Novel Protein Weight Matrix Generated from Amino Acid Indices

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

Abstract— **In recent years, numerous protein weight matrices have been developed that include physical characteristics of proteins, such as local sequence-structure information, alpha-helix information, secondary structure information and solvent accessibility states. These protein weight matrices are shown to have generally improved protein sequence alignments over classical protein weight matrices, like Point Accepted Mutation (PAM), Blocks of Amino Acid Substitution (BLOSUM), and GONNET matrices, where important limitations have been observe in recent works.**

In this paper, a novel protein weight matrix is constructed and presented. This protein weight matrix is not considered based on the mutation rate, like PAM or BLOSUM matrices, but on the physicochemical properties of each amino acid. In the literature, over 500 amino acid indices exist, each one representing a unique biological protein feature. For this study, 25 amino acid indices were selected. These amino acid indices represent general and widely accepted features of the amino acids.

By using the proposed protein weight matrix the following advantages can be obtained compared to the classical protein weight matrices. The proposed protein weight matrix is not biased to specific groups of protein sequences as the values are calculated from the amino acid indices, and not from the protein sequences. Additionally, for the proposed protein weight matrix, the same matrix can be considered regardless of the protein sequence's homology to be aligned or the mutation rate presented. A correlation to the physical characterisations of the amino acids that the protein weight matrix derived from can be achieved. Different similarity matrices can be generated when different physical characterisations of amino acids are considered.

I. INTRODUCTION

In bioinformatics, a protein weight matrix [1] expresses the rate at which one member in a sequence changes to another over time. A protein weight matrix can be created by using the 20 standard amino acid indices, generating a 20 by 20 matrix where each position represents the probability of a given amino acid to be substituted by one of the remaining. Protein weight matrices are generally used in protein or DNA sequence alignments, where the similarity between sequences relies upon their mutation rates as characterised in the matrix. In the literature, various protein weight matrices exist such as, Point Accepted Mutation (PAM), Blocks of Amino Acid Substitution (BLOSUM), and GONNET.

Originally, Point Accepted Mutation (PAM) matrices [1] were introduced in the late 1970s and used in protein sequence alignments. In order to create the PAM matrix of protein mutation [2], a Markov chain model [3] was utilised. These original PAM matrices were calculated based on 1572 measurements on mutation of 71 families with high similarity protein sequences.

For practical uses and to be able to compare different PAM matrices extracted from protein sequences with various lengths, PAM matrices are normalised. Therefore, the PAM matrix that is calculated from protein sequences with only one mutation occurring for every 100 amino acids will be called PAM1. Another example of the PAM matrix, PAM30, is normally used in the literature for a protein sequence's alignment [1]. The PAM30 matrix supplies substitution probabilities for sequences where 30 mutations occur for every 100 amino acids.

The Blocks of Amino Acid Substitution (BLOSUM) matrix that is a protein weight matrix used for protein sequence alignments was introduced in the early 90s [4]. In contrast to other protein weight matrices like PAM, which are generated from comparisons of high similarity protein sequences, BLOSUM matrices are supported on observed alignments. In the literature, BLOSUM matrices are commonly used in multiple alignments between biological diverging protein sequences [5].

To be able to create BLOSUM matrices, the BLOCKS database [6] was used to extract protein families with preserved regions. By using these protein sequences the relative frequencies of amino acids and their substitution probabilities were calculated. Furthermore, for the 210 possible substitutions of the 20 standard amino acids, the log-odds score was measured.

Various BLOSUM matrices were constructed by using different protein families in recent years, and a numeric system based on protein similarity is used to differentiate these matrices. A high number attached to BLOSUM matrix, like BLOSUM80 will indicate that this matrix was designed for aligning protein sequences with high similarity, in contrast to a low number, like BLOSUM45 that indicates being designed for aligning low similarity protein sequences.

The GONNET matrix was introduced in 1992 by Gonnet, Cohen and Benner [7]. This type of matrix calculates the differences between amino acids by using exhaustive protein pairwise alignments. The first step to derive the GONNET matrix is to align the given protein sequences by using other protein weight matrices like PAM or BLOSUM. The next step is to estimate the distance matrix by using the alignment, and interactively refine the alignment to calculate a new distance matrix. All the resulting matrices are normalised to 250 PAMs.

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

[2] Department of Computer Science and Digital Technologies, Faculty of Engineering and Environment, The University of Northumbria at Newcastle, NE1 8ST, Newcastle-upon-Tyne, The United Kingdom

cc390@le.ac.uk, huseyin.seker@northumbria.ac.uk

*Corresponding Author

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

As the authors indicated in the original description of the algorithm [7] the matrix is affected by the homology of proteins used. For this reason it is proposed for the initial alignment, the PAM250 protein weight matrix to be used, and for the iterative alignment refinements, a PAM matrix to be used that is appropriate to the homology of the protein sequences used.

Important limitations have been reported in recent works [8], [9]. Main disadvantages of PAM are that it assumes only uniform distribution of all mutation types and uses only high homology proteins in order to deduce relationships in diverse proteins [8], [9]. BLOSUM Matrix is limited to only a subset of conserved domains and ignores the closeness of relationship between the proteins [8], [9].

In order for GONNET matrix to be constructed, PAM and BLOSUM matrices are used along with exhaustive pairwise sequence alignments, thus it will inherit the same general disadvantages as the protein weight matrix it uses.

In recent years, numerous protein weight matrices have been developed that include physical characteristics of proteins, such as local sequence-structure information [10], alpha-helix information and secondary structure information [11]. These protein weight matrices are shown to have generally improved protein sequence alignments [10], [11]. However, these matrices have considered only one or two physical characteristics of proteins and therefore can not be generalised to model diverse set of protein families.

In this paper a novel generalised method is developed and presented in order to construct novel protein weight matrices based on the physicochemical properties of amino acids. 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 the literature, over 500 amino acid indices exist [13], 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 [14], [15] of the amino acids, like size [16], volume [17], molecular weight [18] and hydrophobicity [18], [19], [20], [21]. The complete list of the amino acid indices used for this analysis is presented in Table I and Tables II.

B. Novel Protein Weight Matrix generated from Amino Acid Indices

In this paper, a novel protein weight matrix is constructed and presented. This protein weight matrix is not considered based on the mutation rate, like PAM [1] or BLOSUM [4] matrices, but on the physicochemical properties of each amino acid. In order to calculate the protein weight matrix, the amino acids need to be converted to numerical values. These values can be derived from the amino acid indices. Each amino acid index represents a unique biological feature, which can encode amino acids. By using the numerical representation of the amino acids the protein weight matrix can be calculated. Further information regarding the amino acid indices used can be found in Section II-A. The following algorithm was proposed and used to construct the novel protein weight matrix:

TABLE I
Amino Acid Indices Used For The Alignment

ID	Name	Description	Reference
1	ZIMJ680102	Bulkiness	[22]
2	ZIMJ680104	Isoelectric point	[22]
3	HUTJ700102	Absolute entropy	[23]
4	DAWD720101	Size	[16]
5	GRAR740102	Polarity	[17]
6	GRAR740103	Volume	[17]
7	FASG760101	Molecular weight	[18]
8	FASG760102	Melting point	[18]
9	FASG890101	Hydrophobicity index	[18]
10	ZHOH040101	The stability scale from the knowledge-based atom-atom potential	[24]
11	OOBM770103	Long range non-bonded energy per atom	[25]
12	MANP780101	Average surrounding hydrophobicity	[19]
13	WOLR790101	Hydrophobicity index	[20]
14	FAUJ880101	Hydration potential	[26]
15	FAUJ880102	Smoothed upsilon steric parameter	[27]
16	ARGP820101	Hydrophobicity index	[21]
17	VELV850101	Electron-ion interaction potential	[28]
18	FAUJ880111	Positive charge	[27]
19	FAUJ880112	Negative charge	[27]
20	FAUJ880109	Number of hydrogen bond donors	[27]
21	KYTJ820101	Hydropathy index	[29]
22	BHAR880101	Average flexibility indices	[30]
23	Proscale_4	Recognition factors	[31]
24	Nl	Long-range contacts	[32]
25	Rk	Relative connectivity	[33]

1) As each amino acid index used in this paper, originated from different sources, each amino acid index is normalised using z-score [12], as shown in Equation 1

$$E' = \frac{E - \mu(E)}{\sigma(E)} \quad (1)$$

where E, μ and σ correspond to index value, mean value and standard deviation for a each amino acid index, respectively.

2) Calculate the protein weight matrix (W) using Euclidean distance. The Euclidean distance between two amino acids x and y where $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ and n is the number of features, can be calculated by Equation 2.

$$W = d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (2)$$

3) Scale the protein weight matrix using the following formula:

$$W' = -\frac{W}{M} \quad (3)$$

where M represents the maximum value of the protein weight matrix W.

TABLE II

AMINO ACID INDICES

ID	Name	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	ZIMJ680102	11.5	14.28	12.82	11.68	13.46	14.45	13.57	3.4	13.69	21.4	21.4	15.71	16.25	19.8	17.43	9.47	15.77	21.67	18.03	21.57
2	ZIMJ680104	6	10.76	5.41	2.77	5.05	5.65	3.22	5.97	7.59	6.02	5.98	9.74	5.74	5.48	6.3	5.68	5.66	5.89	5.66	5.96
3	HUTJ700102	30.88	68.43	41.7	40.66	53.83	46.62	44.98	24.74	65.99	49.71	50.62	63.21	55.32	51.06	39.21	35.65	36.5	60	51.15	42.75
4	DAWD720101	2.5	7.5	5	2.5	3	6	5	0.5	6	5.5	5.5	7	6	6.5	5.5	3	5	7	7	5
5	GRAR740102	8.1	10.5	11.6	13	5.5	10.5	12.3	9	10.4	5.2	4.9	11.3	5.7	5.2	8	9.2	8.6	5.4	6.2	5.9
6	GRAR740103	31	124	56	54	55	85	83	3	96	111	111	119	105	132	32.5	32	61	170	136	84
7	FASG760101	89.09	174.2	132.12	133.1	121.15	146.15	147.13	75.07	155.16	131.17	131.17	146.19	149.21	165.19	115.13	105.09	119.12	204.24	181.19	117.15
8	FASG760102	297	238	236	270	178	185	249	290	277	284	337	224	283	284	222	228	253	282	344	293
9	FASG890101	-0.21	2.11	0.96	1.36	-6.04	1.52	2.3	0	-1.23	-4.81	-4.68	3.88	-3.66	-4.65	0.75	1.74	0.78	-3.32	-1.01	-3.5
10	ZHOH040101	2.18	2.71	1.85	1.75	3.89	2.16	1.89	1.17	2.51	4.5	4.71	2.12	3.63	5.88	2.09	1.66	2.18	6.46	5.01	3.77
11	OOBM770103	-0.491	-0.554	-0.382	-0.356	-0.67	-0.405	-0.371	-0.534	-0.54	-0.762	-0.65	-0.3	-0.659	-0.729	-0.463	-0.455	-0.515	-0.839	-0.656	-0.728
12	MANP780101	12.97	11.72	11.42	10.85	14.63	11.76	11.89	12.43	12.16	15.67	14.9	11.36	14.39	14	11.37	11.23	11.69	13.93	13.42	15.71
13	WOLR790101	1.12	-2.55	-0.83	-0.83	0.59	-0.78	-0.92	1.2	-0.93	1.16	1.18	-0.8	0.55	0.67	0.54	-0.05	-0.02	-0.19	-0.23	1.13
14	FAUJ880101	1.28	2.34	1.6	1.6	1.77	1.56	1.56	0	2.99	4.19	2.59	1.89	2.35	2.94	2.67	1.31	3.03	3.21	2.94	3.67
15	FAUJ880102	0.53	0.69	0.58	0.59	0.66	0.71	0.72	0	0.64	0.96	0.92	0.78	0.77	0.71	0	0.55	0.63	0.84	0.71	0.89
16	ARGP820101	0.61	0.6	0.06	0.46	1.07	0	0.47	0.07	0.61	2.22	1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32
17	VELV850101	0.037	0.096	0.004	0.126	0.083	0.076	0.006	0.005	0.024	0	0	0.037	0.082	0.095	0.019	0.083	0.094	0.055	0.052	0.006
18	FAUJ880111	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
19	FAUJ880112	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
20	FAUJ880109	0	4	2	1	0	2	1	0	1	0	0	2	0	0	0	1	1	1	1	0
21	KYTJ820101	1.8	-4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2
22	BHAR880101	0.357	0.529	0.463	0.511	0.346	0.493	0.497	0.544	0.323	0.462	0.365	0.466	0.295	0.314	0.509	0.507	0.444	0.305	0.42	0.386
23	Proscale_4	78	95	94	81	89	87	78	84	84	88	85	87	80	81	91	107	93	104	84	89
24	Nl	3.92	3.78	3.64	2.85	5.55	3.06	2.72	4.31	3.77	5.58	4.59	2.79	4.14	4.53	3.57	3.75	4.09	4.83	4.93	5.43
25	Rk	1.05	0.94	0.93	0.88	1.17	0.93	0.85	0.99	0.99	1.11	1.07	0.88	1.04	1.07	0.92	0.96	0.99	1.05	1.05	1.12

III. RESULTS AND DISCUSSION

In order to calculate the protein weight matrix using the proposed method, the amino acids need to be converted to numerical sequences. By using Table I that lists the selected 25 amino acid indices, each amino acid was converted into 25 numerical values. By using the numerical representation of the amino acids, the pairwise Euclidean distance between all amino acids was calculated. Table III shows the generated protein weight matrix.

As Table III shows the amino acids with the greatest distance as calculated by the proposed method are

- Between G and W amino acids, Distance: -1
- Between G and R amino acids, Distance: -0.93
- Between D and I amino acids, Distance: -0.88
- Between D and W amino acids, Distance: -0.89

The amino acids with the smallest distance as calculated are

- Between I and V amino acids, Distance: -0.17
- Between L and V amino acids, Distance: -0.21
- Between M and F amino acids, Distance: -0.21
- Between N and Q amino acids, Distance: -0.23

By using the proposed protein weight matrix the following advantages can be obtained compared to the classical protein weight matrices such as PAM, BLOSUM and GONNET matrices:

- The proposed protein weight matrix is not biased to specific groups of protein sequences [34] as the values are calculated from the amino acid indices, and not from the protein sequences.
- By using the classical protein weight matrix, the use of a different matrix can have a major impact on the alignment [34]. For the proposed protein weight matrix, the same matrix can be considered regardless of the protein sequence's homology to be aligned or the mutation rate presented.
- A correlation to the physical characterisations of the amino acids that the protein weight matrix derived from can be achieved.
- Different similarity matrices can be generated when different physical characterisations of amino acids are considered. These characteristics are represented by the amino acid indices.

IV. CONCLUSIONS

In recent years, numerous protein weight matrices have been developed that include physical characteristics of proteins, such as local sequence-structure information [10], alpha-helix information and secondary structure information [11]. These protein weight matrices are shown to have generally improved protein sequence alignments over classical protein weight matrices, like PAM, BLOSUM and GONNET matrices, where important limitations have been observe in recent works [8], [9].

As stated in the paper more than 500 unique amino acid indices exist that represents unique physicochemical properties of amino acids. Future works need to be performed to find the optimal and universal set of features to represent amino acids. Finally, the proposed protein weight matrix will be compared to PAM, BLOSUM and GONNET matrices in aligning different classes of protein sequences and variant homology levels between groups of proteins.

REFERENCES

[1] M. Dayhoff, R. Schwartz, and B. Orcutt, "A model of evolutionary change in proteins," *Atlas of protein sequence and structure*, vol. 5, pp. 345–352, 1972.

[2] P. Lio and N. Goldman, "Models of molecular evolution and phylogeny," *Genome research*, vol. 8, no. 12, pp. 1233–1244, 1998.

[3] S. Meyn, R. Tweedie, and P. Glynn, *Markov chains and stochastic stability*. Cambridge University Press Cambridge, 2009, vol. 2.

[4] S. Henikoff and J. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, p. 10915, 1992.

TABLE III

SIMILARITY MATRIX GENERATED USING THE 25 AMINO ACID INDICES

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	-0.82	-0.46	-0.57	-0.47	-0.52	-0.57	-0.38	-0.55	-0.6	-0.49	-0.67	-0.43	-0.56	-0.45	-0.45	-0.38	-0.77	-0.56	-0.49
R	-	0	-0.56	-0.74	-0.82	-0.49	-0.7	-0.93	-0.46	-0.86	-0.83	-0.36	-0.72	-0.78	-0.71	-0.66	-0.6	-0.73	-0.64	-0.84
N	-	-	0	-0.46	-0.62	-0.23	-0.39	-0.54	-0.44	-0.72	-0.64	-0.44	-0.57	-0.68	-0.4	-0.31	-0.31	-0.73	-0.56	-0.65
D	-	-	-	0	-0.74	-0.44	-0.31	-0.64	-0.63	-0.88	-0.8	-0.63	-0.67	-0.77	-0.58	-0.49	-0.47	-0.89	-0.7	-0.82
C	-	-	-	-	0	-0.61	-0.75	-0.67	-0.6	-0.47	-0.46	-0.78	-0.37	-0.43	-0.59	-0.57	-0.49	-0.59	-0.54	-0.41
Q	-	-	-	-	-	0	-0.4	-0.65	-0.45	-0.74	-0.67	-0.39	-0.53	-0.64	-0.45	-0.37	-0.31	-0.71	-0.56	-0.68
E	-	-	-	-	-	-	0	-0.67	-0.55	-0.8	-0.72	-0.54	-0.63	-0.74	-0.54	-0.55	-0.5	-0.83	-0.64	-0.76
G	-	-	-	-	-	-	-	0	-0.74	-0.84	-0.76	-0.81	-0.72	-0.84	-0.53	-0.47	-0.57	-1	-0.79	-0.75
H	-	-	-	-	-	-	-	-	0	-0.63	-0.55	-0.36	-0.44	-0.54	-0.53	-0.58	-0.44	-0.6	-0.45	-0.58
I	-	-	-	-	-	-	-	-	-	0	-0.25	-0.8	-0.4	-0.36	-0.64	-0.75	-0.58	-0.47	-0.4	-0.17
L	-	-	-	-	-	-	-	-	-	-	0	-0.73	-0.28	-0.3	-0.61	-0.7	-0.53	-0.46	-0.34	-0.21
K	-	-	-	-	-	-	-	-	-	-	-	0	-0.64	-0.73	-0.56	-0.59	-0.52	-0.75	-0.62	-0.78
M	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.21	-0.55	-0.61	-0.42	-0.45	-0.31	-0.35
F	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.61	-0.72	-0.52	-0.36	-0.29	-0.36
P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.43	-0.37	-0.72	-0.56	-0.59
S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.29	-0.78	-0.66	-0.67
T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.63	-0.46	-0.49
W	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.35	-0.5
Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-0.41
V	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0

[5] A. Zomaya, *Handbook of nature-inspired and innovative computing: integrating classical models with emerging technologies*. Springer, 2005.

[6] J. Henikoff, S. Henikoff, and S. Pietrokovski, "New features of the blocks database servers," *Nucleic acids research*, vol. 27, no. 1, pp. 226–228, 1999.

[7] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database." *Science*, vol. 256, no. 5062, pp. 1443–1445, Jun 1992.

[8] Y. Yu and S. Altschul, "The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions," *Bioinformatics*, vol. 21, no. 7, pp. 902–911, 2005.

[9] D. Horner, W. Pirovano, and G. Pesole, "Correlated substitution analysis and the prediction of amino acid structural contacts," *Briefings in bioinformatics*, vol. 9, no. 1, pp. 46–56, 2008.

[10] Y. Huang and C. Bystroff, "Improved pairwise alignments of proteins in the twilight zone using local structure predictions," *Bioinformatics*, vol. 22, no. 4, pp. 413–422, 2006.

[11] D. Rice and D. Eisenberg, "A 3d-1d substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence1," *Journal of molecular biology*, vol. 267, no. 4, pp. 1026–1038, 1997.

[12] M. Marx and R. Larsen, *Introduction to mathematical statistics and its applications*. Pearson/Prentice Hall, 2006.

[13] 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.

[14] 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.

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

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

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

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

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] S. Henikoff and J. Henikoff, "Performance evaluation of amino acid substitution matrices," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 1, pp. 49–61, 1993.