different amino acids are concatenated in the order of occurrence of the amino acids. The resulting

numerical sequence then becomes the representation of the protein sequence. The operation was carried out by having one level wavelet decomposition of the protein sequences of the numerical sequence $x(n)$ for every residue position of multiple protein sequences and sum of the square of the approximation coefficients is plotted against relative residue location. The highly conserved regions shows high peak that is used to obtain the conservation factor C. This is normalized to the range [0, 1]. A threshold ε is then applied to C for which $C > \varepsilon$ being dismissed as being dissimilar. The algorithm then outputs the conserved motifs across the sequences.

The existing technique [9] can identify the functional similarity of two protein sequences based on their sequence-scale similarity vector; it does not say anything about specific regions of conserved motifs in the two sequences. This is more important when comparing multiple sequences with very little sequence similarity. The proposed technique gives the better interpretation conserved of protein sequences as well as it is fast.

IV. Results and Discussion

The proposed motif prediction methods described in this paper was validated on light harvesting chlorophyll polypeptide complexes (LHC). The sensitivity results for the different motif lengths shows high peak in the plot. From the plot, it was observed that as the motif length increases, sensitivity increases. Specificity also increases as the motif length increases.

The conserved residues extracted from the plots are Asp(Asn) 85, 89, 233, 258, 281; Glu(Gln) 113,181, 189, 252, 269, 280; His 116, 285. These constitute the set of residues involved in binding to central metal ion Mg of chlorophylls a/b [18]. Hence these are the functionally important residues. Further, some conserved motifs of sequences are found. These are 83-93 (PGDY(F)GW(F)DP(T)A(L)GL); 113-126 (EV (L) IHC(S/G/A)RWAMLGALG); 232-237 (FDPLGL); 281-286 (NLA(F/L)ADHL). The extent of conservation is quantified at each position along the sequence and then normalized. A threshold is determined to show highly conserved regions. Conservation curve is obtained by plotting magnitude versus residue locations.

Fig.2 (a) shows the results from the application of the DFT-based motif detection technique and Fig.2 (b) shows the results from the application of the wavelet-based motif detection technique to the data set discussed above. As can be observed from the figures, the algorithm is able to correctly identify the different conserved regions. The figures also show that not all conserved motifs are seen in the same frequency band. The results in Fig.2 (a) and (b) indicate that wavelet transforms can detect conserved motifs in proteins with low sequence similarities. Most conserved motifs indicated by DFT are clearly identifiable by our method. However, there is a prominent discrepancy in Figure 2, where wavelet method predicts at location 242 and 245 conserved motif. These two motifs are not indicated by DFT method. This regions, residue 242, 245, 259 are predicted as conserved regions. The wavelet analysis itself is unable to align sequences. Its strength is to determine whether a given set of sequence has some common feature that can only be seen at certain decomposition levels. db2 wavelet has also shown a better correlation with the experimental results. For sequences having high similarity shown in Fig. 2, the wavelet method clearly identified all conserved motifs. Hence it is to be noted that ten bit numerical representation of each residue of the protein sequence has significant correlation with its biological activity.

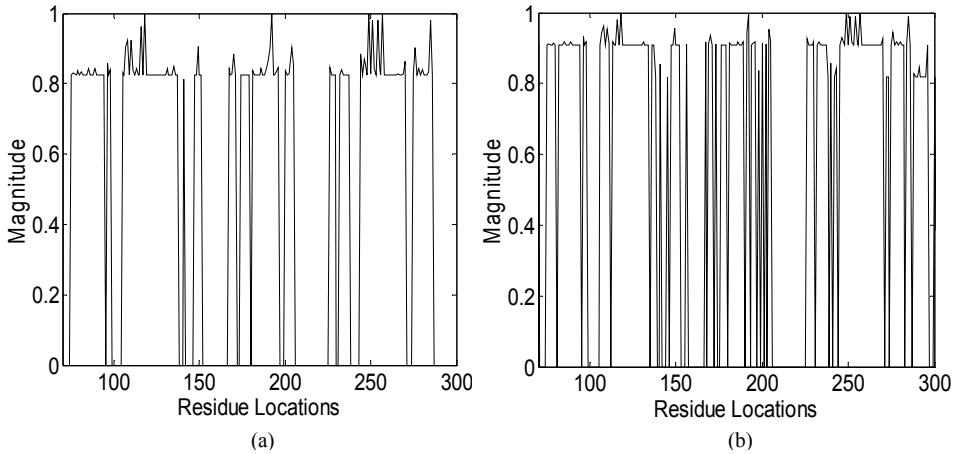


Fig. 2. Conserved motif detection of LHC using: (a) DFT , (b) DWT with ten bit binary representation

V. Conclusion

Discrete Fourier transform and wavelet transform method can successfully yield conserved sequence motifs in proteins with distantly related or even unrelated on the basis of function. It is to be noted that ten bit numerical representation of amino acids of the protein sequence based on physic-chemical properties has significant correlation with its biological activity. The proposed techniques are not only fast but also give the better interpretation in detecting conserved motifs of protein sequences. This will be helpful in investigating the sequence motifs for specific localized functions namely, ligand binding or conserved structural moiety.

TABLE I: TEN BIT BINARY REPRESENTATION OF AMINOACID CHARACTERS

Amino acid	MSB	Positive	Negative	Polar	Charged	Small	Extremely small	Hydrocarbon chain	Aromatic	Alicyclic
Ala	1	0	0	0	0	1	1	0	0	0
Arg	1	1	0	1	1	0	0	0	0	0
Asn	1	0	0	1	0	1	0	0	0	0
Asp	1	0	1	1	1	1	0	0	0	0
Cys	1	0	0	0	0	1	0	0	0	0
Gln	1	0	0	1	0	0	0	0	0	0
Glu	1	0	1	1	1	0	0	0	0	0
Gly	1	0	0	0	0	1	1	0	0	0
His	1	1	0	1	1	0	0	0	1	0
Ile	1	0	0	0	0	0	0	1	0	0
Leu	1	0	0	0	0	0	0	1	0	0
Lys	1	1	0	1	1	0	0	0	0	0
Met	1	0	0	0	0	0	0	0	0	0
Phe	1	0	0	0	0	0	0	0	1	0
Pro	1	0	0	0	0	1	0	0	0	1
Ser	1	0	0	1	0	1	1	0	0	0
Thr	1	0	0	1	0	1	0	0	0	0
Trp	1	0	0	1	0	0	0	0	1	0
Tyr	1	0	0	1	0	0	0	0	1	0
Val	1	0	0	0	0	1	0	1	0	0
Gap	0	0	0	0	0	0	0	0	0	0

ACKNOWLEDGMENT

The author would like to express his special thanks to Prof. (Dr.) M. R. Panigrahi, Principal, Vikash College of Engg for Women, Bargarh, India for constant support and strong encouragements to carry out the research studies. The author is indebted to Shri. D. Murali Krishna, Chairman of VCEW for providing the facilities and constructive comments for such an achievement.

REFERENCES

[1] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status," *Nucleic Acids Res*. vol. 30, pp. 235-238, 2002.

[2] C. G. Nevill-Manning, T. D. Wu, and D. L. Brutlag, "Highly specific protein sequence motifs for genome analysis,"in *Proc. Natl. Acad. Sci. USA* 95, 5865-5871, 1998.

[3] C. O. Pabo and R. T. Saue, "Transcription factors: structural families and principles of DNA recognition," *Annu. Rev. Biochem.* 61, pp. 1053-1095, 1992.

[4] I. B. Dodd and J. B. Egan, "Improved detection of helix-turn-helix DNA-binding motifs in protein sequences," *Nucleic Acids Res*. **vol.** 18, pp. 5019-5026, 1990.

[5] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," in *Proc. Natl. Acad. Sci. USA* 84, pp. 4355-4358, 1987.

[6] N. Leibowitz, R. Nussinov, and H. J. Wolfson, "MUSTA - a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins," *J. Comput. Biol*. vol. 8, pp. 93-121, 2021.

[7] T. L. Bailey and C. P. Elkan, "Fitting a mixture model by expectation-maximization to discover motifs in biopolymers," in *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 28-36, 1994.

[8] T. L. Bailey, M. E. Baker, and C. P. Elkan, "An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases," *J. Steroid Biochem. Mol. Biol*. vol. 62, pp. 29-44, 1997.

[9] C. H. D. Trad, Q. Fang, and I. Cosic, "An overview of protein sequence comparisons using wavelets," in *Proceedings of the IEEE-EMBS*, pages 115-119, 2001.

[10] H. Hirakawa and S. Kuhara, "Prediction of hydrophobic cores of proteins using wavelet analysis," *Genome Inform. Ser Workshop Genome Inform*, vol. 8, pp. 61–70, 1997.

[11] H. Hirakawa, S. Muta, and S. Kuhara, "The hydrophobic cores of proteins predicted by wavelet analysis," *Bioinformatics*, vol. 15, pp. 141–148, 1999.

[12] C. de Trad, Q. Fang, and I. Cosic, "Protein sequence comparision based on the wavelet transform approach," *Protein Eng*, vol. 15, pp. 193–203, 2002.

[13] K. B. Murray, D. Gorse, and J. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," *J. Mol. Biol*, vol. 316, pp. 341–363, 2002.

[14] A. Krishnan, K. B. Li, and P. Issac, "Rapid detection of conserved regions in protein sequences using wavelets," *In Silico Biol.* 4 0013, 2004.

[15] M. J. Zvelebil, G. Barton, F. R. Tayler, and M. J. E. Sternberg, *J .Mol Bio*, vol. 195: 957-961, 1987.

[16] J. D. Thompson, G. D. Higgins and T. J. Gibson, *Nucleic Acids Res*: vol. 22, pp. 4673–4680, 1994.

[17] S. Mallat. *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.

[18] W. Kuhlbrandt, D. N. Wang and Y. Fusiyoshi, "Tuning analysis for the high-*Q* class-E power amplifier," *Nature*, London. vol. 367, pp. 614-621, 1994.

Jayakishan Meher was born in Odisha, India on August 21, 1967. He has obtained M.Tech in Electronics and Telecommunication Engineering from University College of Engineering, Burla, Odisha, India in 2002 and M.Tech in Computer Science and Engineering from RV University, India in the year 2007. In year 2012 he received Ph.D from Sambalpur University, Odisha, India. Currently he is Associate Professor and Head of the Department of Computer Science and Engg, Vikash College of Engg for Women, Bargarh, Odisha, India. His research interests include digital signal

processing, genomic and proteomic data analysis, microarray data analysis . Recently he has taken interest in working with drug design and analusis and architecture design specifically on signal processing based bioinformatics applications.

Mukesh Kumar Raval received his PhD from Sambalpur University, Orissa, India. He was Professor and Head of the Department of Chemistry, Gangadhar Meher College, Sambalpur. He received his training in Molecular Biophysics from the Molecular Biophysics Unit, Indian Institute of Science, Bangalore. He has research interest in the area of photosynthesis, structure and function of proteins, molecular modelling and drug design.

Pramod Kumar Meher has received the first-class degrees of BSc (Honours) and MSc in Physics, and PhD in Science, all from Sambalpur University, Sambalpur, India, in 1976, 1978 and 1996, respectively. 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.

Gana Nath Dash received his PhD in Physics from the Sambalpur University 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.