Construction of Protein Dendrograms Based on Amino Acid Indices and Discrete Fourier Transform

Charalambos Chrysostomou[1]* and Huseyin Seker[2]

Abstract— **From the literature, existing methods use pairwise percent identity to identify the percentage of similarity between two protein sequences, in order to create a dendrogram. As this is a parametric method of measuring the similarities between proteins, and different parameter may yield different results, this method does not guarantee that the global optimal similarity values will be found. As protein dendrogram construction is used in other areas, such as multiple protein sequence alignments, it is very important that the most related protein sequences to be identified and align first. Furthermore, by using the pairwise percent identity of the protein sequences to construct the dendrograms, the physical characteristics of protein sequences and amino acids are not considered.**

In this paper, a new method was proposed for constructing protein sequence dendrograms. For this method, Discrete Fourier Transform, was used to construct the distance matrix in combination with the multiple amino acid indices that were used to encode protein sequences into numerical sequences. In order to show the applicability and robustness of the proposed method, a case study was presented by using nine Cluster of Differentiation 4 protein sequences extracted from the UniProt online database.

I. INTRODUCTION

A dendrogram is a branching diagram used to visualise the arrangement of the clusters based on the degree of similarity of certain characteristics. In recent years, the construction and use of dendrogram in protein sequences has become an important tool and has successfully been applied in various areas of research such as multiple protein sequence alignment [1].

In order to create a dendrogram, existing methods use pairwise percent identity to identify the percentage of similarity between two protein sequences. However, this method of measuring the similarities between proteins does not guarantee that the global optimal similarity values will be found [2]. As protein dendrogram construction is used in other areas, such as multiple protein sequence alignments, it is very important that the most related protein sequences be identified and align first. By using only the pairwise percent identity of the protein sequences the physical characteristics of protein sequences and amino acids are not considered. By combining different physical characteristics of protein sequences hidden connections between protein sequences may be identified that do not rely on the pairwise percent identity of the protein sequences [3]. Additionally, current methods depend on the nature of the proteins and the appropriate parameters to produce accurate results. Two of the main issues in selecting the appropriate parameters for the construction of protein dendrograms are: a) widely debated which are the optimal parameters for a specific class of proteins and b) In many occasions the optimal results depend on the judgement of the researcher contacting the experiment.

Previous work had shown that by combining different amino acid indices, which represent different physical characteristics of protein sequences, and signal-processing techniques the protein distance matrix can be constructed [4]. These distances between protein sequences can be used to assembly the protein dendrogram. In this paper, a novel approach is proposed that effectively utilises of signal-processing techniques and multiple amino acid indices to create protein dendrograms.

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

II. METHODS AND MATERIALS

A. Amino Acid Indices

In order to apply signal processing techniques, the protein sequences need to be encoded to numerical sequences. By using amino acid indices, each amino acid of the protein can be converted into a specific number. In the literature, more than 500 amino acid indices exist [5], where each index represents a unique biological protein feature. As Table I shows, 25 amino acid indices were selected for this study.

Specifically, these amino acid indices represent commonly used and widely accepted physical characteristics of the amino acids, like [6], [7], [8], [9] of the amino acids, like hydrophobicity [10], [11], [12], [13], molecular weight [10], size [14] and volume [15].

B. Construction of Dendrogram

The first step in constructing a dendrogram is to calculate the distance matrix between all the protein sequences. The following steps need to be completed in order to calculate the distance matrix, based on the Discrete Fourier Transform (DFT) and multiple amino acid indices.

- By using the 25 amino acid indices as described in Table I, each protein sequence can be converted into numerical sequence. The sequences are converted by replacing each corresponding alphabetical letter for each protein

[1] Department of Genetics, University of Leicester, University Road Leicester, LE1 7RH, United Kingdom

[2] Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK

cc390@le.ac.uk, hseker@dmu.ac.uk

*Corresponding Author

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

TABLE I

AMINO ACID INDICES USED FOR THE DENDROGRAM CONSTRUCTION

ID	Name	Description	Reference
1	ZIMJ680102	Bulkiness	[16]
2	ZIMJ680104	Isoelectric point	[16]
3	HUTJ700102	Absolute entropy	[17]
4	DAWD720101	Size	[14]
5	GRAR740102	Polarity	[15]
6	GRAR740103	Volume	[15]
7	FASG760101	Molecular weight	[10]
8	FASG760102	Melting point	[10]
9	FASG890101	Hydrophobicity index	[10]
10	ZHOH040101	The stability scale from the knowledge-based atom-atom potential	[18]
11	OOBM770103	Long range non-bonded energy per atom	[19]
12	MANP780101	Average surrounding hydrophobicity	[11]
13	WOLR790101	Hydrophobicity index	[12]
14	FAUJ880101	Hydration potential	[20]
15	FAUJ880102	Smoothed upsilon steric parameter	[21]
16	ARGP820101	Hydrophobicity index	[13]
17	VELV850101	Electron-ion interaction potential	[22]
18	FAUJ880111	Positive charge	[21]
19	FAUJ880112	Negative charge	[21]
20	FAUJ880109	Number of hydrogen bond donors	[21]
21	KYTJ820101	Hydropathy index	[23]
22	BHAR880101	Average flexibility indices	[24]
23	Proscale_4	Recognition factors	[25]
24	Nl	Long-range contacts	[26]
25	Rk	Relative connectivity	[27]

sequence with the corresponding number for each of the 25 amino acid indices.

- Each converted numerical sequence needs to be zero-padded to the length of the longest protein sequence. This is an important step for comparing sequences with different lengths.
- The absolute frequency spectra is calculated for each of the converted numerical sequence for each protein of the dataset. DFT is described in Equations 1 and 2.
- The 25 absolute frequency spectra calculated in the previous step is combined into one vector for each protein sequence.
- The distance matrix can be calculated for all protein sequences using the correlation distance as described in Equation 3.

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)nm} \quad n = 0, 1, ..., N-1 \quad (1)$$

where $x(m)$ is the mth member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. From this point forward, only the first half of the series $(N/2)$ will be considered as the DFT coefficients consists of two mirror parts. The absolute frequency spectrum can be determined from the following formula:

$$S_{(n)} = X(n)X^*(n) = |X(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \quad (2)$$

where $S_{(n)}$ is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X^*(n)$ are the complex conjugate.

The correlation distance can be calculated as follows

$$D(X,Y) = 1 - \frac{(X - \bar{X}) \cdot (Y - \bar{Y})}{\parallel (X - \bar{X}) \parallel_2 \parallel (Y - \bar{Y}) \parallel_2} \quad (3)$$

where $\bar{X}$ and $\bar{Y}$ represent the mean values of vectors X and Y, respectively.

After calculating the distance matrix, the dendrogram can be constructed. The method used to build the guide tree is the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). Moreover, UPGMA is a hierarchical clustering method created by Sokal and Michener [28] and used in bioinformatics for the creation of phenetic trees. A detailed description of the procedure that the UPGMA algorithm follows to create a rooted dendrogram is given below:

1) Calculate the distance matrix D (Eq. 3) for all the protein sequences.
2) Initialise the dendrogram by creating one leaf node for each protein sequence in the dataset.
3) Find X and Y protein sequences that have the smallest distance $D_{X,Y}$ between them.
4) Create a new node XY.
5) Connect X and Y, and give the two branches connecting X and Y to XY with the length $D_{X,Y}/2$.
6) Compute the distance from the new node to all the remaining nodes of the dendrogram as a weighted average using Equation 4.

$$D_{XY,Z} = (\frac{n_X}{n_X + n_Y})D_{X,Z} + (\frac{n_Y}{n_X + n_Y})D_{Y,Z} \quad (4)$$

7) Append the distances calculated above in D matrix and delete the distances that correspond to the nodes X and Y.

8) Repeat Steps 3-7 until all the protein sequences are covered.

C. Case Study: Dendrogram Construction of Cluster of Differentiation 4 Proteins

The Cluster of Differentiation 4 (CD4) [29] 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. Human immunodeficiency virus (HIV) uses CD4 receptor by biding to gp120 to infect a host T-cell, which in recent years has become subject to an intense research in the pursuit of a cure.

In this study, nine CD4 protein sequences, as listed in Table II, will be used. These protein sequences were collected from UniProt [30].

TABLE II

CD4 PROTEINS

ID	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

III. RESULTS AND DISCUSSIONS

By using the algorithm as described in this paper, protein dendrograms can be created, using any given set of amino acid indices, each one of which represents a unique biological feature of protein sequences. The results are based on the combination of 25 widely accepted amino acid indices (Table I), which produced the best results, according to the biological relationships between proteins, for constructing dendrograms for protein sequences. The produced dendrogram can be found in Figure. 1.

For comparison purposes, ClustalW [31] was used being a freely available and popular tool for constructing protein dendrograms and protein sequence alignments. Table III represents the pairwise percent identity of the protein sequences used, that derived from ClustalW. As ClustalW is a parametric process the default parameters are used. The results from ClustalW are shown in Figure. 2.

The results suggest that the proposed method for constructing dendrograms can match the results produced from ClustalW with the suggested parameters. For ClustalW, depending on the nature of the protein sequences used different parameters needs to be used, which can influence the results produced. For the case study, the proposed method produced identical results, but no parameters need to be optimised. Additionally, the dendrogram produced by the proposed method can be linked to the physical characteristics of the amino acids / protein sequence by the used set of amino acid indices.

TABLE III

PAIRWISE PERCENT IDENTITY OF CD4 PROTEINS

	2	3	4	5	6	7	8	9
1	98	91	91	91	91	80	90	81
2		91	91	91	90	80	90	80
3			99	99	98	79	95	80
4				99	98	79	95	80
5					98	79	95	80
6						79	94	79
7							79	90
8								80

TABLE IV

DISTANCE MATRIX OF CD4 PROTEINS

	2	3	4	5	6	7	8	9
1	0.006	0.119	0.112	0.11	0.114	0.591	0.109	0.584
2		0.125	0.117	0.115	0.119	0.589	0.112	0.583
3			0.015	0.014	0.018	0.633	0.063	0.629
4				0.003	0.012	0.631	0.055	0.622
5					0.01	0.63	0.054	0.62
6						0.629	0.059	0.623
7							0.629	0.148
8								0.614

IV. CONCLUSIONS

In this paper, a new method was proposed for constructing protein sequence dendrograms. For this method Discrete Fourier Transform was used to construct the distance matrix in combination with the multiple amino acid indices that were used to encode protein sequences into numerical sequences.

In order to show the applicability and robustness of the proposed method, a case study was successfully presented by using nine CD4 protein sequences extracted from the UniProt online database. The results produced were compared with a popular and freely available tool for constructing protein dendrograms and performing protein sequence alignments.

Further research needs to be carried out to compare the results between ClustalW and the proposed method. As future work, larger dataset of CD4 protein sequences needs to be used which may include more diverse species. Finally, as protein dendrograms can be created as guide trees in protein sequence alignments, further research is required to investigate how on how the proposed method can affect these alignments.

REFERENCES

[1] D. W. Mount, *Bioinformatics: sequence and genome analysis.* CSHL press, 2004.

[2] E. Althaus, A. Caprara, H.-P. Lenhof, and K. Reinert, "Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics." *Bioinformatics*, vol. 18 Suppl 2, pp. S4–S16, 2002.

[3] Z. Wu, X. Xiao, and K. Chou, "2d-mh: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids," *Journal of theoretical biology*, vol. 267, no. 1, p. 29, 2010.

[4] C. Chrysostomou and H. Seker, "Construction of protein distance matrix based on amino acid indices and discrete fourier transform," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.* IEEE, 2013, pp. 4066–4069.

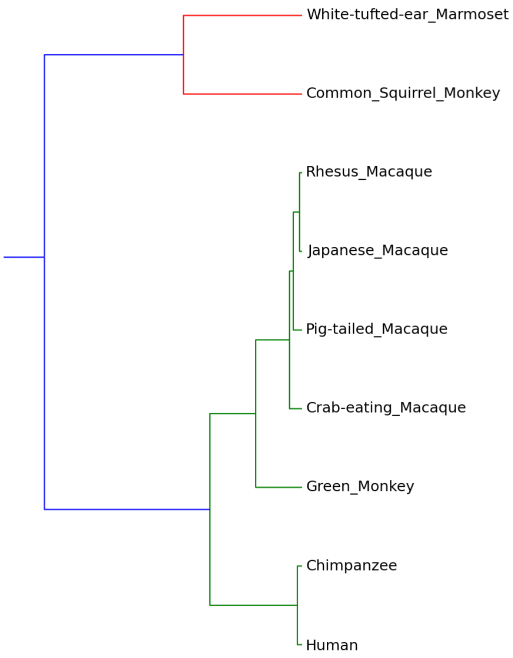

Fig. 1. Dendrogram Constructed using the Proposed Method

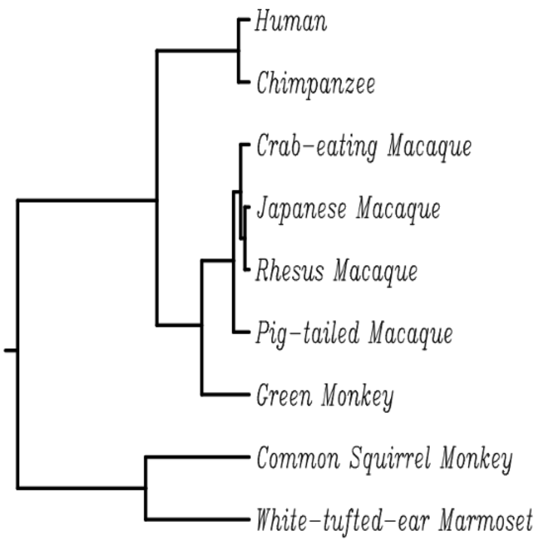

Fig. 2. ClustalW Dendrogram

[5] 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.

[6] X. Xia and W. H. Li, "What amino acid properties affect protein evolution?" *J Mol Evol*, vol. 47, no. 5, pp. 557–564, Nov 1998.

[7] S. Woolley, J. Johnson, M. J. Smith, K. A. Crandall, and D. A. McClellan, "Treesaap: selection on amino acid properties using phylogenetic trees." *Bioinformatics*, vol. 19, no. 5, pp. 671–672, Mar 2003.

[8] G. Singh, *Chemistry of amino-acids and proteins*. Discovery Publishing House, 2007.

[9] A. Hughes, *Amino Acids, Peptides and Proteins in Organic Chemistry: Building Blocks, Catalysis and Coupling Chemistry*. Wiley-VCH, 2011, vol. 3.

[10] G. Fasman, *Practical handbook of biochemistry and molecular biology*. CRC, 1989.

[11] P. Manavalan and P. Ponnuswamy, "Hydrophobic character of amino acid residues in globular proteins," 1978.

[12] R. Wolfenden, P. Cullis, and C. Southgate, "Water, protein folding, and the genetic code," *Science*, vol. 206, no. 4418, p. 575, 1979.

[13] P. ARGOS, J. Rao, and P. HARGRAVE, "Structural prediction of membrane-bound proteins," *European Journal of Biochemistry*, vol. 128, no. 2-3, pp. 565–575, 1982.

[14] O. Mayo and D. Brock, *The biochemical genetics of man*. Cambridge Univ Press, 1972.

[15] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, p. 862, 1974.

[16] J. ZimmermanNaomi and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of theoretical biology*, vol. 21, no. 2, pp. 170–201, 1968.

[17] L. Acid, D. Citrulline, and D. HCI, "Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds," *Handbook of biochemistry and molecular biology*, vol. 1, no. 154.33, p. 109, 1984.

[18] H. Zhou and Y. Zhou, "Quantifying the effect of burial of amino acid residues on protein stability," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 54, no. 2, pp. 315–322, 2004.

[19] M. Oobatake and T. Ooi, "An analysis of non-bonded energy of proteins," *Journal of Theoretical Biology*, vol. 67, no. 3, pp. 567–584, 1977.

[20] R. Wolfenden, L. Andersson, P. Cullis, and C. Southgate, "Affinities of amino acid side chains for solvent water," *Biochemistry*, vol. 20, no. 4, pp. 849–855, 1981.

[21] J. FAUCHÈRE, M. Charton, L. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *International journal of peptide and protein research*, vol. 32, no. 4, pp. 269–278, 1988.

[22] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. LalovicC, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?" *IEEE Transaction on Biomedical Engineering*, vol. 32, no. 5, pp. 337–341, 1985.

[23] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of molecular biology*, vol. 157, no. 1, pp. 105–132, 1982.

[24] R. Bhaskaran and P. Ponnuswamy, "Positional flexibilities of amino acid residues in globular proteins," *International Journal of Peptide and Protein Research*, vol. 32, no. 4, pp. 241–255, 1988.

[25] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. Wilkins, R. Appel, and A. Bairoch, "Protein identification and analysis tools on the expasy server," *The proteomics protocols handbook*, pp. 571–607, 2005.

[26] L. Fernández, J. Caballero, J. Abreu, and M. Fernández, "Amino acid sequence autocorrelation vectors and bayesian-regularized genetic neural networks for modeling protein conformational stability: Gene v protein mutants," *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 834–852, 2007.

[27] J. Huang, S. Kawashima, and M. Kanehisa, "New amino acid indices based on residue network topology," *Genome Informatics*, vol. 18, pp. 152–161, 2007.

[28] R. Sokal and C. Michener, "A statistical method for evaluating systematic relationships," *Univ. Kans. Sci. Bull.*, vol. 38, pp. 1409–1438, 1958.

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

[30] A. Bairoch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.*, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D154–D159, 2005.

[31] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.