Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

PMME 2016

Protein Sequence Similarity Analysis Using Computational Techniques*

Nikhila K. S., Dr. Vrinda V. Nair*

Department Of Electronics And Communication, College Of Engineering, Trivandrum-695016, India
Department Of Electronics And Communication, College Of Engineering, Trivandrum-695016, India

Abstract

Protein sequence basically consists of sequence of amino acid bases. Computational analysis of sequences requires mapping of amino acid bases into corresponding numerical values. Mapping involves various techniques which use the physicochemical and biochemical properties of amino acids in protein sequences, such as polarity, solubility, hydropathy, and so forth. Sequence analysis can be done using techniques like Fourier transform, Wavelet transform and the like. In this work a novel sequence similarity algorithm is developed incorporating an optimal mapping and corresponding analysis method which is superior in time of execution in comparison to existing methods.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Protein sequence similarity; Hurst exponent, Entropy, Phylogenetic tree;

1. Introduction

Proteins are the fundamentals of life and hence they are subject to a wide variety of research activities within molecular biology. Numerous proteins are synthesized by all living organisms based on 20 different amino acids. One of the key tasks related to proteins is the similarity comparison of protein sequences, which helps the prediction and classification of protein structure and function. Generally, the biological function of a protein is determined by

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

its three-dimensional structure which is dependent on the linear sequence of amino acids. Rigden [1] presented that one of the fundamental principles of molecular biology is that proteins having similar sequence possess similar functions.

The existing mathematical approaches for protein sequence similarity analysis are based on sequence alignments [2]. For example, the methods developed for classifying protein sequences by searching against protein databases by their sequence alignments are PSI-BLAST [3] and methods based on hidden Markov models (HMM) [4]. But these methods build a model for a single protein family and check out the fitness of each candidate sequence in the model [5]. So they fail when the query protein lacks proper sequence similarity. Wavelet transform, an effective tool in signal processing is commonly used in bioinformatics to analyse the protein sequences [6], [9]. Based on discrete wavelet transform (DWT), a new concept of similarity of protein sequence, sequence-scale similarity, has been proposed in [13] to identify the functional similarity of two protein sequences. Lina Yang presented sequence analysis using a hybrid method involving discrete wavelet transform and fractal dimension [7]. Protein sequence comparison can also be done based on sparse representation [8]. A compound method for protein secondary structure prediction was proposed in [14]. A frequency domain approach to protein sequence similarity analysis and functional classification is represented in [10]. A new approach to analyse similarity between protein sequences is empirical mode decomposition [11]. Charalambos Chrysostomou used Discrete Fourier transform for protein sequence analysis [12]. In this work phylogenetic tree is obtained which is similar to the reference tree generated using alignment tools which requires parameter optimization.

This work aims at developing a novel sequence similarity algorithm which uses an optimal mapping and feature extraction method. An evolutionary tree is generated that possess more similarity to the reference tree in comparison with the existing ones. For analyzing the protein sequences by computational methods they are to be converted to numerical signals. Various mapping techniques which involve the physicochemical and biochemical properties of the amino acids in the protein sequence chains are used. After obtaining numerical sequences in correspondence with the amino acid sequences they are analyzed using any of the computational tools. Analysis techniques used in the proposed method are Hurst exponent, entropy and standard deviation which results in an improved phylogenetic tree construction.

The paper is organised as follows: Section II presents the materials and methods used, Section III presents the results obtained. Finally, concluding remarks are given in Section IV.

2. Materials And Methods

2.1 Dataset

2.1.1 Protein sequences

Nine CD4 protein sequences are used in this study. CD4 (Cluster of Differentiation 4) [15] is a glycoprotein and was first discovered in late 1970. The main role of CD4 is to act as a co-receptor along with the T-cell receptor with an antigen-presenting cell. These sequences are collected from the databases UNIPROT and NCBI. Protein sequences used in this work are listed in table I.

2.1.2 Amino acid indices

Mapping of the amino acid bases into numerical values is an important step in the computational analysis of the protein sequences. In this work, amino acid indices are used for sequence mapping, where each index represents a unique biological feature. The indices are derived from the databases AAINDEX [16] and APDbase. Around 500 indices are present, among which we took different combinations of 25 indices.

TABLE 1
CD4 PROTEINS

Sl. No.	Uniprot ID	Organism	Protein length
1	P01730	Human	458
2	P16004	Chimpanzee	458
3	P79185	Crab-eating Macaque	458
4	P79184	Japanese Macaque	458
5	P16003	Rhesus Macaque	458
6	Q08340	Pig-tailed Macaque	458
7	Q29037	Common Squirrel Monkey	457
8	Q08338	Green Monkey	458
9	Q8HZT8	White-tufted-ear Marmoset	457

2.2 Methodology

2.2.1 Mapping

Protein sequences are normally represented as character data, as they are the linear combinations of around twenty amino acids. Analysis using signal processing techniques is difficult if the sequences are obtained as character data. For analysis to become easier these are to be converted to signals. The physicochemical properties, which are parameters having predefined values are used for the conversion to numerical data. Mapping of all the protein sequences in table I is done using combinations of amino acid indices in table II. The nine amino acid sequences are now converted to nine numerical data series.

TABLE 2
AMINO ACID INDICES USED

Sl. No.	Name	Description
1	ZIMJ680102	Bulkiness
2	ZIMJ680104	Isoelectric point
3	HUTJ700102	Absolute entropy
4	DAWD720101	Size
5	GRAR740102	Polarity
6	GRAR740103	Volume
7	FASG760101	Molecular weight
8	FASG760102	Melting point
9	FASG890101	Hydrophobicity index
10	ZHOH040101	The stability scale from the knowledge-based atom-atom potential
11	OOBM770103	Long range non-bonded energy per atom
12	MANP780101	Average surrounding hydrophobicity
13	WOLR790101	Hydrophobicity index
14	FAUJ880101	Hydration potential
15	FAUJ880102	Smoothed upsilon steric parameter
16	ARGP820101	Hydrophobicity index
17	VELV850101	Electron-ion interaction potential
18	FAUJ880111	Positive charge
19	FAUJ880112	Negative charge

20	FAUJ880109	Number of hydrogen bond donors
21	KYTJ820101	Hydropathy index
22	BHAR880101	Average flexibility indices
23	Proscale 4	Recognition factors
24	Nl	Long-range contacts
25	Rk	Relative connectivity

2.2.2 Feature extraction

In the next stage feature extraction of these numerical series is done using various sequence analysis techniques. Three analysis techniques are mainly used, Hurst exponent, Entropy, Standard deviation.

i) HURST EXPONENT

The Hurst exponent is used as a measure of long term memory of time series. It relates to the autocorrelations of the time series, and the rate at which these decrease as the lag between pairs of values increases. The Hurst exponent is referred to as the "index of dependence" or "index of long range dependence". It quantifies the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction. A value *H* in the range 0.5–1 indicates a time series with long term positive autocorrelation, meaning both that a high value in the series will probably be followed by another high value and that the values a long time into the future will also tend to be high. A value in the range 0 – 0.5 indicates a time series with long term switching between high and low values in adjacent pairs, meaning that a single high value will probably be followed by a low value and that the value after that will tend to be high, with this tendency to switch between high and low values lasting a long time into the future. The Hurst Exponent, *H*, is defined in terms of the asymptotic behaviour of the rescaled range as follows[17]:

$$E\left[\frac{R(n)}{S(n)}\right] = Cn^H \; as \; n \to \infty \qquad (1)$$

where, $\left[\frac{R(n)}{S(n)}\right]$ is the rescaled range, E[x]is the expected value, *n* is the time of the last observation (e.g. it corresponds to X_n in the input time series), *C* is a constant. Thus the Hurst Exponent is estimated by calculating the average rescaled range over multiple regions of the data. Then the mean value of rescaled range is equal to the number of regions raised to the Hurst Exponent, multiplied by a constant.

Hurst Exponent is approximated using the expression:

$$\text{Log (R/S)} = \log (\text{C}) + \text{H} * \log (\text{n}) \qquad (2)$$

Thus to estimate the Hurst Exponent, we plot (R/S) versus n in log-log axes. The slope of the regression line approximates the Hurst Exponent. In this context Hurst exponent is used as a characteristic parameter to describe long range correlation in protein sequences. We compute the Hurst exponents of each protein sequences, so that we can obtain the features of Hurst exponents in each sequence. The Hurst exponent describes the degree of self similarity of the protein sequences data set. The Hurst exponent of a data series with long range dependence is between 0.5 and 1. An increased Hurst exponent indicates an increase in the degree of self similarity and long range dependence.

ii) ENTROPY

Quantifying the amount of regularity for a time series is an essential task in understanding the behaviour of a system. One of the most popular regularity measurements for a time series is the sample entropy. Entropy is a measure of the average uncertainty of symbols or outcomes.

Given a random variable X with a set of possible symbols or outcomes = $[a_1, a_2,...,a_j]$, having probabilities $[p_1, p_2,.. , p_j]$, with $P(x = a_i) = p_i$, $p_i > 0$ and $\sum p(x)=1$, entropy is defined as

$$H(x) = \sum p(x) \log_2 p(x) \quad (3)$$

iii) STANDARD DEVIATION

Standard deviation is a widely used measure of the variability or dispersion. It shows how much variation there is from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

$$\sigma = \sqrt{\frac{1}{N}} \sum_{i=1}^{N} (x_i - \mu)^2 \quad (4)$$

2.2.3 Distance matrix

One of the important measures in evaluating the difference between sequences is to obtain the distance between them. In this work, since we are considering nine sequences for the similarity analysis the distances between them are represented in distance matrix. Commonly used measures for generating distance matrix are Euclidean distance, Hamming distance, Manhattan distance, Tanimoto coefficient and Jaccard index. Manhattan distance is adopted in the proposed method. We used Manhattan distance since the features are obtained as scalar values. Distance between a vector X=(X_1, X_2 etc.) and another Y= (Y_1, Y_2 etc.) is:

$$d = \sum |x_i - y_i| \quad (5)$$

2.2.4 Phylogenetic tree

Phylogenetic tree is a graph that indicates how a family of related sequences have been derived during evolution. The branches of the tree represent the evolutionary relationship between the sequences. The degree of similarity between the sequences is represented as the length of branches. The main aim of phylogenetic analysis is to find the length of the branches and to derive the evolutionary relationship. Neighbouring branches on a tree indicates sequences that are closely related. Phylogenetic analysis can be used to determine the sequences having equivalent functions.

3. Results and Discussions

In this work, nine CD4 (Cluster of differentiation 4) protein sequences are selected from the Uniprot protein database [15]. They have nearly same length.

3.1 Feature extractor outputs

Samples of the feature extractor outputs of each sequence are shown below. Following values are obtained by the feature extraction of the sequences which are mapped using hydropathy values.

TABLE 3
FEATURE VALUES

	Standard deviation	Hurst exponent	Entropy
P01730	3.1141	0.454552	3.7766
P16004	3.1074	0.440770	3.7629
P79185	3.0608	0.430249	3.7675
P79184	3.0711	0.434012	3.7838
P16003	3.0755	0.408890	3.7810
Q08340	3.0588	0.442800	3.7751
Q29037	3.0882	0.472924	3.7687
Q08338	3.0640	0.525790	3.7450
Q8HZT8	3.0965	0.490750	3.8233

3.2 Distance matrix

A matrix which includes the distance between various sequences is represented in the distance matrix. Each entries in the distance matrix indicates the evolutionary distance between each of the sequences in the dataset.

Sample distance matrix is shown in table 4.

TABLE 4
DISTANCE MATRIX

	P01730	P16004	P79185	P79184	P16003	Q08340	Q29037	Q08338	Q8HZT8
P01730	0	0.006	0.119	0.112	0.11	0.114	0.591	0.109	0.584
P16004	0.006	0	0.125	0.117	0.115	0.119	0.589	0.112	0.583
P79185	0.119	0.125	0	0.015	0.014	0.018	0.633	0.063	0.629
P79184	0.112	0.117	0.015	0	0.003	0.012	0.631	0.055	0.622
P16003	0.11	0.115	0.014	0.003	0	0.01	0.63	0.054	0.62
Q08340	0.114	0.119	0.018	0.012	0.01	0	0.629	0.059	0.623
Q29037	0.591	0.589	0.633	0.631	0.63	0.629	0	0.629	0.148
Q08338	0.109	0.112	0.063	0.055	0.054	0.059	0.059	0	0.614
Q8HZT8	0.584	0.583	0.629	0.622	0.62	0.623	0.623	0.614	0

3.3 Phylogenetic tree

In this work phylogenetic trees are drawn from the distance matrices by using functions in the PHYLIP software package. PHYLIP, the Phylogeny Inference Package, is a package of programs for inferring phylogenies (evolutionary trees). It can compute distances between trees, draw trees, edit trees and compute distance matrices. The functions *neighbour*, *fitch* and *kitsch* are the tree drawing functions, which draws a phylogenetic tree when the distance matrix is given as input. The trees generated from these programs are in the *Newick* format. It is a method of representing trees with branch lengths using parentheses and commas. The methods are evaluated by looking at how closely the phylogenetic trees constructed using the values obtained by that method correspond with the available reference phylogenetic tree. Sample phylogenetic tree generated using the 25 amino acid indices and the feature standard deviation is shown in Fig. 1.

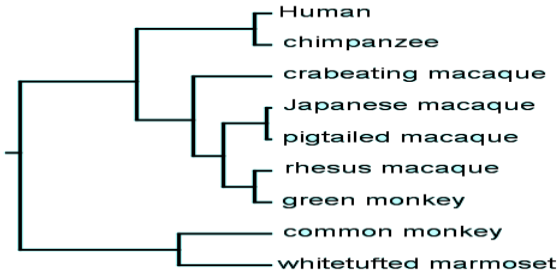

Fig. 1. Phylogenetic tree I

3.4 Evaluation

The phylogenetic trees drawn using different combinations of mapping and feature extractors were then compared with a reference tree generated by CLUSTAL software which uses conventional sequence alignment. Tree generated using CLUSTAL software is shown in Fig. 2.

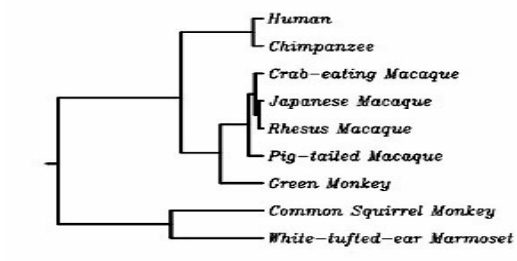

Fig.2. CLUSTAL Tree

The *treedist* function of the *PHYLIP PACKAGE* is used to compare the phylogenetic trees generated using the proposed method. The trees were compared using the *branch score distance* option of the *treedist* program in PHYLIP package. The input tree file of the treedist program should be in the Newick Format. The smaller the distance value closer will be the tree to the reference. Screenshots of the result of comparison with generated trees and reference tree is shown in Fig. 3.

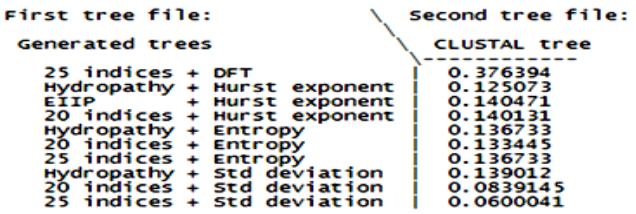

Fig. 3. Comparison with CLUSTAL reference tree

From the figure we can see that the tree generated using combination of measures 25 amino acid indices and standard deviation is most similar to the clustal reference tree with the least distance value of 0.06.
Due to the biological significance of hydropathy and the other 25 amino acid indices used, we can infer that these are relevant mapping techniques. Entropy and standard deviation quantifies the regularity and dispersion of the time series respectively. From the results it is clear that these are excellent features in comparison with features mentioned in literature[12]. Distance from the tree drawn using Fourier transform to the reference tree is more compared to that with the trees drawn using the proposed method[12]. Hence we can infer that these are relevant feature extractors.

3.5 Comparison of time complexity

The time of execution of different algorithms used in this method as well as the time for generation of reference tree is computed. Screenshots of reference tree generation time as well as time of tree generation using proposed method is given in Fig. 4 and Fig. 5 .

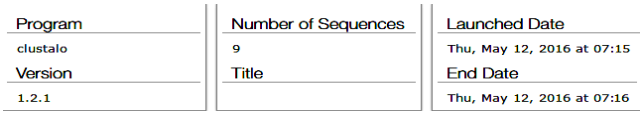

Fig. 4. Screenshot of CLUSTAL reference tree generation time:
Time for execution=60 seconds

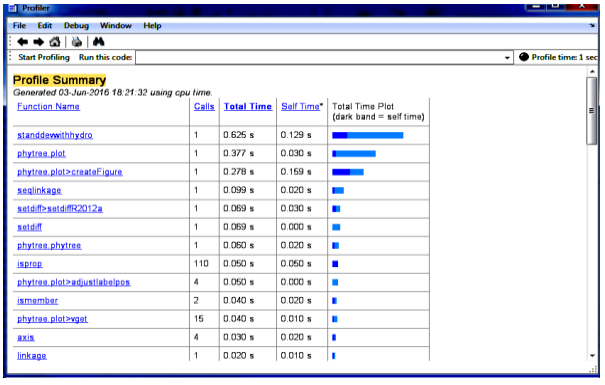

Fig. 5. Screenshot of tree generation time using the proposed method:
Time for execution=1 second

From the above results it is clear that the time for execution in *CLUSTAL SOFTWARE* is around 60 seconds, whereas using the proposed method we obtain phylogenetic trees which are closer to the reference tree within around 1 second. This is because the former use protein sequence alignment to obtain the sequence similarity which is a time consuming process.

Conclusion

In this work we have done protein sequence analysis using signal processing techniques. Here we used Hurst exponent, Entropy and Standard deviation to construct the distance matrix. In addition to that multiple amino acid indices were used to encode protein sequences into numerical sequences. A case study was done using nine CD4 protein sequences derived from the UniProt online protein sequence database, inorder to present its applicability.
The results were compared with freely available tool CLUSTAL software which uses multiple sequence alignment for protein sequence similarity analysis. From the comparison it is clear that the proposed method generates phylogenetic trees which are similar to the reference tree. Also the time required for the tree generation using proposed method is much less than the time for reference tree generation.

References

[1] D. J. Rigden, :From protein structure to function in bioinformatics NEWYORK,NY: Springer-verlag 2009.

[2] Y. Shibberu and A. Holder, “A spectral approach to protein structure alignment,” IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 8, no.4, pp.867-875, Jul. 2011.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, “Gapped blast and psi-blast: A new generation of protein database search programs,: Nucleic Acids Res., vol. 25, no. 17, pp. 3389-3402,1997.

[4] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, “Hidden Markov models in computational biology: Applications to protein modeling,” J. Mol. Biol., vol. 235, pp. 1501-1531, 1994.

[5] D. S. Huang, X.-M. Zhao, G.-B. Huang, and Y.-M. Cheung, ”Classifying protein sequences using hydropathy blocks,” Pattern Recog., vol. 39, no. 12, pp. 2293-2300, 2006.

[6] P. Lio, “Wavelets in bioinformatics and computational biology: State of art and perspectives,” Bioinformatics, vol. 19. No. 1, pp. 2-9, 2003.

[7] Lina Yang, Yuan Yan Tang, Yang Lu, and Huiwu Luo, “A Fractal dimension and wavelet transform based method for protein sequence similarity analysis,” IEEE/ACM Trans. On comput. Biol. And bioinformatics, vol. 12, no. 2, March/April 2015.

[8] Lina Yang, Yuan Yan Tang, Yulong Wang, Huiwu Luo, Jianjia Pan, Haoliang Yuan, Xianwei Zheng, Chunli Li and Ting Shu, ”Similarity analysis based on sparse representation for protein sequence comparison,” 2015 IEEE.

[9] Jie Su and Junpeng Bao, ”A wavelet transform based protein sequence similarity model”, Applied Mathematics & Information Sciences;May2013, Vol. 7 Issue 3, p1103

[10] Anu Sabarish.R and Tessamma Thomas, “A frequency domain approach to protein sequence similarity analysis and functional classification,” signal and image processing: An international journal, vol. 2, no. 1, March 2011.

[11] Shao-Ming Zhu, Zu-Guo Yu, Vo Anh, Sheng-Yuan Yang, “Analysing the similarity of proteins based on a new approach to empirical mode decomposition,” 2010 IEEE.

[12] Charalambos Chrysostomou and Huseyin Seker, “Construction of Protein Dendrograms based on amino acid indices and discrete fourier transform,” 2014 IEEE.

[13] C. H. de Trad, Q. Fang, and I. Cosic, “Protein sequence comparison based on the wavelet transform approach,” protein Eng., vol. 15, no. 3, pp. 193-203,2002.

[14] H. Chen, F. Gu, and Z. Huang, “A compound method of protein secondary structure prediction and its implementation,” in Proc. First IEEE Int. Multi-symp. Comput, Comput. Sci., vol. 1, 2006, pp. 104-109.

[15] A. Bernard, Leucocyte typing: human leucocyte differentiation antigens detected by monoclonal antibodies: specification, classification, nomenclature, Springer, 1984.

[16] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. katayama, and M. Kanehisa, “Aaindex: amino acid index database, progress report 2008,” Nucleic acids research, vol. 36, no. suppl 1, p. D202, 2008.

[17] Vrinda V. Nair, Achuthsankar S. Nair, Anita mallya, Bhavya sebastian , “Hurst CGR (HCGR) – A Novel Feature Extraction Method from Chaos Game Representation of Genomes,” Conference Paper *in* Communications in Computer and Information Science 190:302-309 · July 2011.