MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 717-730

ISSN 0340 - 6253

A Discrete Hidden Markov Model for Detecting Histone Crotonyllysine Sites

Guohua Huang[1*], Wenfei Zeng

Department of Mathematics, Shaoyang University, Shaoyang, Hunan 42200, China

(Received November 14, 2015)

Abstract

Crotonyllysine is a new type of post-translational modifications that is responsible for promoter and enhancer region of gene transcription. Due to little knowledge about its sophisticated mechanism, accurate identification of crotonyllysine still remains challenging. We presented a discrete hidden Markov model to address this problem. We reached a predictive sensitivity of 0.7941 by the leave-one-out cross validation, more than those predicted by the representation-based support vector machine and random forest. The large-scale prediction confirmed most of computer-annotated crotonyllysine sites of five protein sequences in the Uniprot database. We demonstrated that disorder, physicochemical properties and position-specific distribution of amino acids around lysine appeared not to be strongly linked to crotonylation. These results and analysis indicated that it is effective for the presented method to detect crotonyllysine sites. The predicting tool is freely available for academic research at http://yun.baidu.com/share/link?shareid=442733655&uk=1460570570.

1. Introduction

Lysine Crotonylation (Kcr) is a newly identified histone post-translational modifications (PTMs) where crotonyl functional groups are added to the lysine residues of proteins. Tan *et al.* [1] reported that Kcr is presented in the eukaryotic cell from yeast to human and that Kcr is different from lysine acetylation in genomic distribution and regulation [2, 3]. These studies suggested also that histone Kcr is closely associated with active gene promoters and potential

[1] [*]To whom correspondence should be addressed: Guohua Huang (guohuahhn@163.com)

enhancers in mammalian cell genomes as well as male germ cell differentiation [1]. Tan and co-authors firstly identify crotonyllysine as a new type of PTM by using a mass spectrometry-based approach that combined analysis of histone peptides [1]. Bao *et al.* [4] used chemical proteomics approach to identify some 'eraser' enzymes for lysine crotonylated histone marks. However, the mechanism of catalyzing crotonylation by enzymes is unknown, greatly increasing difficulty in experimentally detecting crotonyllysine sites. This hinders a better understanding of the physiological roles and regulation of this PTM [3, 5]. In the past decade, many *in silico* techniques have been proposed to aid one to detect PTM sites, and achieved successes as expected. For example, Chuang *et al.* [6] achieved the accuracy of 0.687 on predicting the N-linked glycosylation sites, Chen *et al.* [7] about 0.975 on predicting Sumoylation Sites and Shi *et al.* [8] 0.8599 on predicting palmitoylation site, *etc.* Following these successful cases, we first presented a discrete hidden Markov model (DHMM) for *in silico* prediction of crotonyllysine sites. The method is based on the assumption that both crotonylated and non-crotonylated peptides are generated by two distinct DHMMs respectively. We trained two DHMMs by using crotonylaed and non-crotonylated samples respectively. For an unknown sample, we determined whether it is crotonylated according to probabilities of generating it.

2. Method and materials

2.1 Data

Crotonylated Proteins were collected from the Uniprot database (Release 2015_09) [9-12] which is a comprehensive repository dedicated to protein sequences and functional annotations. The process of collecting data was described as follows. First, we searched the Uniprot database with the keyword "crotonyllysine" and retrieved 92 manually reviewed protein sequences. Then, removing non-experimentally verified crotonyllysine sites, we got 57 unique protein sequences. Next, the sequence cluster program CD-HIT [13] was applied to reduce homology of 57 proteins sequences. The clustering parameter (cutoff) is set to 0.7. We obtained 6 unique protein sequences including 35 crotonylated sites. We slid an 11-mer window along each protein sequence and extracted peptides that center lysine, and have five residues in the upstream and downstream of it, respectively. 34 peptides undergoing the

lysine-crotonylated event were considered as positive samples and other 90 peptides as negative ones. All the samples constituted the training set. Table 1 listed all the positive and negative samples in the proteins sequences.

Table 1. Modification and non-modification sites in the training set

Protein	Crototyllysine sites	Non-crototyllysine sites
P70696	7, 13, 14, 17, 18, 22, 25, 36	26, 30, 32, 45, 48, 59, 87, 110, 118, 122
Q6DN03	6, 12, 13,16, 17, 21, 24, 35	25, 28, 29, 31, 44, 47, 58, 86, 152, 164
Q96QV6	37, 119, 120	6, 10, 14, 16, 75, 76, 96
P16403	34, 64, 85, 90, 97, 159, 168	17, 21, 22, 23, 26, 27, 46, 52, 63 75, 81, 106, 109, 110, 117, 119, 121, 122, 127, 129, 130, 136, 137, 139, 140, 148, 149, 152, 153, 156, 157, 160, 169, 172, 175, 176, 178, 181, 183, 184, 187, 191, 194, 196, 199, 201, 204, 206, 207
P68431	10, 19, 24, 28, 57	15, 37, 38, 65, 80, 116, 123
P62805	6, 9, 13	21, 32, 45, 60, 78, 80, 92

2.2 Method

The hidden Markov model (HMM) was a statistical learning algorithm which has a theoretical mathematical foundation and was thus applicable to a wide range of problems of interest, particularly to speech recognition [14]. A HMM was universally expressed as a five-element array $\lambda = (N, M, \pi, A, B)$, where N refers to the set of hidden states, M the set of observation symbols, π the initial states distribution, A the matrix of state transition probability, and B the observation symbol probabilities distribution per state . We designed the structure of the HMM as shown in Fig. 1, which have two hidden states $N = \{C, F\}$ where C and F stand for conservation and non-conservation respectively, and have 20 discrete observation symbols corresponding to twenty amino acids. The HMM was also expressed compactly as $\lambda = (\pi, A, B)$. Given observation symbol sequences (here peptides), by using E-M algorithm we may estimate the parameters, i.e. π, A, B, which corresponds to the problem 3 in the standard HMM. Assume that the positive and negative samples were respectively generated from two HMMs that have the same structure but differ in aspect of parameters. We used the positive samples to learn the positive HMM λ_1, and the negative samples to learn the negative HMM λ_2. Given a testing peptide P, we calculated the probabilities of generating it under the λ_1 and λ_2 respectively. This corresponds to the first problem in the standard HMM. The testing

sample was predicted to be positive if $\lambda_1$more likely generated P than λ_2. And it was predicted negative otherwise.

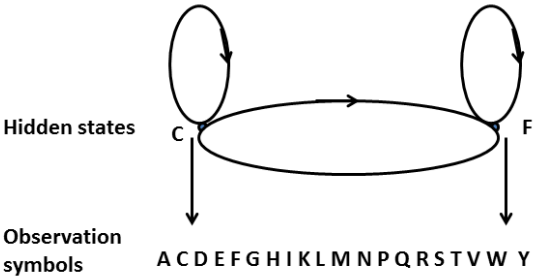

Figure 1. The structure of the presented DHHM**.** There are two hidden states and there are 20 observation symbols corresponding to 20 amino acids per state.

2.3 Cross validation and evaluation

We adopted leave-one-out cross-validation to examine the presented method. In the leave-one-out cross validation, the training set are classified into *n* parts (*n* is the number of samples in the training set), each one of which in turn serves a testing role and the other *n*-1 of which serve a training role. The sensitivity (*SN*), specificity (*SP*), accuracy (*ACC*) and Matthews's correlation coefficient (*MCC*) are used to assess the predictive performance, which are computed as follows:

TP and TN correspond to the numbers of true positive and true negative samples, respectively. FP and FN are the numbers of false positive and false negative samples respectively.

Table 2. The perfomances of different methods by leave-one-out cross validation

Learning algorithm	**Representation**	***SN***	***SP***	***ACC***	***MCC***
SVM	Disorder	0.1176	0.7222	0.5565	-0.1688
	BES	0.1471	0.8778	0.6774	0.0331
	AAPP	0.7647	0.2667	0.4032	0.0320
	CKAAP	0.2647	0.8222	0.6694	0.0967
	BES+AAPP+Disorder	0.7647	0.2667	0.4032	0.0320
random forest	disorder	0.2647	0.8556	0.6935	0.1404
	BES	0.4412	0.8778	0.7581	0.3495
	AAPP	0.3235	0.9111	0.7500	0.2906
	CKAAP	0.5588	0.8889	0.7984	0.4718
	BES+AAPP+Disorder	0.2941	0.9333	0.7581	0.3027
DHMM		0.7941	0.7778	0.7823	0.5259

3. Result and discussion

The leave-one-out cross-validation performance of the DHMM on the training set was shown in Table 2. The predictive *ACC* was 0.7823, meaning that we predicted correctly 97 of 124 samples. Assume that the probability of successfully guessing one positive or negative sample was 0.5. Continually repeating such guess for 124 times is equivalent to a binomial distribution. Therefore, the probability of identifying correctly more than 97 samples is 9.2728×10^{-11}, much lower than the *ACC*. The results indicate the promising performance of the presented method.

3.1 Comparison with representation–based learning methods

Recently, a large number of approaches have been presented to predict post-translational modification sites including S-nitrosylation sites [15], sulfotyrosine sites [16], ubiquitination sites [17] and N-acetylation sites[18]. Most predictive approaches followed such a framework that peptides are first represented by numerical vectors, then the machine learning algorithm such as support vector machine (SVM) which is widely applied in the area of bioinformatics [19] and random forest are applied to learn a classifier. We called these approaches the representation-based learning methods and used them as the baseline for comparison. Different types of representation for peptides were described as follows.

3.1.1 Binary encoding scheme (BES)

BES is an intuitive representation of protein sequences. In the BES, each amino acid is encoded into a 20-dimensional binary vector. For example, Alanine (A) is represented by $(1,0,0,\ldots,0)$, Cysteine(C) by $(0,1,0,\ldots,0)$. Each sample corresponds to a 220-dimensional vector.

3.1.2 Amino acid physicochemical properties (AAPP)

AAindex [20-22] is a comprehensive repository compiling physicochemical and biochemical properties of single amino acid and amino acid pair. Atchley *et al. [23]* carried out multivariate statistical analyses on 494 amino acid attributes to produce five interpretable numeric patterns that correspond to polarity, second structure, molecular volume, codon diversity and electrostatic charge respectively. These five types of properties are widely

applied to predictions of palmitoylation sites [24], S-nitrosylation sites [15, 25] and carbamylation sites [26].

3.1.3 Disorder

Intrinsically disordered region causes unstable or flexible 3D structures. Many studies reported that some functions of protein are specific to such dynamics of structures rather than stability[27]. Therefore, disorder states of protein are used as determining factors to differentiate between post-translational modification and non-post-translational modifications [15, 24, 28]. Here, we used the VSL2 program [29] to predict disorder of peptides. Each residue corresponds to a number and thus a peptide is an 11-dimensional vector.

3.1.4 Composition of k-spaced amino acid pairs (CKAAP)

The CKAAP of protein sequence was widely used for predicting mucin-type O-glycosylation sites [30], palmitoylation sites [31], methylation sites [32], ubiquitination sites [33], pupylation sites [34] and Phosphorylation Sites [35]. Given a peptide sequence, its CKAAP is represented as occurrence frequencies of *k*-spaced amino acid pairs such as $AX_1 \cdots X_k A$, $AX_1 \cdots X_k C$, and $AX_1 \cdots X_k C$ where $X_1,\ldots,X_k$ refers to one of 20 amino acids, respectively and *k* is set to 0, 1 and 2. Therefore, each peptide corresponds to a 1200-dimensional vector.

SVM is a classical machine-learning algorithm that maximizes margins between two groups. Combining both least risks in structure and in experiences, the SVM is applicable to a widely range of problem of interests [36]. Random forest is an ensemble machine learning algorithm which comprises various decision trees [37]. The random forest has successfully been employed for predictions of phosphorylation site [38], γ-carboxylation sites [39], glycosylation sites [40] and SUMOylation sites [41]. We used the two popular algorithms across the above different representations as the baseline for comparison. The performances of the leave-one-out cross validations were listed in Table 2. Obviously, in terms of *SN* and *MCC*, the DHMM is best. Although the SVM with the BES and with CKAAP, and the Random Forest reached higher *SP*s than the DHMM, the former performed much worse than the latter in the identification of crotonylation sites. For example, the random forest with the CKAAP got a *SP* of 0.8889, but obtained a *SN* of less than 0.56. This is a seriously unbalanced performance. It is more important to identify crotonyllysine sites than to

recognize the non-crotonyllysine sites. These results indicate advantages of the HMM over the state of the art in the prediction of crotonyllysine sites.

Table 3 The computer-annotated crotonyllysine sites in the Uniprot database and in the paper

Protein Identifier	Uniprot database	the paper
P02253	34, 64, 85, 90, 97,159, 168	34, 63, 64, 85, 90, 97, 137,139, 140, 148, 149, 153, 168
P0C169	37, 119, 120, 126	6, 10, 14, 16, 119, 120
Q00729	7, 13, 14, 17, 18, 22, 25, 36	13, 14, 17, 18, 22, 25, 26, 45,59, 122
P68432	5,10, 19, 24, 28, 57	10, 15, 19, 24, 57
P62803	6, 9, 13, 17	6, 9, 13, 45

3.2 Large-scale identification of crotonyllysine sites

As mentioned previously, 92 crotonylated proteins were downloaded from the Uniprot database, 35 of which however did not contain any experimentally validated crotonylated sites. We used the sequence cluster program CD-HIT[13] to reduce homology among 35 sequences and obtained five unique protein sequences including 29 computer-annotated crotonyllysine sites (see the second column in Table 3). Using the positive and negative samples as the training set, the crotonyllysine sites predicted by the DHMM were listed in the third column of Table 3. Interestingly, we found that most of computed-annotated sites in the Uniprot database were confirmed by the presented method. Histone H1 (P02253) protein is known for binding chromatin DNA and poly (A) RNA because it is necessary for the condensation of nucleosome chains into higher-order structured fibers. H1 serves a negative regulation of transcription from RNA polymerase II promoter by ways of chromatin remodeling, nucleosome spacing and DNA methylation [10]. The manual assertion of Kcr in the histone H1 inferred by similarity include K34, K64, K85, K90, K97, K159 and K168. Except the K159, we confirmed all the lysine-crotonylated sites. Histone H2A type 1-C (P0C169) is annotated with the Kcr at the K37, K119, K120 and K126 by similar comparison to the Histone H2A type 1 (P0C0S8) where the same four positions are experimentally detected as Kcr by Tan *et al.*[1]. We omitted Kcr at K37 and K126, but successfully identified Kcr at K119 and K120. Kcr at K6, K10, K14 and K16 needs further experimental validation. We confirmed Kcr at the K13, K14, K17, K18, K22 and K25 of Histone H2B type 1-A (Q00729) and omitted only two Kcr sites (K7, K36). K10, K19, K24 and K57 of Histone H3.1 (P68432) were confirmed as crotonylated sites annotated by the DHMM. Three fourths of Kcr site of Histone H4 (P62803) by similarity inference were confirmed by the presented

method. These results suggested the presented is feasible to large-scale identification of crotonyllysine sites.

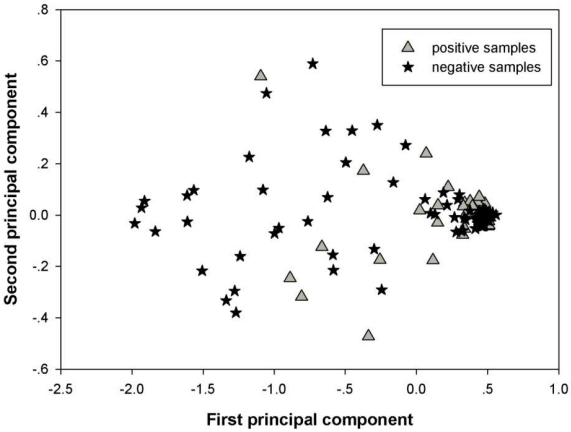

Figure 2. The scatter diagram of first two components of disorder. The first principle component explains 94.42 % of all the disorder information, and the second 4.31%.

3.3 Discussion

Crotonyllysine is a newly discovered type of histone PTMs. Compared with other well-studied PTMs such as phosphorylation and methylation, Kcr is foreign even to most researchers. Moreover, little is known about the modification mechanism of Kcr. The representation-based approaches which have been developed over the past decades provide a well-defined framework to theoretically identify and analyze PTMs, such as the nearest neighbor-based method for predicting and analyzing S-nitrosylation proposed by Li *et al.[25]* and PMeS by Shi *et al.* [42]. Following the route, we attempted to identify crotonyllysine sites and to reveal the dominant factors influencing crotonylation from disorder, physicochemical properties and position-specific distribution of amino acids. Unfortunately, the performance of these approaches is discouraged. As shown in Table 3, the disorder-based method performed worst, whether for SVM or for random forest. To demonstrate abilities of disorder to distinguish between positive and negative samples, we drawn the scatter diagram of first two principal components from the principal component analysis (PCA) shown in Figure 2. The first two principal components accounted for 0.9873 of valuable information of disorders.

Obviously, it is difficult to differentiate between positive and negative samples from the point of view of first two components, indicating that disorder appears not to be necessarily related to crotonylation. Kcr is similar with lysine acetylation in structure. Therefore, it is a natural idea to employ factors determining acetylation to identify Kcr. It however got the opposite of what one wants. Suo *et al.* [43] demonstrated a significant difference between binary encodings of acetylation and non-acetylation, Hou *et al.* [44] employed amino acid physicochemical property to identify acetylation sites, while physicochemical property and binary encode of amino acid seemed not to a dominant factor for identification of crotonyllysine sites. This is explained by the facts that lysine crotonylation substantially differs from lysine acetylation in genomic distribution and regulation [2,3]. CKAAP of amino acids characterized well palmitoylation [31], but is invalid for identification of crotonylated protein. The sophistication of Kcr and difference from most PTMs make the existing approaches more difficultly applicable to identification of it. Therefore, we adapted the DHMM instead of the representation-based leaning algorithms to identify crototyllysine sites. For a complicated system or process, the HMM is always one of best ways to represent it in practice. Our results also demonstrated this view. To investigate difference between the positive and negative DHMM, we used the web program [45] to draw a two-sample Logo of positive versus negative samples, as shown in Figure 3. Obviously, the produced observation sequences by the positive DHMM differed widely from those by the negative DHMM. At the first position in the downstream of the centered lysine, the positive DHMM is enriched with Glycine (G), while the negative DHMM depleted Arginine (R) and Proline (P). The negative DHMM depleted Lysine (K) at the second position in the downstream the centered lysine against the positive DHMM. Except the above two sites, no depleted mark was observed in the negative DHMM. On the contrary, the positive DHMM is enriched with R and Glutamine (Q) at the first position, with Leucine (L) at the third position, and with R at the fifth position in the upstream of the centered lysine and with Threonine (T) and Glutamic acid (E) at the second position, with Tyrosine (Y) and E at the third position, and with Histidine (H) at the fifth position in the downstream. These analyses support the previous hypothesis that the positive and negative samples are generated by the two widely distinct HMM and also explain the good performance of the presented method.

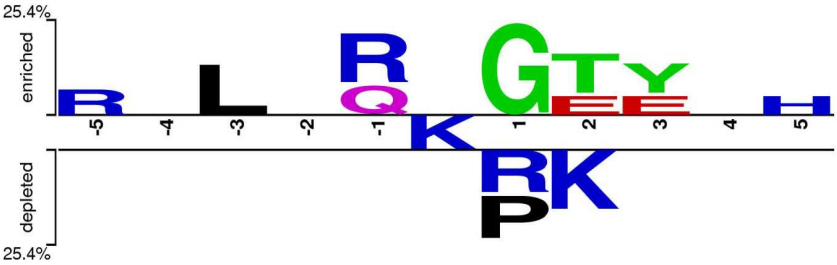

Figure 3. Two sample Logo of 34 positive versus 90 negative samples. Only amino acid residues are significantly enriched or depleted (P-value o 0.05; t-test) around lysine.

3.4 CrotPred software

The software CrotPred for predicting crotonyllysine sites is implemented by the matlab program language with the aid of the HMM toolbox for matlab which is available at http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html. As shown in Figure 4, the software may run at any computers on the Windows platform where Matlab is installed and is very easy to use. One may either enter protein sequences or open a sequence file in the fasta format on the staring interface of the CrotPred, and then select a path to save the predictive results. Clicking the button "Predict", one obtained information about crotonylation sites. Time for which the software runs depends on the number of predicted protein sequences.

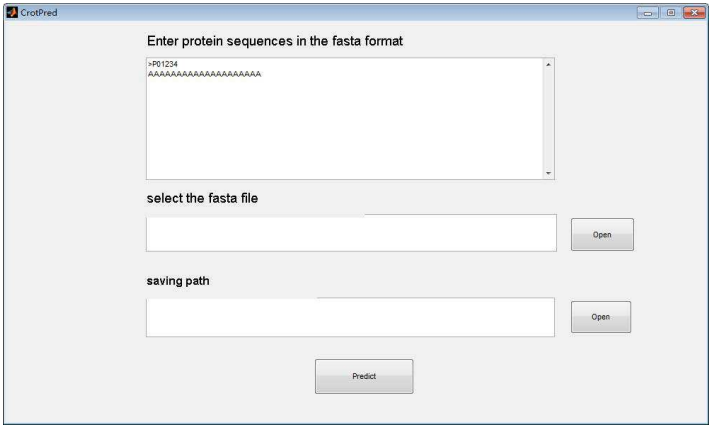

Figure 4. The user graphical interface of the software CrotPred

Conclusion

Crotonyllysine is one type of histone PTMs. The regulating role and modification mechanism

about it is unclear. We explored potential dominant factors influencing crotonylation and found that the disorder state of proteins or regions and amino acid physicochemical properties seemed not to significantly be associated with it. Therefore, we used the well-studied DHMM instead of representation-based learning algorithms to characterize crotonyllysine sites. The performance demonstrated the efficiency of the presented method. In addition, we implemented a software named CrotPred which is available at http://yun.baidu.com/share/link?shareid=442733655&uk=1460570570.

Acknowledgements: This work is in part supported by Scientific Research Fund of Hunan Provincial Education Department (15B216), by Scientific Research Fund of Shaoyang Science and Technology Dureau (2015NC44), by Scientific Research Fund of Hunan Provincial Science and Technology Department (2014FJ3013, 2014FJ3106) and by Hunan Natural Science Foundation (14JJ7078).

Reference

[1] M. Tan, H. Luo, S. Lee, F. Jin, J. S. Yang, E. Montellier, T. Buchou, Z. Cheng, S. Rousseaux, N. Rajagopal, Z. Lu, Z. Ye, Q. Zhu, J. Wysocka, Y. Ye, S. Khochbin, B. Ren, Y. Zhao, Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification, *Cell* **146** (2011) 1016–1028.

[2] M. Muers, Chromatin: a haul of new histone modifications, *Nat. Rev. Genet.* **12** (2011) 744–744.

[3] C. H. Kim, M. Kang, H. J. Kim, A. Chatterjee, P. G. Schultz, Site–specific incorporation of ε–N-crotonyllysine into histones, *Angew. Chem. Int. Ed.* **51** (2012) 7246–7249.

[4] X. Bao, Y. Wang, X. Li, X. M. Li, Z. Liu, T. Yang, C. F. Wong, J. Zhang, Q. Hao, X. D. Li, Identification of 'erasers' for lysine crotonylated histone marks using a chemical proteomics approach, *Elife* **3** (2014) #e02999.

[5] C. A. Olsen, Expansion of the lysine acylation landscape, *Angew. Chem. Int. Ed.* **51** (2012) 3755–3756.

[6] G. Y. Chuang, J. C. Boyington, M. G. Joyce, J. Zhu, G. J. Nabel, P. D. Kwong, I. Georgiev, Computational prediction of N-linked glycosylation incorporating structural properties and patterns, *Bioinformatics* **28** (2012) 2249–2255.

[7] Y. Z. Chen, Z. Chen, Y. A. Gong, G. Ying, SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties, *PLoS ONE* **7** (2012) #e39195.

[8] S. P. Shi, X. Y. Sun, J. D. Qiu, S. B. Suo, X. Chen, S. Y. Huang, R. P. Liang, The prediction of palmitoylation site locations using a multiple feature extraction method, *J. Mol. Graph. Model.* **40** (2013) 125–130.

[9] C. UniProt, Activities at the Universal Protein Resource (UniProt), *Nucleic Acids Res.* **42** (2014) D191–D198.

[10] M. Magrane, U. Consortium, UniProt knowledgebase: a hub of integrated protein data, Database (Oxford) 2011 (2011) bar009.

[11] C. UniProt, The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res.* **38** (2010) D142–D148.

[12] C. UniProt, Update on activities at the Universal Protein Resource (UniProt) in 2013, *Nucleic Acids Res.* **41** (2013) D43–D47.

[13] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics* **22** (2006) 1658–1659.

[14] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceed. IEEE* **77** (1989) 257–286.

[15] G. Huang, L. Lu, K. Feng, J. Zhao, Y. Zhang, Y. Xu, N. Zhang, B. Q. Li, W. Huang, Y. D. Cai, Prediction of S-nitrosylation modification sites based on kernel sparse representation classification and mRMR algorithm, *BioMed Res. Int.* **2014** (2014) #438341.

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

[17] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, Y. Li, Prediction of lysine ubiquitination with mRMR feature selection and analysis, *Amino Acids* **42** (2012) 1387–1395.

[18] T. Y. Lee, J. B. Hsu, F. M. Lin, W. C. Chang, P. C. Hsu, H. D. Huang, N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites, *J. Comput. Chem.* **31** (2010) 2759–2771.

[19] A. G. Bari, M. R. Reaz, B. S. Jeong, Effective DNA encoding for splice site prediction using SVM, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 241–258.

[20] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, *Nucleic Acids Res.* **28** (2000) 374–374.

[21] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: amino acid index database, *Nucleic Acids Res.* **27** (1999) 368–369.

[22] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, *Nucleic Acids Res.* **36** (2008) D202–D205.

[23] W. R. Atchley, J. Zhao, A. D. Fernandes, T. Druke, Solving the protein sequence metric problem, *Proc. Natl. Acad. Sci. USA* **102** (2005) 6395–6400.

[24] L. L. Hu, S. B. Wan, S. Niu, X. H. Shi, H. P. Li, Y. D. Cai, K. C. Chou, Prediction and analysis of protein palmitoylation sites, *Biochimie* **93** (2011) 489–496.

[25] B. Q. Li, L. L. Hu, S. Niu, Y. D. Cai, K. C. Chou, Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches, *J. Proteomics* **75** (2012) 1654–1665.

[26] G. Huang, Y. Zhou, Y. Zhang, B. Q. Li, N. Zhang, Y. D. Cai, Prediction of carbamylated lysine sites based on the one-class k-nearest neighbor method, *Mol. Biosys.* **9** (2013) 2729–2740.

[27] H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol.* **6** (2005) 197–208.

[28] L. L. Hu, Z. Li, K. Wang, S. Niu, X. H. Shi, Y. D. Cai, H. P. Li, Prediction and analysis of protein methylarginine and methyllysine based on multisequence features, *Biopolymers* **95** (2011) 763–771.

[29] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, Z. Obradovic, Length–dependent prediction of protein intrinsic disorder, *BMC Bioinformatics* **7** (2006) #208.

[30] Y. Z. Chen, Y. R. Tang, Z. Y. Sheng, Z. Zhang, Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, *BMC Bioinformatics* **9** (2008) #101.

[31] X. B. Wang, L. Y. Wu, Y. C. Wang, N. Y. Deng, Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs, *Protein Eng. Des. Sel.* **22** (2009) 707–712.

[32] W. Zhang, X. Xu, M. Yin, N. Luo, J. Zhang, J. Wang, Prediction of methylation sites using the composition of k-spaced amino acid pairs, *Protein Pept. Lett.* **20** (2013) 911–917.

[33] Z. Chen, Y. Zhou, J. Song, Z. Zhang, hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties, *Biochim. Biophys. Acta* **1834** (2013) 1461–1467.

[34] C. W. Tung, Prediction of pupylation sites using the composition of k-spaced amino acid pairs, *J. Theor. Biol.* **336** (2013) 11–17.

[35] X. Zhao, W. Zhang, X. Xu, Z. Ma, M. Yin, Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs, *PLoS ONE* **7** (2012) #e46302.

[36] V. N. Vapnik, V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[37] L. Breiman, Random forests, *Machine Learn.* **45** (2001) 5–32.

[38] B. Trost, A. Kusalik, Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights, *Bioinformatics* **29** (2013) 686–694.

[39] N. Zhang, B. Q. Li, S. Gao, J. S. Ruan, Y. D. Cai, Computational prediction and analysis of protein γ-carboxylation sites based on a random forest method, *Mol. Biosyst.* **8** (2012) 2946–2955.

[40] S. E. Hamby, J. D. Hirst, Prediction of glycosylation sites using random forests, *BMC Bioinformatics* **9** (2008) #500.

[41] A. Ijaz, SUMOhunt: Combining spatial staging between lysine and SUMO with random forests to predict SUMOylation, *ISRN Bioinformatics* **2013** (2013) 1–11.

[42] S. P. Shi, J. D. Qiu, X. Y. Sun, S. B. Suo, S. Y. Huang, R. P. Liang, PMeS: prediction of methylation sites based on enhanced feature encoding scheme, *PLoS ONE* **7** (2012) #e38772.

[43] S. B. Suo, J. D. Qiu, S. P. Shi, X. Y. Sun, S. Y. Huang, X. Chen, R. P. Liang, Position-specific analysis and prediction for protein lysine acetylation based on multiple features, *PLoS ONE* **7** (2012) #e49108.

[44] T. Hou, G. Zheng, P. Zhang, J. Jia, J. Li, L. Xie, C. Wei, Y. Li, LAceP: lysine acetylation site prediction using logistic regression classifiers, *PLoS ONE* **9** (2014) #e89575.

[45] V. Vacic, L. M. Iakoucheva, P. Radivojac, Two sample logo: a graphical representation of the differences between two sets of sequence alignments, *Bioinformatics* **22** (2006) 1536–1537.