35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

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

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

Abstract— **Protein distance matrix is widely used in various protein sequence analyses, and mainly obtained by using pairwise sequence alignment scores or protein sequence homology, which fail to take into consideration of individual physical characteristics of protein sequences and amino acids, or a combination of these features. In this paper, a new method is therefore proposed for constructing protein distance matrix based on natural amino acid indices in combination with Discrete Fourier Transform (DFT).**

For the proposed method, protein distance matrices can be generated using any given set of amino acid indices, each one of which represents a unique biological feature of protein sequences. In this study, the results are based on the combination of 25 widely accepted amino acid indices, which produced the best results, according to the biological relationships between proteins. As a case study 26 Cluster of Differentiation 4 (CD4) protein sequences were used in order to construct a distance matrix based on the proposed method.

The results show that the pairwise relationship between CD4 protein sequences remain the same in comparison with their pairwise percent identity. For another group of protein sequences the pairwise relationship between CD4 protein sequences dramatically changed with the proposed method in comparison to the pairwise percent identity. The proposed distance matrix has been shown to have a positive impact on these case studies and therefore is expected to be useful in several fields such as multiple protein sequence alignment and phylogenetic analysis, where an accurate distance matrix based on natural generalized protein properties plays an important role.

Index Terms— **Amino Acid Indices, Cluster of Differentiation 4 (CD4), Distance Matrix, Discrete Fourier Transform (DFT), Protein Sequences**

I. Introduction

In bioinformatics, a protein distance matrix [1] is a matrix containing the pairwise distances of a set of given protein sequences. This matrix will have a size of NxN, where N is the number of protein sequences.

In recent years, the use of distance matrices in protein sequences has become an important tool in bioinformatics and has successfully been applied in various fields such as multiple protein sequence alignment [1] and phylogenetic analysis [2].

However, existing methods used to create protein distance matrices like pairwise percent identity do not guarantee that the global optimal protein similarity values will be found [3]. Commonly, this issue will raise additional problems in other areas like phylogenetic analysis and multiple protein sequence alignments when the relationships of protein sequences in the phylogenetic tree are not in the correct order or if errors occur in the early pairwise alignments and by adding less related proteins, these errors will increase. Furthermore, existing methods do not take into consideration of individual physical characteristics of protein sequences and amino acids, or a combination of these features when the distance matrix is constructed. By taking into consideration of a combination of proteins physical characteristics, hidden associations between protein sequences may be revealed that do not necessarily rely on the homology or pairwise percent identity of the protein sequences [4]. In order to overcome these problems a novel approach that utilises signal-processing techniques and multiple amino acid indices is created and described in this paper.

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 encode protein sequences to numerical sequences, amino acid indices need to be selected. In the literature, more than 500 amino acid indices exist [5], each one representing a unique biological protein feature. For this study, 25 amino acid indices were selected as shown in Table I. These amino acid indices represent general and widely accepted features [6] of the amino acids, like size [7], volume [8], molecular weight [9] and hydrophobicity [9]–[11]. The complete list of the amino acid indices used for this analysis is presented in Table I.

B. Discrete Fourier Transform

The Discrete Fourier Transform (DFT) is defined as follows

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

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. The following formula determines the absolute frequency spectrum

[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

charalambos.chrysostomou@gmail.com, hseker@dmu.ac.uk

*Corresponding Author

978-1-4577-0216-7/13/$26.00

TABLE I

AMINO ACID INDICES USED FOR THE ALIGNMENT

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

$$S_a(n) = X(n)X*(n) = |X(n)|^2, \quad n = 1, 2, ..., N/2 \quad (2)$$

where S_a 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.

C. Construction of Protein Distance Matrix

The following steps need to be completed in order to calculate the distance matrix.

- Each protein sequence in the dataset is converted into 25 numerical sequences using the amino acid indices shown in Table I.
- As different protein sequences are likely to have distinct lengths, each numerical sequence is zero-padded to match the length of the longest protein sequence in the dataset [25]. This step is essential for comparing multiple proteins with different lengths.
- By using DFT as described in Equations 1 and 2, the absolute frequency spectra is calculated for each of the 25 numerical sequences for all the protein sequences.
- For each protein sequence, the 25 absolute frequency spectra is combined into one vector. The correlation distance matrix for all the protein sequences can be calculated by using the correlation distance as given in Equation 3.

$$D(X, Y) = 1 - \frac{(X - \bar{X}) \cdot (Y - \bar{Y})}{\| (X - \bar{X}) \|_2 \, \| (Y - \bar{Y}) \|_2} \quad (3)$$

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

D. Case Study: Construction of Protein Distance Matrix for Cluster of Differentiation 4 Proteins

The Cluster of Differentiation 4 (CD4) [26] is a glycoprotein and was discovered in late 1970. CD4 is expressed on the surface of T helper cells, monocytes, macrophages, and dendritic cells. The main function of CD4 is to act as a co-receptor that supports the T-cell receptor with an antigen-presenting cell. In recent years, CD4 has become subject to an intense research towards finding a potential cure for Human immunodeficiency virus (HIV). Protein sequence alignment is important as it can identify regions of similarity that can be considered significant in regards to the functional, structural or evolutionary relationships between the protein sequences. HIV-1 uses CD4 to infect a host T-cell and accomplishes this by binding gp120, a known viral envelope protein with CD4. In this study, 26 CD4 protein sequences, as listed in Table II, will be used in order to construct a distance matrix based on the proposed method. These protein sequences were collected from UniProt [27] and belongs to various animals as listed in Table II.

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	P06332	Mouse	457
10	P46630	Rabbit	459
11	P05540	Rat	457
12	Q6R3N3	Pig	417
13	A7YY52	Bovine	395
14	NP_001123374	Sheep	455
15	ACG76115	Goat	455
16	Q8HZT8	White-tufted-ear Marmoset	457
17	P33705	Dog	463
18	AAB24450	Cat	474
19	Q9XS78	Beluga Whale	455
20	Q71QE2	Bottle-nosed Dolphin	455
21	NP_001092760	Gray short-tailed Opossum	461
22	ABR22561	Tammar Wallaby	464
23	AAS67020	Chicken	487
24	CAP04927	Turkey	487
25	B8YEL3	Domestic Duck	482
26	AAW63065	Muscovy Duck	482

III. RESULTS AND DISCUSSIONS

By using the algorithm as described in this paper, protein distance matrices can be generated by using any given set of amino acid indices, each one of which represents a unique biological feature of protein sequences. In this paper, the results are based on the combination of 25 widely accepted amino acid indices, which produced the best results, according to the biological relationships between proteins.

Table III represents the pairwise percent identity of the CD4 protein sequences and Table IV represents the proposed DFT based protein distance matrix. A high-quality distance matrix gives lower value to the protein sequences with high biological relationship and a higher value to protein sequences with low biological relationship. For percent

TABLE III

PAIRWISE PERCENT IDENTITY OF CD4 PROTEINS

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	98	60	90	61	91	91	91	91	54	61	54	80	60	23	23	16	18	81	24	60	57	60	61	22	60
2		60	90	61	91	91	91	90	54	62	53	80	61	23	21	18	18	80	24	60	57	60	61	22	59
3			59	63	59	60	59	59	49	58	46	56	63	18	20	14	18	56	21	61	56	62	71	20	61
4				60	95	95	95	94	54	61	53	79	60	23	24	20	18	80	23	59	57	59	62	21	58
5					61	60	61	60	50	60	50	59	97	19	19	17	18	59	23	73	67	73	63	24	72
6						99	99	98	53	61	53	79	60	24	23	16	17	80	22	60	58	60	62	21	60
7							99	98	54	61	53	79	60	24	23	16	18	80	22	60	58	60	62	22	59
8								98	53	61	53	79	60	24	23	16	18	80	22	60	58	60	62	21	59
9									54	60	53	79	60	24	23	15	18	79	21	60	58	60	62	21	59
10										52	74	54	49	22	22	16	14	53	23	49	46	50	51	23	50
11											52	59	60	22	22	19	17	59	24	59	55	60	59	24	60
12												53	49	18	21	17	11	53	21	49	46	50	50	21	49
13													58	23	24	15	17	90	22	58	55	57	59	22	57
14														19	23	17	18	58	21	72	68	73	64	23	72
15															91	15	17	22	62	24	23	24	22	62	24
16																14	18	23	64	24	25	25	22	63	25
17																	47	15	14	18	18	14	16	16	14
18																		17	18	17	20	18	13	17	19
19																			21	58	56	57	59	22	56
20																				23	22	23	21	83	23
21																					63	68	60	22	68
22																						84	58	23	83
23																							62	23	97
24																								22	62
25																									22

TABLE IV

DISTANCE MATRIX OF CD4 PROTEINS

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	0.01	0.78	0.11	0.81	0.12	0.11	0.11	0.12	0.81	0.8	0.8	0.59	0.8	0.9	0.9	0.95	0.98	0.58	0.94	0.76	0.75	0.79	0.76	0.92	0.79
2		0.78	0.11	0.81	0.13	0.12	0.12	0.12	0.81	0.8	0.8	0.59	0.8	0.9	0.9	0.95	0.98	0.58	0.94	0.76	0.76	0.8	0.76	0.92	0.8
3			0.76	0.78	0.76	0.74	0.75	0.75	0.75	0.67	0.76	0.88	0.79	1.01	1	0.92	0.96	0.87	0.94	0.78	0.75	0.8	0.49	0.95	0.8
4				0.81	0.07	0.06	0.06	0.06	0.82	0.8	0.81	0.63	0.8	0.93	0.93	0.96	0.98	0.61	0.93	0.77	0.75	0.8	0.74	0.92	0.81
5					0.81	0.8	0.8	0.81	0.8	0.73	0.84	0.77	0.03	0.96	0.95	0.97	0.97	0.78	1.02	0.45	0.4	0.35	0.81	1.01	0.35
6						0.02	0.01	0.02	0.84	0.82	0.83	0.63	0.79	0.95	0.95	0.95	0.98	0.63	0.94	0.77	0.75	0.79	0.74	0.93	0.8
7							0	0.01	0.83	0.81	0.82	0.63	0.79	0.94	0.94	0.95	0.98	0.62	0.94	0.77	0.74	0.78	0.73	0.93	0.79
8								0.01	0.83	0.81	0.82	0.63	0.79	0.95	0.95	0.96	0.98	0.62	0.94	0.76	0.74	0.78	0.73	0.93	0.79
9									0.83	0.81	0.82	0.63	0.8	0.95	0.94	0.95	0.98	0.62	0.94	0.76	0.74	0.78	0.73	0.93	0.79
10										0.68	0.49	0.83	0.79	0.95	0.95	0.99	0.94	0.81	0.99	0.78	0.8	0.78	0.71	1.01	0.78
11											0.65	0.73	0.73	0.96	0.94	0.92	0.88	0.7	0.98	0.74	0.79	0.8	0.71	0.96	0.79
12												0.8	0.84	0.92	0.92	0.94	0.91	0.8	0.97	0.82	0.89	0.85	0.71	0.99	0.85
13													0.77	0.93	0.92	0.93	0.94	0.15	1	0.81	0.79	0.78	0.85	1.01	0.79
14														0.96	0.95	0.97	0.96	0.78	1.02	0.44	0.38	0.34	0.81	1.02	0.34
15															0.12	0.94	0.96	0.92	0.76	0.96	0.95	0.95	1.02	0.76	0.95
16																0.95	0.94	0.91	0.76	0.94	0.93	0.94	1.01	0.76	0.93
17																	0.81	0.91	0.97	0.97	0.97	0.99	0.99	0.96	1
18																		0.93	0.96	1	0.94	0.95	0.97	0.99	0.96
19																			1	0.79	0.79	0.79	0.86	1	0.8
20																				1.03	0.98	1.01	0.96	0.23	1.01
21																					0.48	0.48	0.76	1.01	0.46
22																						0.18	0.76	0.98	0.2
23																							0.79	1.01	0.04
24																								0.97	0.79
25																									1.01

identity, a high percentage will indicate that two protein sequences have high similarity in their amino acid arrangements while a low percentage will indicate low similarity.

As the results indicate in Tables III and IV by using the CD4 protein sequences, two important observations can be made in relation to the proposed distance matrix and the pairwise percent identity. The first observation is that pairwise relationship for a group of protein sequences remained the same between percent identity and the proposed distance matrix. An example can be seen in

- Sheep - Rhesus Macaque, with 97% and 0.03,
- Rabbit - Pig-tailed Macaque, 98% and 0.02,
- Bottle-nosed Dolphin - Domestic Duck, 83% and 0.23,
- Chicken - Muscovy Duck, 97% and 0.04,
- Dog - Crab-eating Macaque, 14% and 0.92,
- Dog - White-tufted-ear Marmoset, 14% and 0.95

for percent identity and the proposed distance, respectively.

The second observation is that pairwise relationship for a group of protein sequences has significant difference between percent identity and the proposed distance matrix. An example can be seen in

- Beluga Whale - Turkey, 59% and 0.86,
- Beluga Whale - Pig-tailed Macaque, 80% and 0.63,
- Beluga Whale - Bottle-nosed Dolphin, 21% and 1.0,
- Pig-tailed Macaque - Sheep, 60% and 0.79,
- Pig-tailed Macaque - Bottle-nosed Dolphin, 80% and 0.63
- Turkey - Rhesus Macaque, 63% and 0.81,

for percent identity and the proposed distance, respectively.

As the results show, the relationship between specific animals against the remaining animals under investigation like Beluga Whale, Pig-tailed Macaque and Turkey significantly changed with the proposed distance matrix. Further investigation needs to be conducted in order to determine the

effects of these differences in various application areas like multiple sequence alignment.

IV. Conclusions

Protein distance matrix is widely used in various protein sequence analyses, but depends generally on pairwise sequence alignment scores or protein sequence homology. However, they don't consider individual and natural physical characteristics of protein sequences and amino acids, or a combination of these features. In this paper, a new method is therefore proposed for constructing protein distance matrix based on natural amino acid indices in combination with DFT. As the results show for the proposed method for a group of protein sequences the pairwise relationship between CD4 protein sequences remain the same as pairwise percent identity. For another group of protein sequences the pairwise relationship between CD4 protein sequences dramatically changed with the proposed method in comparison to the pairwise percent identity. The proposed distance matrix will have an impact on several fields it is used in, such as multiple protein sequence alignment [1] and phylogenetic analysis [2], where an accurate distance matrix plays an important role.

The distance matrix is important as it can be used to construct a dendrogram that will act as a guide for multiple sequence alignments (MSA) in which the global alignment is estimated by a series of pairwise alignments. For MSA, for a reliable global alignment to be performed, the closest sequences needs to align first. The results of the MSA to be obtained through the proposed method of constructing the protein distance matrices need to be compared with those of the state-of-art MSA programs. In addition, the proposed distance matrix, dendrograms and MSA to be produced by the suggested method can be directly linked to the physicochemical feature of proteins that the amino acid indices used represent. Further research is necessary to establish a biological link between the protein sequences and amino acid indices used and the proposed distance matrix.

In this paper 25 widely used and accepted amino acid indices were used to construct the distance matrix. In the literature more than 500 unique amino acid indices exist [5]. Further study needs to be performed in order for a universal set of amino acid indices to be selected that can effectively represent all types of protein sequences. Additionally, as each amino acid index represent a unique physicochemical feature of proteins, the effect of each individual amino acid index used in the proposed distance needs to be further examined.

References

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

[2] A. Phillips, D. Janies, and W. Wheeler, "Multiple sequence alignment in phylogenetic analysis," *Molecular Phylogenetics and Evolution*, vol. 16, no. 3, pp. 317–330, 2000.

[3] 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.

[4] 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.

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

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] C. Chrysostomou, H. Seker, and N. Aydin, "Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis," in *Proceedings of EMBC 2011*, Boston, USA, August 2011, pp. 4955–8.

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

[27] 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.