Prediction of Lysine Glycation PTM site in Protein using Peptide Sequence Evolution based Features

S.M. Shovan
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
sm.shovan@gmail.com

Md. Al Mehedi Hasan
Department of Computer Science & Engineering
University of Rajshahi
Rajshahi, Bangladesh
mehedi_ru@yahoo.com

***Abstract*— Glycation is a post-translational modification which is non-enzymatic in nature. It is closely associated with different biological functions and responsible for many diseases, for example, diabetes, renal failure etc. Identification of Glycation sites is very important in the development of drugs and the research areas but identifying manually in laboratory is laborious, costly and time consuming. Development of a computational tool will be very useful for the Glycation sites prediction with high accuracy. In our experiment, a new feature extraction technique, called peptide sequence evolution based feature representation, is introduced which gave an Accuracy of 95.94 ± 0.54%, a Sensitivity of 98.20% and a Specificity of 90.67±1.07% after running 10-fold cross-validation five times. This result outperforms the previously developed tools BPB_GlySite, NetGlycate, PreGly and Gly_PseAAC.**

Keywords*— *Lysine Glycation site, post-translational modification, data imbalance issue, support vector machine, peptide sequence based feature.

I. INTRODUCTION

Post-translational modifications, abbreviated as PTMs, are the alternation of protein after the formation of protein in the process of Central Dogma. Normally these modifications or alternations occur on the side chain of specific protein units called as amino acids. For example, Lysine residue is responsible for Glycation PTM. PTMs can be several types. For example, hydroxylation, carbonylation, nitration, glycation and so on. These PTMs are responsible for the diversity of protein in terms of their functions and structures additionally, the plasticity of active cells are greatly influenced by these alternations [1]. Expansion of genetic code and regulation of physiology of the living cells are the results of PTMs [2,3].

Lysine glycation is a very common post-translational modification which adds sugar molecule to a peptide without the help of enzyme. Efficacy, conformation and immunogenicity are different biological processes, those are effected significantly by lysine glycation [4]. Glycosylation is a process of covalent bonding of different sugar molecule such as fructose or glucose) which does not require the influence of enzymes. Unlikely, glycosylation requires the controlling influence of enzymes. So, sometime glycation is called the non-enzymatic glycosylation. The process of glycation takes two steps, 1) A more steady Amadori product is formed from the rearrangement of less steady Schiff base, 2) Advanced Glycation End, abbreviated as AGEs, products can't be reversed, are constructed from Amadori product [5,6].

Different previous researches have shown that glycation PTM can cause several human disease, such as, vascular complications for diabetes [7], Parkinson's disease and Alzheimer's disease [8], renal failure [9]. So it will be very useful to discover the hidden molecular mechanisms of glycation for the treatment of previously mentioned disease as well as in the drug development. Sadly, the incident mechanism is still unknown.

The initial step for the better understanding of molecular mechanisms is to recognize the glycated portion of the protein segment and the responsible glycation sites with high accuracy. For the detection of glycation sites, mass spectrometry is one of the most popular method that has been used these days [10,11]. But this manual experimental method is costly, laborious as well as time consuming. So developing a computational tool for predicting glycation sites will be very much useful. So far several computational tool have been developed to predict glycation sites from protein sequences. First predictor NetGlycate [12] used ensemble artificial neural network algorithm in 2006. In 2015 another computational tool PreGly [13] is introduced for the prediction of glycation sites which indicated that for the prediction of glycation sites, the composition of k-spaced amino acid pairs features contributed most. Another tool named Gly-PseAAC [14] is developed for the prediction of glycation sites which uses position-specific amino acid propensity (PSAAP) and sequence order information for the problem of glycation residue prediction. Another tool was developed named BPB_GlySite [15] was developed that uses Bi-Profile Bayes (BPB) as a feature extraction technique for encoding the training protein segment. So far the performance of these tools mentioned above is not quite satisfactory. The further improvement is possible.

For improving the performance, it is essential to find a good technique which will contain sufficient information that helps us to effectively separate the glycated sites from non-glycated sites. In our study, we have introduced a new feature extraction technique, called peptide sequence evolution based feature. The typical sequential evolution method produces Position-Specific Scoring Matrix (PSSM) [16,17] that acts as a feature extractor. In contrast, our new proposed method generates the matrix which uses the concept of PSSM but in the different way. Support Vector Machine, in short SVM, with Radial Basis Function, abbreviated as RBF, kernel is chosen for classification in the proposed method. PTM dataset faces data imbalance issue. Imbalanced data set contains unequal number of data in each classes. This imbalance data

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

set can bias the result if not properly handled. For balancing out the data, oversampling technique has been used.

For dataset, we have taken the dataset used by the BPB_GlySite [15] to compare the result. For tuning the parameter of support vector machine, 10-fold cross validation is applied five times to avoid any unexpected results as well as to find a concrete solution. Each time data is split into different 10 segments and a complete run is performed.

For building a successful sequence based predictor, we have followed Chou's five-step rules: [18,19] (i) For successful training and testing, a valid and reliable dataset should be chosen or built, (ii) biological sequence, for instance protein sequence, should be represented into mathematical model that is capable of separating the classes effectively, (iii) performing the prediction, a solid and robust algorithm should be introduced or developed, (iv) for measuring the most probable performance of the model, a model validation test should be performed accurately, (v) At last, develop a human interactive web-server which should be accessible to public. For this paper, the fifth rule is absent because, the web-server is under construction.

II. Material and Methods

A. Benchmark dataset

Benchmark dataset of BPB_GlySite [15] has also been used in our experiment. The dataset contains total 696 peptide sequences where 223 peptide sequences are positive (glycated peptide) and 446 are negative (non glycated peptide). The benchmark data set has been collected from http://123.206.31.171/BPB_GlySite/data.html. Each row consists of protein sequence which is constructed using sliding window of window size 15. The responsible residue lysine (K) is kept in the middle. The length of both upstream and downstream is kept 7. For ensuring the uniformity of length of each protein sequence, necessary number of dummy residue, denoted by "X", has been padded in both sides when needed.

A peptide sample is represented as follows,

$$P_\xi(\odot) = R_{-\xi}R_{-(\xi-1)} \dots R_{-2}R_{-1}\Theta R_1R_2 \dots R_{+(\xi-1)}R_{+\xi} \quad (1)$$

Here, Θ denotes amino acid lysine (K). ξ is an integer. $R_{-\xi}$ represents upstream and $R_{+\xi}$ represents downstream of peptide sample. Total length of the peptide, substring of protein, sample is (2ξ+1).
Each peptide sample falls under one of the two categories as follows,

$$P_\xi(\odot) \in \begin{cases} P_\xi^+(\odot), \text{if central residue is a glycation site} \\ P_\xi^-(\odot), \text{else} \end{cases} \quad (2)$$

$P_\xi^+(\odot)$ represents positive glycation segments and $P_\xi^-(\odot)$ represents negative glycation segments and ∈ refers to membership of set theory.

Benchmark dataset has constructed as follows,

$$S_\xi(K) = S_\xi^+(K) \cup S_\xi^-(K), \text{ when } \Theta = K \quad (3)$$

Where, $S_\xi^+(\odot)$ contains glycated segments, $P_\xi^+(\odot)$ and $S_\xi^+(\odot)$ contains non-glycated segments, $P_\xi^-(\odot)$ and ∪ is the union operation of set theory.

B. Feature extraction

Feature extraction is the key for building a successful predictor as it contains the information which is useful to separate positive class from negative class, in our case, to identify whether a protein is glycated or not. As the protein contains the sequence of each amino acids, it is required to represent the peptide sequences into mathematics because most of the machine learning algorithm take numeric inputs. The representation should carry enough information that is sufficient to make enough distinction between positive and negative classes.

Protein sequences go through several random evolutions. This evolution includes single or multiple alternations, insertions or deletions of amino acids. In results, these create completely new sequences, though these came from the same origin. To find these similarity among the protein/peptide sequences, a sophisticated tool named BLAST, stands for Basic Local Alignment Search Tool, which finds the resemblance of a query sequence with the database of sequences. PSI-BLAST [20], stands for Position-Specific Iterative Basic Local Alignment Search Tool, provides a PSSM [20] (Position-Specific Scoring Matrix) for each query sequence only if the query is not eliminated for being similar to another sequence with a predefined threshold. Then the PSSM [20] matrices of all selected protein sequences are further processed to represent as feature vector which will eventually be fed into classifier.

1) Peptide sequence evolution based feature extraction

The typical sequential evolutionary feature extraction technique uses complete database of protein sequences. In contrast, we have created a custom database from the peptide sequences of our dataset which gives an acceptable performance as well as eliminates the necessity of the premade protein database. The PSI-BLAST [20] is run against the custom made database where each peptide sequence is treated individually as a query sequence which provides PSSM like matrix as an output. Further processing is done to extract the features from these matrices. This method is named as peptide sequence evolution based feature because it takes only peptide sequences to make custom database instead of considering the whole protein database.

The following technique is used to generate feature vector from the dataset.

a) A query peptide sequence is denoted by P and can be represented as

$$P = R_1R_2R_3R_4R_5 \dots \dots \dots R_L \quad (4)$$

From the study of Schaffer et al. [20] we know that, the information of sequential study evaluation of P, can be represented by a 20*L dimensional matrix according to the equation (5).

$$\begin{bmatrix} \dot{E}_{1\to1} & \dot{E}_{2\to1} & \dots\dots & \dot{E}_{L\to1} \\ \dot{E}_{1\to2} & \dot{E}_{2\to2} & \dots\dots & \dot{E}_{L\to2} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \dot{E}_{1\to20} & \dot{E}_{2\to20} & \dots\dots & \dot{E}_{L\to20} \end{bmatrix} \quad (5)$$

Where, 20 refers to the 20 different amino acids according to the alphabetic order, L refers to the length of P, $\dot{E}_{i\to j}$ denotes the propensity of the amino acid residue at position 'i' being mutated to the amino acid at position 'j' at the time of evaluation process. The custom database has been created by all the peptide sequence in the benchmark dataset [15] and each sequence is individually treated as a query sequence against the custom database. The search method used two iterations and the cutoff value for E-value is chosen as 0.001 for it's widely usage.

b) The new matrix can be derived from the matrix in equation (5) as follows,

$$\begin{bmatrix} E_{1\to1} & E_{2\to1} & \dots\dots & E_{L\to1} \\ E_{1\to2} & E_{2\to2} & \dots\dots & E_{L\to2} \\ \vdots & \vdots & & \vdots \\ . & . & & . \\ . & . & & . \\ E_{1\to20} & E_{2\to20} & \dots\dots & E_{L\to20} \end{bmatrix} \quad (6)$$

With

$$E_{i\to j} = \frac{\dot{E}_{i\to j} - \bar{\dot{E}}_j}{SD(\bar{\dot{E}}_j)} \qquad \text{i= 1,2,…..,L ; j= 1,2,…, 20}$$

Where

$$\bar{\dot{E}}_j = \frac{1}{L}\sum_{i=1}^{L} \dot{E}_{i\to j} \quad \text{j= 1,2,…..,20}$$

Here $\bar{\dot{E}}_j$ refers to the mean of $\dot{E}_{i\to j}$ for i= 1,2,…20 and the standard deviation is denoted and defined by the following equation,

$$SD(\bar{\dot{E}}_j) = \sqrt{\sum_{i=1}^{L}\left[\dot{E}_{i\to j} - \bar{\dot{E}}_j\right]^2 / L} \quad (7)$$

c) The new matrix MM^T is measured by multiplying M with the transpose of M that becomes (20×20) matrix of 400 elements. Moreover the resultant matrix is a symmetric matrix contains 210 unique information, 20 comes from the diagonal and 190 = (400-20)/2 elements from lower or upper triangle matrix. In our study, the lower triangular matrix with the diagonal has been considered, as follows,

$$\begin{bmatrix} (1) & & & & \\ (2) & (3) & & & \\ (4) & (5) & (6) & & \\ \vdots & \vdots & \vdots & & \\ (191) & (192) & (193) & \dots. & (210) \end{bmatrix} \quad (8)$$

The matrix of equation converted into vector representation of 210 elements as show below,

$$P_{evo} = [\Theta_1^E \Theta_2^E \dots \Theta_u^E \dots \Theta_{210}^E]^T \quad (9)$$

C. Imbalance dataset management

The benchmark dataset [15] contains 223 positive and 446 negative sites. For balancing the imbalanced data, oversampling technique has been used. After performing the feature extraction step, one negative site is eliminated and now there are 445 negative sites and 223 positive sites for which the P_{evo} has been calculated. The ratio of negative and positive is 1.9955:1 ≈ 2:1. For making the dataset balanced, we took each positives two times while keeping the negatives untouched for making the ratio almost 2:2, equivalently 1:1. So total number of data is 223*2+445= 891, where 223*2= 446 are positives and 445 are negatives. This oversampling technique helped us resolving the data imbalance problem.

D. SVM classification

Support Vector Machine, also known as SVM in short, is a binary classifier that gives the optimal hyper plane for separating the classes by the margin with maximum width [21]. So the constraint problem becomes as follows,

$$\text{Minimize}_{\,w,b}\ \frac{1}{2}\|W\|^2 \quad (10)$$

$$\text{Subject to } y_i(w^T x_i + b) \geq 1, \quad \text{i= 1,2,3,…..,n}$$

A penalty term is added to allow errors for finding the wider margin.

$$\text{Minimize}_{\,w,b}\ \frac{1}{2}\|W\|^2 + \text{C}\sum_{i=1}^{n}\xi_i \quad (11)$$

$$\text{Subject to} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \text{i= 1,2,3,…..,n}$$
$$\xi_i \geq 0, \quad \text{i= 1,2,3,…..,n}$$

From the Lagrange multipliers, the dual formulation is obtained and represented with respect to α_i variable. [21,22].

$$\text{Maximize}_{\,\alpha}\ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\ \sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j x_i^T x_j \quad (12)$$

$$\text{Subject to } \sum_{i=1}^{n} y_i\alpha_i = 0\ ,\ \ 0 \leq\ \alpha_i \leq C$$
$$\text{For i = 1,2,3,…..,n}$$

The final form of the discriminant function which has linearity property becomes as follows,

$$\text{F(x)} = \sum_i^n y_i\alpha_i x_i^T x\ + b \quad (13)$$

To make the linear function nonlinear, we need to use a nonlinear function Ø: X→ F, which projects from input space X to another space called feature space F. So, the form of the optimization function becomes as follows using the kernel function [22-23],

$$\text{Maximize}_{\,\alpha}\ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\ \sum_{j=1}^{n}\alpha_i\,\alpha_j y_i k(x_i, x_j) \quad (14)$$

$$\text{Subject to } \sum_{i=1}^{n} y_i\alpha_i = 0\ ,\ \ 0 \leq\ \alpha_i \leq C$$
$$\text{For i = 1,2,3,…..,n}$$

In terms of kernel function, the discriminant function becomes,

$$\text{F(x)} = \sum_i^n y_i\alpha_i k(x_i, x_j) + b \quad (15)$$

In our experiment, radial basis function, in short RBF, is used as a kernel function which is defined by the following formula,

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i, x_j\|^2}{2\sigma^2}\right) \quad (16)$$

Here, σ is the width for the kernel function.

E. Experimental setting

For getting the best parameter value for classifier or the predictor, three most used methods are k-fold cross-

validation (subsampling), independent dataset test and jackknife test [24-26]. Jackknife test estimates with (n-1) sampling leaving one sample for testing. This estimates gives same result every time it runs for a specific dataset but the limitation is, it runs 'n' times for which the computational time is very high for large dataset. In contrast, k-fold cross-validation takes very low computational time for large dataset compared to jackknife testing.

In our study, k-fold cross-validation, more specifically 10-fold cross-validation, method has been used, where the whole dataset is divided into 10-splits and each time 9 splits are considered for training and one split left for testing. The 10-fold cross-validation is done for 5 times for ensuring the stability.

F. Measuring matries

For calculating the predictive performance and capability of a classifier or predictor, there are several measuring matrices. In our study, we have used: i) Total accuracy denoted as Acc, ii) sensitivity denoted as Sn, and iii) specificity denoted as Sp [27-30]. The equations are given below.

$$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

$$\text{Sn} = \frac{TP}{TP+FN} \quad (18)$$

$$\text{Sp} = \frac{TN}{TN+FP} \quad (19)$$

Where, TP, stands for True Positive, means the number of glycated peptides are correctly predicted, TN, stands for True Negative, means the number of non-glycated peptides are correctly predicted, FP, stands for False Positive, means the number of non-glycated peptides wrongly predicted as glycated peptides, and FN, stands for False Negative, means the glycated peptides wrongly predicted as the non-glycated peptides.

III. RESULTS AND DISCUSSION

Efficient model selection is required for the best performance of SVM. 5 complete runs of the 10-fold cross validation is computed.

TABLE I. OPTIMAL C AND **σ** (SIGMA) OF 5 COMPLETE RUN OF THE 10-FOLD CROSS-VALIDATION**.**

No. of complete run	Lysine (K)	
	C	σ
1st	2^{-8}	2^{1}
2nd	2^{-8}	2^{1}
3rd	2^{-8}	2^{1}
4th	2^{-8}	2^{1}
5th	2^{-8}	2^{1}

Here, C is the penalty term for soft margin and sigma σ is the width of Radial Basis Function (RBF). To find the optimal C and σ, we search for the best value for both of these from 2^{-8} to 2^{8}. In our model, we considered the optimal value of C and σ are those values for which the model gives maximum accuracy. As we have run the system 5 times, we got 5 optimal C and σ values for each time. Those values are shown in the table 1.

IV. COMPARISON WITH EXISTING MODELS

For the comparison of our model with the performance of existing tools BPB_GlySite[15] and Gly-PseAAC [14], we found the following table.

TABLE II. COMPARISON AMONG GLY-PSEAAC, BPB_GLYSITE AND OUR MODEL IS SHOWN IN THE FOLLOWING TABLE

Predictor	Matrices	Lysine (K)
Gly-PseAAC	Acc(%)	68.69(± 0.92)
BPB_GlySite		69.63(± 0.74)
Proposed Predictor		95.94(± 0.54)
Gly-PseAAC	Sn(%)	57.48(± 1.75)
BPB_GlySite		63.68(± 1.40)
Proposed Predictor		98.20(± 0.00)
Gly-PseAAC	Sp(%)	74.30(±1.50)
BPB_GlySite		72.60(± 0.65)
Proposed Predictor		90.67(±1.07)

The table II contains the comparative performance of the previously developed predictors BPB_GlySite[15] and Gly-PseAAC [14] with our proposed method. For comparing the performance among them, Acc, Sn and Sp has been considered. The values of Acc, Sn and Sp are calculated by running 10-fold cross validation 5 times and represented as mean(± standard deviation). The mean and standard deviation is calculated from the 5 results. The complete execution is run 5 times for the sake of getting a concrete solution.

From the table II, it is obvious that in all cases, our proposed model outperforms the existing models BPB_GlySite [15] and Gly-PseAAC [14] to a great extent. PreGly is another predictor which predicts glycated sites with an accuracy of 85.51% [13]. Our proposed method has the accuracy of 95.94(± 0.54)%, which is noticeably higher compared to the PreGly [13] predictor. The first predictor, NetGlycate [12] developed in 2006, provides only information of Matthews correlation coefficient of 0.58, but does not provide any information about Acc, Sn or Sp so that we can compare with our proposed method. The tools developed after the NetGlycate [12] are better in terms of performance. According to the performance analysis, our proposed technique for feature extraction is performing superior compared to the tools developed after NetGlycate [12]. So, we can come to a statement that, our proposed technique for feature extraction is better in terms of accuracy, specificity and sensitivity compared to the all previously developed tools NetGlycate [12], PreGly [13], Gly-PseAAC [14] and BPB_GlySite [15].

V. CONCLUSION

The BPB_GlySite has used bi-profile bayes feature extraction whereas we have used a new feature extraction technique called peptide sequence evolution based feature. Other than the feature extraction technique, the dataset and the classifier (Support Vector Machine) are same in both of the models. As the performance is better on our proposed method, a conclusion can be drawn that the peptide sequence evolution based feature extraction technique preserves more information than the bi-profile bayes feature extraction

technique for separating the positive and negative classes for the dataset of glycation sites.

VI. References

[1] Y. Xu, J. Ding, L. Wu and K. Chou, "iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition", *PLoS ONE*, vol. 8, no. 2, p. e55844, 2013.

[2] C. Walsh, S. Garneau-Tsodikova and G. Gatto, "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications", *Angewandte Chemie International Edition*, vol. 44, no. 45, pp. 7342-7372, 2005.

[3] E. Witze, W. Old, K. Resing and N. Ahn, "Mapping protein post-translational modifications with mass spectrometry", *Nature Methods*, vol. 4, no. 10, pp. 798-806, 2007.

[4] A. Miller, D. Hambly, B. Kerwin, M. Treuheit and H. Gadgil, "Characterization of Site-Specific Glycation During Process Development of a Human Therapeutic Monoclonal Antibody", *Journal of Pharmaceutical Sciences*, vol. 100, no. 7, pp. 2543-2550, 2011.

[5] S. Cho, G. Roman, F. Yeboah and Y. Konishi, "The Road to Advanced Glycation End Products: A Mechanistic Perspective", *Current Medicinal Chemistry*, vol. 14, no. 15, pp. 1653-1671, 2007.

[6] A. Lapolla, D. Fedele, L. Martano, N. Arico', M. Garbeglio, P. Traldi, R. Seraglia and D. Favretto, "Advanced glycation end products: a highly complex set of biologically relevant compounds detected by mass spectrometry", *Journal of Mass Spectrometry*, vol. 36, no. 4, pp. 370-378, 2001.

[7] N. Ahmed, R. Babaei-Jadidi, S. Howell, P. Beisswenger and P. Thornalley, "Degradation products of proteins damaged by glycation, oxidation and nitration in clinical type 1 diabetes", *Diabetologia*, vol. 48, no. 8, pp. 1590-1603, 2005.

[8] e. Ling X, "Immunohistochemical distribution and subcellular localization of three distinct specific molecular structures of advanced glycation end products in... - PubMed - NCBI", *Ncbi.nlm.nih.gov*, 2018.

[9] S. Agalou, "Profound Mishandling of Protein Glycation Degradation Products in Uremia and Dialysis", *Journal of the American Society of Nephrology,* vol. 16, no. 5, pp. 1471-1485, 2005.

[10] Q. Zhang, J. Ames, R. Smith, J. Baynes and T. Metz, "A Perspective on the Maillard Reaction and the Analysis of Protein Glycation by Mass Spectrometry: Probing the Pathogenesis of Chronic Disease", *Journal of Proteome Research*, vol. 8, no. 2, pp. 754-769, 2009.

[11] P. Thornalley and N. Rabbani, "Detection of oxidized and glycated proteins in clinical samples using mass spectrometry — A user's perspective", *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1840, no. 2, pp. 818-829, 2014.

[12] M. Johansen, L. Kiemer and S. Brunak, "Analysis and prediction of mammalian protein glycation", *Glycobiology*, vol. 16, no. 9, pp. 844-853, 2006.

[13] Y. Liu, W. Gu, W. Zhang and J. Wang, "Predict and Analyze Protein Glycation Sites with the mRMR and IFS Methods", *BioMed Research International*, vol. 2015, pp. 1-6, 2015.

[14] Y. Xu, L. Li, J. Ding, L. Wu, G. Mai and F. Zhou, "Gly-PseAAC: Identifying protein lysine glycation through sequences", *Gene*, vol. 602, pp. 1-7, 2017.

[15] Z. Ju, J. Sun, Y. Li and L. Wang, "Predicting lysine glycation sites using bi-profile bayes feature extraction", *Computational Biology and Chemistry*, vol. 71, pp. 98-103, 2017.

[16] P. Du and C. Xu, "Predicting multisite protein subcellular locations: progress and challenges", *Expert Review of Proteomics,* vol. 10, no. 3, pp. 227-237, 2013.

[17] K. Chou, Z. Wu and X. Xiao, "iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins", *PLoS ONE*, vol. 6, no. 3, p. e18258, 2011.

[18] K. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236-247, 2011.

[19] J. Jia, Z. Liu, X. Xiao, B. Liu and K. Chou, "iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC", *Journal of Theoretical Biology*, vol. 377, pp. 47-56, 2015.

[20] A. Schaffer, "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994-3005, 2001.

[21] N. Cristianini, J. Shawe-Taylor, A. Elisseeff and J. Kandola, "On Kernel-Target Alignment", *Papers.nips.cc,* 2018. [Online]. Available: https://papers.nips.cc/paper/1946-on-kernel-target-alignment. [Accessed: 27- Oct- 2018].

[22] Shibin Qiu and T. Lane, "A Framework for Multiple Kernel Support Vector Regression and Its Applications to siRNA Efficacy Prediction", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 190-199, 2009.

[23] X. Liu, L. Zhou, L. Wang, J. Zhang, J. Yin and D. Shen, "An efficient radius-incorporated MKL algorithm for Alzheimer's disease prediction", *Pattern Recognition*, vol. 48, no. 7, pp. 2141-2150, 2015.

[24] J. Jia, Z. Liu, X. Xiao, B. Liu and K. Chou, "iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC", *Oncotarget*, vol. 7, no. 23, 2016.

[25] J. Jia, Z. Liu, X. Xiao, B. Liu and K. Chou, "pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach", *Journal of Theoretical Biology*, vol. 394, pp. 223-230, 2016.

[26] Z. Ju, J. Cao and H. Gu, "Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC", *Journal of Theoretical Biology*, vol. 397, pp. 145-150, 2016.

[27] H. Lv, J. Han, J. Liu, J. Zheng, R. Liu and D. Zhong, "CarSPred: A Computational Tool for Predicting Carbonylation Sites of Human Proteins", *PLoS ONE*, vol. 9, no. 10, p. e111478, 2014.

[28] Y. Xu, Y. Ding, N. Deng and L. Liu, "Prediction of sumoylation sites in proteins using linear discriminant analysis", *Gene*, vol. 576, no. 1, pp. 99-104, 2016.

[29] B. Liu, Y. Liu, X. Jin, X. Wang and B. Liu, "iRSpot-DACC: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance", *Scientific Reports*, vol. 6, no. 1, 2016.

[30] Z. Liao, Y. Ju and Q. Zou, "Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest", *Scientifica*, vol. 2016, pp. 1-10, 2016.