MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **81** (2019) 193-207

ISSN 0340 - 6253

APGC: Universal Protein Prediction Based on Compositional, Physico–Chemical and Structural Information

Meng Kong, Yusen Zhang [1], Wei Chen

School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China.

(Received July 25, 2018)

Abstract

Since the traditional experimental methods are expensive and the computational methods predict specific types of proteins, it is important to develop computational methods to predict multiple types of proteins. In this paper, we propose a sequence-based model called APGC using the compositional, physico-chemical and structural information. Initially, we develop a novel sequence representation based on probability distribution (PD) and n-gap index. Afterwards, combinations of auto covariance (AC), probability distribution, n-gap index and composition of the moment vector (CMV) features are used to map the peptide sequences onto numeric feature vectors, which are subsequently used as input in support vector machine for prediction. The prediction results obtained in this study are significantly more universal and accurate than those of previously developed methods.

1 Introduction

Since predicting and classification of specific proteins is of fundamental importance to cure cancer, design new drugs, and understand the molecular mechanism in biological systems, many researchers have focused on this area in postgenome era. With the explosive growth of biological sequences in the post genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, preserving considerable sequence order information or key pattern characteristic.

[1] Corresponding author: `zhangys@sdu.edu.cn`

Most of the existing methods for predicting protein classification are based on the amino acid composition (AAC) [1]. As is well known, by using AAC to represent the sample of a protein, all its sequence-order information would be missing. To avoid complete loss of the sequence order information, the pseudo amino acid (PseAA) composition has been widely applied to various biomedical areas and computational proteomics [2,3]. Juan Mei et al. predict HIV-1 and HIV-2 proteins based on the concept of Chou's [4] PseAA composition and increment of diversity (ID), support vector machine (SVM), logistic regression (LR), and multilayer perception (MP) [5]. Using amino acid composition and binary profiles as the input of SVM, Tyagi et al. [6] proposed a model to identify anticnacer peptides (ACPs). Hajisharifi et al. [7], using Chou's pseudo amino acid composition and the local alignment kernel based method, also proposed a model to do the same. Leyi Wei et al. [8] present a sequence-based feature representation algorithm called adaptive k-skip-n-gram that sufficiently captures the intrinsic correlation information of Cell-penetrating peptides (CPPs). Y. Zhang et al. [9] put forward the so-called q-Wiener index by using hypergeometric series. This concept has also been applied for performing sequence analysis. In addition, more and more researchers have already studied physicochemical properties of 20 amino acids, such as hydrophobicity values, isoelectric point, relative molecular mass and ionization equilibrium constant (pKa values) [10-12] when it comes to sequences comparison. Using indexes of some physicochemical properties of 20 amino acids, Liu [13], Randić [14], Wu [15], Czerniecka [16] have proposed a number of different graphical representations of proteins, respectively. Extracting the features based on the properties of amino acid is essential and reasonable to compare proteins and study their function [17].

The above mentioned methods have their own advantages in generating knowledge for the prediction of specific types of proteins. However, all of these methods are not universal for multiple types of proteins. In this study we propose a novel model called APGC for universal protein prediction with high accuracy. Initially, in order to capture as much information of protein sequences as possible, we extract the 14 physicochemical properties features from the 531 indices by principal component analysis (PCA) [18,19]. Then, combinations of auto covariance (AC), probability distribution (PD), n-gap index and composition of the moment vector (CMV) features are used to map the peptide sequences onto numeric feature vectors, which are subsequently used as input in SVM for prediction. The APGC model structure is shown in Fig. 1. Our results are interpreted in the Section 'Results and Conclusions'.

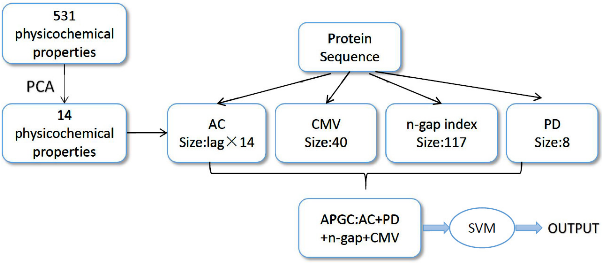

Figure 1. APGC model structure.

2 Methods

2.1 Auto covariance features (AC)

In order to capture as much information of protein sequences as possible, a variety of physicochemical properties are used in the procedure of feature extraction. All physicochemical properties used can be found in the Amino Acid index (AAindex) database, which store physicochemical or biochemical properties of amino acids or pair of amino acids. For the purpose of amino acid sequence transformation, we only consider the 544 amino acid properties (i.e., indices in AAindex1). Of the 544 indices, 13 have incomplete data or an over-representation of zeros, hence are removed. Thus 531 indices are evaluated for potential use in the dimensionality reduction of PCA. The normalization ensures that all properties are expressed as dimensionless numbers. Then, we extract the 14 features from the 531 indices which accumulated contribution rate reach 95%.

As a statistical tool for analyzing sequences of vectors developed by Wold et al. [20], AC has been adopted by more and more leading investigators for protein classification [21-23]. Given a protein sequence, AC variables describe the average interactions between residues, a certain lag apart throughout the whole sequence. Here, lag is the distance between one residue and its neighbour, a certain number of residues away.

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}\right) \times \left(X_{(i+lag),j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}\right) \quad (1)$$

where j represents one descriptor, i the position in the sequence X, n the length of the sequence X and lag the value of the lag.

2.2 Probability distribution features (PD)

Initially, twenty different kinds of amino acids can be divided into four classes based on DHP [24]: non-polar class (A): P, A, L, V, I, F, W and M; uncharged polar class (B): Q, S, T, Y, C, N and G; positive polar class (C): H, K and R; negative polar class (D): D and E.

Then, based on DHP,

protein I = {GLWSKIKEVGKEAAKAAAKAAGKAALGAVSEAV}

can be represented as:

reduced sequence I = {BAABCACDABCDAACAAACAABCAAABAABDAA}

C. Yu et al. [25] proposed a DNA sequence comparison by a probabilistic method. We improve his method and apply it to protein sequences. Fig.1 shows the four amino acids classes A, B, C and D are as follows: $A(1, 0.8), B(1, 0.6), C(1, 0.4), D(1, 0.2)$. The points in the graphical representation are obtained by summing the vectors representing descriptors in the reduced amino acid sequences. The endpoint of every vector represents one descriptor. Fig. 2 shows the graphical representation of the reduced amino acid sequences ($BAABCACDAB$).

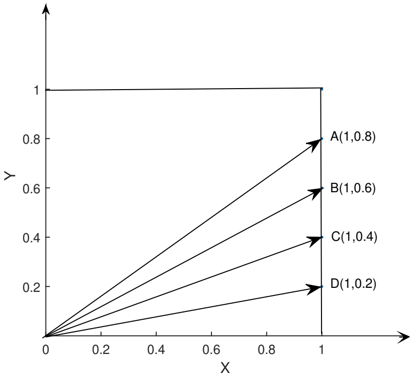

Figure 2. Amino acids vector system based on A(1,0.8), B(1,0.6), C(1,0.4), D(1,0.2).

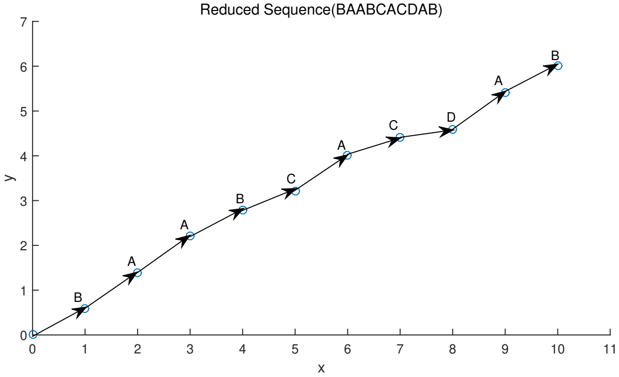

Figure 3. Graphical representation of reduced amino acids sequence (BAABCACDAB) based on the vector system.

For a reduced amino acids sequence of length n, we define its probability distribution as $(p_1, p_2, \cdots, p_n)$.

$$p_i = \frac{x_i - \overrightarrow{y}_i}{\frac{1}{2}n(n+1) - y_n} \tag{2}$$

where (x_i, y_i) represents the position of the ith descriptor in the protein graphical curve, $\overrightarrow{y}_i$ represents the y-coordinate value at the ith descriptor in the protein graphical curve according to Fig. 3. We can prove that this distribution is a discrete probability distribution [25].

We transform a protein sequence into a discrete probability distribution using our graphical representation. However, the probabilistic distribution of a protein sequence $(p_1, p_2, \cdots, p_n)$ is related to its length n. To overcome this limitation, we construct a slipping window. For a protein sequence of length n and a specific $N < n$, consider the $n - N + 1$ subsequences of length N. Thus we average these probabilistic distributions $(p_1, p_2, \cdots, p_N)$ that each of subsequences of length N.

2.3 N-gap index features

Considering previous reduced sequences based on DHP, we generate four types of reduced amino acid sequences according to their physicochemical properties, including polarity, acidity, charge and DHP [24]. The classifications of amino acids based on four properties [26] are shown in Table 1.

Table 1. amino acid classification.

Property	Classification
Polarity/acidity	DE RHK WYF SCMNQT GAVLIP
Acidity	DE KHR ACFGILMNPQSTVWY
Charge	KR AVNCQGHILMFPSTWY DE
DHP	PALVIFWM QSTYCNG HKR DE

In order to contain as much of the sequence-order effects as possible, we propose a simplified n-gap index model based on pseudo amino acid composition [27]. The sequence-order-correlated indexes from Equations (3) and (4) extract the sequence features.

Suppose a reduced protein sequence of L amino acid residues:

$$R_1R_2R_3R_4R_5R_6R_7\cdots\cdots R_L \tag{3}$$

where R_1 represents the character at sequence position 1, R_2 the character at position 2, and so forth. Schematic of the extraction scheme for n-gap features is shown in Fig. 4.

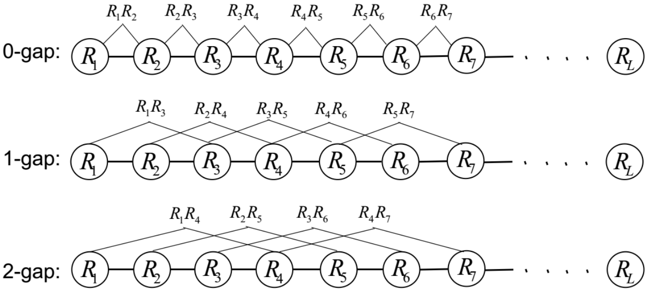

Figure 4. Schematic of the extraction scheme for n-gap features.

$$F_{n-gap} = \{\cup_{i=1}^{L-i-1} R_iR_{i+n+1} \mid i = 1, 2, \cdots, m\} \tag{4}$$

where m represents the number of amino acids groups of the reduced sequence, as well as the number of characters types in reduced sequence. The occurrences of characters and n-gap di-characters can be counted by $|R_i|$ and $|R_iR_j|$ in the set F_{n-gap}, where R_i and R_j represent i-th and j-th kind of character of the reduced sequence.

The element in the n-gap index vector is defined as:

$$V_{n-gap}(i) = \frac{|R_iR_j|_{F_{n-gap}}}{|R_i + c - n|_{F_{n-gap}}}, i = 1, 2, \cdots, m; j = 1, 2, \cdots, m \tag{5}$$

where $n = 0, 1, 2, \cdots$ and $c = max(n) + 1$. In this paper, $n = 0, 1, 2$, the overall dimension of n-gap index is $(5^2 + 3^2 + 3^2 + 4^2) \times 3 = 177$.

2.4 Composition of the moment vector features (CMV)

CMV's [28-30] which can reflect the frequencies and composition of amino acid residues have been widely used to predict certain protein sequences:

$$c_i^{(k)} = \frac{1}{L(L-1)\cdots(L-k)} \sum_{j=1}^{l} p_{ij}^k \quad (i = 1, 2, \cdots, 20) \tag{6}$$

where $k = 0, 1$ is the level of CMV; L is the length of sequence; i is the ith amino acid; l is the total number of i residues in the protein sequence; and p_{ij} is the position of the ith amino acid. Then, we take the 0-level position vector and the 1-level position vector into a 40-dimensional vector to characterize the protein sequence.

2.5 Features for prediction algorithms.

In the section "Auto Covariance(AC)",we extract the 14 features from the 531 physicochemical properties using PCA, then the AC features are created. In order to improve the prediction accuracy, $lag \times 14$ AC, 8 PD, 177 n-gap and 40 CMV vectors are combined, and these parameters are selected as the input parameters of SVM. Performance metrics of SVM in predicting HIV proteins is shown in Table 2.

A standard set of parameters has been used to evaluate the performance of various methods developed in this study. Following is a brief description of the parameters: (i) sensitivity, also referred to as recall, is the percent of correctly predicted allergen epitopes; (ii) specificity is the percent of correctly predicted non-allergen epitopes; (iii) accuracy is the proportion of correctly predicted epitopes; (iv) Matthews correlation coefficient (MCC).

AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming positive ranks higher than negative). AUC is between 0.1 to 1, which can be used to evaluate classifier visually. The larger AUC is, the better classifier is [31]. The parameters may be calculated by the following equations.

$$\begin{cases} Sn = \frac{TP}{TP+FN} \\ \\ Sp = \frac{TN}{TN+FP} \\ \\ Acc = \frac{TP+TN}{TP+FP+TN+FN} \\ \\ Mcc = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{[TP+FP][TP+FN][TN+FP][TN+FN]}} \end{cases} \tag{7}$$

where TP and FN refer to true positive and false negatives and TN and FP refer to true negatives and false positives; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; Mcc, the Mathew's correlation coefficient.

Table 2. Performance metrics of SVM in predicting HIV proteins.

Features	Performance metrics				
	Sn	Sp	Acc	Mcc	AUC
AC	96.57	91.07	95.28	87.17	0.99
AC+PD	98.46	92.89	97.13	91.85	1
AC+PD+n-gap	99.92	98.53	99.56	98.84	1
AC+PD+n-gap+CMV	1	1	1	1	1

3 Materials

The HIV protein dataset. The dataset is downloaded from the Swiss-Prot [32] (`http://www.uniprot.org/`). In order to get enough number of protein sequences, HIV-1 dataset and HIV-2 dataset with $\leq 90\%$ identity were used. In the final datasets, HIV-1 dataset consists of 260 non-redundant protein sequences and HIV-2 dataset consists of 81 non-redundant protein sequences.

The SPAAN dataset. Virulent protein datasets are taken from VirulentPred. The SPAAN dataset contains 469 adhesins and 703 non-adhesins proteins [34].

The Eukaryotic virulent protein dataset. The dataset is generated by employing the NTX-pred method randomly, consisting of 50 neurotoxins and 50 non-virulent proteins [35].

The Anticancer dataset. Hajisharifi et al. originally generated the Anticancer dataset [29]. Afterwards, the data was modified by Wei Chen [36]. It contains 138 anticancer peptides and 206 non-anticancer peptides.

RT proteins of HIV. HIV-1 and HIV-2 Reverse Transcriptase Proteins is downloaded from the Swiss-Prot (`http://www.uniprot.org/`). HIV-1 dataset with $\leq$ 50%

identity and HIV-2 dataset with $\leq 90\%$ identity are used. The final dataset includes 234 RT amino acid sequences from HIV-1 and 237 sequences from HIV-2 [37].

The GPCR dataset. The dataset of GPCR was originally generated by X.Xiao et al. This is a dataset containing G protein coupled receptors (GPCR) and non-GPCRs. None of the proteins included have $\geq$ 40% pairwise sequence identity to any other in the same subset. The final dataset includes 365 GPCR sequences and 365 non-GPCR sequences [38].

The HPV protein dataset. The HPV protein dataset is downloaded from the NCBI database. This dataset contains 444 amino acids sequences from HPV-16 and 470 amino acids sequences from HPV-18.

The Crotonyllysine sites dataset. Qiu et al. originally generated the Crotonyllysine sites dataset [39]. Qiu's training dataset is extracted from Uniprot database, and it consists of 169 experimentally annotated crotonyllysine sites and 847 non-crotonyllysine sites.

The EBNA protein dataset. The HPV protein dataset is downloaded from the Swiss-Prot. This dataset contains 117 amino acids sequences from EBNA-1 and 72 amino acids sequences from EBNA-2.

The N-formylated proteins. The N-formylated protein dataset is originally generated by Zhe Ju et al. It consists of 74 N-formylated sites and 69 non-N-formylated sites [40].

The Hepatovirus proteins. The Hepatovirus protein dataset is downloaded from the Swiss-Prot. This dataset contains 6888 amino acids sequences from Hepatovirus-a and 147045 amino acids sequences from Hepatovirus-b.

The CPP proteins. Cell-penetrating peptides (CPPs) are short peptides (5-30 amino acids) that can enter almost any cell without significant damage. This is a dataset containing 462 CPP proteins and 462 non-CPP proteins [8].

4 Results and Conclusions

4.1 Results

A ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. The Fig. 5 shows the performance of APGC approach for different datasets. The prediction performance of previously mentioned datasets are shown in Table 3. The results that we have generated as a comparison with previously mentioned approaches are shown in Table 4.

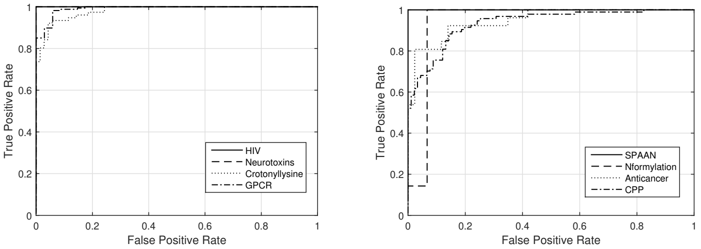

Figure 5. The ROC curves of APGC approach for different datasets.

Table 3. Results of APGC for all datasets.

dataset	Sn	Sp	Acc	Mcc	AUC
HIV	1.00	1.00	1.00	1.00	1.00
RTHIV	99.14	99.18	99.15	98.30	1.00
SPAAN adhesins	85.25	94.69	90.92	80.99	0.95
GPCR	95.54	91.68	93.63	87.37	0.98
Euk neurotoxins	98.55	89.72	94.10	88.22	1.00
Hepatovirus	1.00	99.80	99.90	99.80	1.00
Crotonyllysine	99.44	82.77	96.70	86.44	0.96
HPV	99.57	1	99.79	98.59	1.00
N-formylation sites	98.61	79.42	89.57	80.37	0.92
EBNA	90.23	84.48	87.47	74.33	0.93
Anticancer	82.65	94.11	89.50	78.13	0.96
CPP	88.52	92.28	90.41	80.74	0.96

Table 4. Comparison of APGC with previous approaches.

method	dataset	Sn	Sp	Acc	Mcc	AUC
APGC	SPAAN	**85.25**	**94.69**	**90.92**	**80.99**	**0.95**
VirulentPred [33]	SPAAN	58.44	87.79	70.14	0.46	-
2Gram [41]	SPAAN	49.17	99.30	79.35	0.59	0.94
APGC	Neurotoxins	**98.55**	**89.72**	**94.1**	**88.22**	**1**
VirulentPred	Neurotoxins	96	16	56	-	-
NTXPred(FNN) [42]	Neurotoxins	89.65	78.78	84.19	0.69	-
NTXPred(RNN) [42]	Neurotoxins	89.12	96.35	92.75	0.86	-
AS [40]	Neurotoxins	92.00	1	96.00	0.92	0.99
APGC	HIV	**1**	**1**	**1**	**1**	**1**
ID(SVM) [5]	HIV	97.47	98.59	98.48	96.04	-
ID(LR) [5]	HIV	97.43	97.08	97.87	94.51	-
ID(MLP) [5]	HIV	98.42	98.42	98.78	96.85	-
APGC	GPCR	**95.54**	**91.68**	**93.63**	**87.37**	**0.98**
GPCR-CA [38]	GPCR	91.08	92.22	91.64	0.8330	
APGC	Crotonyllysine	**99.44**	**82.77**	**96.70**	**86.44**	**0.96**
CrotPred [43]	Crotonyllysine	79.41	77.78	78.23	0.5259	-
Qiu [39]	Crotonyllysine	71.69	98.70	94.43	0.7780	-
APGC	Nformylation	**98.61**	**79.42**	**89.57**	**80.37**	**0.92**
Zhou [44]	Nformylation	77.78	95.06	90.74	0.7478	-
Binary [40]	Nformylation	40.00	98.33	88.28	0.5256	0.90
APGC	Anticancer	**82.65**	**94.11**	**89.50**	**78.13**	**0.96**
Hajisharifi [37]	Anticancer	89.70	85.18	92.68	0.784	-
APGC	CPP	**88.52**	**92.28**	**90.41**	**80.74**	**0.96**
NB [7]	CPP	82.7	94.8	88.7	0.781	-
LibSVM [7]	CPP	88.1	92.6	90.4	0.810	-

4.2 Conclusions

This paper provides an alignment-free measure, developing a novel combination feature model for analyzing protein sequences based on physicochemical properties of 20 amino acids. We propose a novel probabilistic method for protein sequence comparison that uses a graphical representation. After constructing the graphical representation, we are able to construct a probability distribution for a protein sequence. Then, we extracted a combined vector by mixing normalized features together (AC, CMV and PD and n-gap index features) for each sequence and used it as input of SVM. The results show that our approach provides a new, powerful tool to predict various types of protein sequences for both molecular biologists and computational scientists.

As pointed out in and realized in a series of recent publications, user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful computational tools and enhancing their impact. Our future efforts will be to establish a web-server for the prediction method reported in this paper.

Acknowledgments: We gratefully acknowledge the anonymous reviewers who read our paper and gave some constructive comments. This work is supported by the National Natural Science Foundation of China (Nos. 61877064)

References

[1] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, *J. Biochem.* **99** (1986) 152–162.

[2] H. Mohabatkar, M. Beigi, K. Abdolahi, S. Mohsenzadeh, Prediction of allergenic proteins by means of the concept of chous pseudo amino acid composition and a machine learning approach, *Med. Chem.* **9** (2013) 133–137.

[3] W. Zhong, S. Zhou, Molecular science for drug development and biomedicine, *Int. J. Mol. Sci.* **15** (2014) 20072–20078.

[4] K. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, *J. Theor. Biol.* **273** (2011) 236–247.

[5] J. Mei, J. Zhao, Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers, *Sci. Rep.* **8** (2018) #2359.

[6] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, G. P. Raghava, In silico models for designing and discovering novel anticancer peptides, *Sci. Rep.* **3** (2013) #2984.

[7] Z. Hajisharifi, M. Piryaiee, B. M. Mohammad, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, *J. Theor. Biol.* **341** (2014) 34–40.

[8] L. Wei, J. Tang, Q. Zou, SkipCPP-Pred: an improved and promising sequence–based predictor for predicting cell–penetrating peptides, *BMC Gen.* **18** (2017) 1–11.

[9] Y. Zhang, I. Gutman, J. Liu, Z. Mu, q-Analog of Wiener index, *MATCH Commun. Math. Comput. Chem.* **67** (2012) 347–356.

[10] D. Sun, C. Xu, Y. Zhang, A novel method of 2D graphical representation for proteins and its application, *MATCH Commun. Math. Comput. Chem.* **75** (2016) 431–446.

[11] X. Xia, W. Li, What amino acid properties affect protein evolution? *J. Mol. Evol.* **47** (1998) 557–564.

[12] Z. Qi, M. Jin, S. Li, J. Feng, A protein mapping method based on physicochemical properties and dimension reduction. *Comput. Biol. Med.* **57** (2015) 1–7.

[13] Y. Liu, D. Li, K. Lu, Y. Jiao, P. He, P-H Curve, a graphical representation of protein sequences for similarities analysis, *MATCH Commun. Math. Comput. Chem.* **70** (2013) 451–466.

[14] M. Randić, 2-D Graphical representation of proteins based on physico-chemical properties of amino acids,*Chem. Phys. Lett.* **444** (2007) 176–180.

[15] Z. C. Wu, X. Xiao, K. C. Chou, 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, *J. Theor. Biol.* **267** (2010) 29–34.

[16] A. Czerniecka, D. Bielińska–Waż, P. Waż, T. Clark, 20D-dynamic representation of protein sequences, *Genomics* **107** (2016) 16–23.

[17] L. Yu, Y.Zhang, I. Gutman, Protein sequence comparison based on physicochemical properties and the position–feature energy matrix, *Sci. Rep.* **7** 2017 #46237

[18] K. Diamantaras, S. Kung, *Principal Component Neural Networks – Theory and Applications*, Wiley, New York, 1996, pp. 74–75.

[19] M. Bishop, Pattern recognition and machine learning, information science and statistics, *Publ. Am. Stat. Assoc.* **103** (2006) 886–887.

[20] S. Wold, J. Jonsson, M. Sjörström, M. Sandberg, S. Rännar, DNA and peptide sequences and chemical processes mutlivariately modelled by principal component analysis and partial least-squares projections to latent structures, *Anal. Chim. Acta.* **277** (1993) 239–253.

[21] Y. Guo, M. Li, M. Lu, Z. Wen, Z. Huang, Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform, *Proteins Struct. Func. Bioinf.* **65** (2010) 55–60.

[22] Z. Wen, M. Li, Y. Li, Y. Guo, K. Wang, Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition, *Amino Acids* **32** (2007) 277–283.

[23] I. Doytchinova, D. Flower, VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines, *BMC Bioinf.* **8** (2007) #4.

[24] G. Han, Z. Yu, V. Anh, A. Krishnajith, Y. Tian, An ensemble method for predicting subnuclear localizations from primary protein structures, *PLoS ONE* **8** (2013) #e57225.

[25] C. Yu, M. Deng, S. Yau, DNA sequence comparison by a novel probabilistic method, *Inf. Sci.* **181** (2011) 1484–1492.

[26] J. Bergera, S. Mitraa, M. Carlib, A. Nerib, Visualization and analysis of DNA sequences using DNA walks, *J. Franklin Inst.* **341** (2004) 37–53.

[27] K. Chou, Y. Cai, Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition, *J. Cell. Biochem.* **90** (2003) 1250–1260.

[28] C. Jia, Y. Zhang, Z. Wang, SulfoTyrP: a high accuracy predictor of protein sulfotyrosine sites, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 227–240.

[29] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, *BMC Bioinf.* **9** (2008) #226.

[30] T. Liu, C. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, *J. Theor. Biol.* **267** (2010) 272–275.

[31] C. Xu, L. Ge, Y. Zhang, M. Dehmer, I. Gutman, Prediction of therapeutic peptides by incorporating q-Wiener index into Chou's general PseAAC, *J. Biomed. Inf.* **75** (2017) 63–69.

[32] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, B. Boeckmann, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.* **31** (2003) 365–370.

[33] A. Garg, D. Gupta, VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens, *BMC Bioinf.* **9** (2008) 62–73.

[34] S. Saha, G. Raghava, Prediction of neurotoxins based on their function and source, *Sillico Biol.* **7** (2007) 369–382.

[35] Z. Hajisharifi, M. Piryaiee, B. M. Mohammad, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, *J. Theor. Biol.* **341** (2014) 34–40.

[36] W. Chen, H. Ding, P. Feng, H. Lin, K. C. Chou, iACP: a sequence–based tool for identifying anticancer peptides, *Oncotarget.* **7** (2016) 16895–16909.

[37] M. Behbahani, H. Mohabatkar, M. Nosrati, Discrimination of HIV-1 and HIV-2 Reverse Transcriptase Proteins Using Chous PseAAC, *Iran. J. Sci. Technol. Trans. Sci.* (2017) 1–7.

[38] X. Xiao, P. Wang, K. Chou, GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes, *J. Comput. Chem.* **30** (2009) 1414–1423.

[39] W. Qiu, B. Sun, H. Tang, J. Huang, H. Lin, Identify and analysis crotonylation sites in histone by using support vector machine, *Artif. Intell. Med.* **83** (2017) 75–81.

[40] Z. Ju, J. Cao, Prediction of protein N-formylation using the composition of k-spaced amino acid pairs, *Anal. Biochem.* **534** (2017) 40–45.

[41] L. Nanni, A. Lumini, S. Brahnam, An empirical study of different approaches for protein classification, *Sci. World J.* **2014** (2014) #e236717.

[42] S. Saha, G. Raghava, Prediction of neurotoxins based on their function and source, *Sillico Biol.* **7** (2007) 369–382.

[43] G. Huang, W. Zeng, A discrete hidden Markov model for detecting histonecrotonyllysine sites, *MATCH Commun. Math. Comput. Chem.* **75** (2016) 717–730.

[44] Y. Zhou, T. Huang, G. Huang, N. Zhang, X. Kong, Y. Cai, Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method, *Neurocomputing* **217** (2016) 53–62.