Send Orders for Reprints to reprints@benthamscience.net

194 *Current Genomics,* 2020, *21,* 194-203

REVIEW ARTICLE

Recent Development of Machine Learning Methods in Microbial Phosphorylation Sites

Md. Mamunur Rashid[1], Swakkhar Shatabda[2], Md. Mehedi Hasan[1,3,*] and Hiroyuki Kurata[1,4,*]

[1]*Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan;* [2]*Department of Computer Science and Engineering, United International University, Plot-2, United City, Madani Avenue, Badda, Dhaka, 1212, Bangladesh;* [3]*Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan;* [4]*Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan*

ARTICLE HISTORY

Received: January 16, 2020
Revised: April 12, 2020
Accepted: April 13, 2020

DOI:
10.2174/1389202921666200427210833

Abstract: A variety of protein post-translational modifications has been identified that control many cellular functions. Phosphorylation studies in mycobacterial organisms have shown critical importance in diverse biological processes, such as intercellular communication and cell division. Recent technical advances in high-precision mass spectrometry have determined a large number of microbial phosphorylated proteins and phosphorylation sites throughout the proteome analysis. Identification of phosphorylated proteins with specific modified residues through experimentation is often labor-intensive, costly and time-consuming. All these limitations could be overcome through the application of machine learning (ML) approaches. However, only a limited number of computational phosphorylation site prediction tools have been developed so far. This work aims to present a complete survey of the existing ML-predictors for microbial phosphorylation. We cover a variety of important aspects for developing a successful predictor, including operating ML algorithms, feature selection methods, window size, and software utility. Initially, we review the currently available phosphorylation site databases of the microbiome, the state-of-the-art ML approaches, working principles, and their performances. Lastly, we discuss the limitations and future directions of the computational ML methods for the prediction of phosphorylation.

Keywords: Microbial phosphorylation, post-translational modifications, feature encoding, machine learning, mycobacterial organisms, proteome analysis.

1. INTRODUCTION

Phosphorylation is an important and most common post-translational modification (PTMs) of proteins, which plays an important role in various aspects of biological processes including cell signaling and gene regulatory functions in both eukaryotes and microbial organisms [1-9], whereas several diseases occur due to abnormal phosphorylation events and different kinase modifications [1, 10, 11]. The phosphorylation events resulting in dysregulation of protein kinases involve a potential signaling mechanism associated with various complex diseases, including cancer development and progression [12]. For example, p53 is a protein critically responsible for tumor suppression, in which multi-site PTMs have been observed, suggesting extensive control of this protein [13]. Due to its pivotal role in various biological cellular processes, the molecular networks of protein phosphorylation in eukaryotes have been studied extensively [14-16]. Phosphorylation studies in mycobacterial organisms have currently demonstrated their critical importance in diverse cellular processes and pathogenesis [17-20]. Since there are numerous unmet clinical needs in bacterial infectious diseases, it is important to study bacterial protein phosphorylation comprehensively [1].

In the last decades, low-throughput experimental techniques were primarily applied to discovering novel phosphorylation sites, but executing these techniques is time-consuming and labor-intensive [21]. Recently, advanced proteome-based high-throughput mass spectrometry methods have greatly accelerated the identification of novel phosphorylation sites [22, 23], which have determined a large number of microbial phosphorylated substrates and PTM sites [15]. With a rapid increase of protein data *via* high-throughput sequencing techniques, it has been anticipated that the number of potential phosphorylation sites increases. This high-throughput method has several limitations: for a given phosphorylation site with specific modified residues, it is unable to identify the responsible protein kinases for such phosphorylation events [3]; it is difficult to pinpoint the exact phosphorylation sites by handling the existing technical challenges [23]; and experimentation processes mostly require expensive types of equipment and often labor-

*Address correspondence to these authors at the Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan; Tel: +81-948-297-828;
E-mail: hasan.md-mehedi922@mail.kyutech.jp and Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan; Tel: +81-948-297-828; E-mail: kurata@bio.kyutech.ac.jp

1389-2029/20 $65.00+.00

intensive, which are not available in ordinal laboratories [3, 16, 24, 25]. To identify novel phosphorylation sites by solving these limitations, machine learning (ML)-based approach has become increasingly popular. Hence, current efforts have primarily been made to develop computational methods, particularly ML-based approaches, to precisely identify the bacterial protein phosphorylation sites, but a limited number of computational tools have been presented so far [16, 24-26].

ML algorithms could greatly reduce the costs and labors in detecting potential phosphorylation sites with existing experimental data [16, 27, 28]. This review summarizes the ML-based computational resources with available databases, general or species-specific prediction models, and kinase-specific prediction models for bacterial proteins. It also discusses the benefits and limitations associated with the ML-based approaches. Therefore, this review can assist scientists to select the best predictor of bacterial phosphorylation sites and suggests the future directions of the ML methods.

2. CURRENT PHOSPHORYLATION DATABASES

In Table **1**, we have listed several protein phosphorylation site databases of mycobacteria, namely Phosphorylation Site Database [29], dbPTM 3.0 [30], PHOSIDA [31], UniProt [32], and SysPTM 2.0 [33] containing 1400, 186, 305, 176, and 348 phosphorylation sites, respectively. To date, the dbPSP has been the most updated available phosphorylation site database for microbes, which was constructed by Pan *et al.* [15] in 2015. It registers 3750 distinct phosphorylated proteins with 7391 phosphorylation sites on different amino acid residues containing arginine (Arg), cysteine (Cys), aspartic acid (Asp), tyrosine (Tyr), serine (Ser), threonine (Thr), and histidine (His) from 96 organisms. This database serves as an extensive data resource for studying bacterial phosphorylation.

3. HOMOLOG REDUNDANCY

In PTM analysis, the curated sequences are often affected by homology and redundancy problems. Therefore, sequence redundancy elimination or homology reduction is a prerequisite to decipher the overfitting problem on the datasets. To shrink the homology sequences, most of phosphorylation prediction tools castoff the flanking sequence windows or protein sequences by using the program CD-HIT (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit) [32] or BLASTCLUST (http://nebc.nox.ac.uk/bioinformatics/docs/blastclust.html). However, there is no cutoff standard program to filter the high sequence similarity. Therefore, a rigorous investigation on the benchmark dataset is essential to build a precise prediction model.

4. PROTEIN ENCODING SCHEME

ML-algorithms are not able to directly handle sequence data, which need to be transformed into numeric feature vectors using different encoding methods [16, 34-59]. Three popular feature encoding schemes, consisting of evolutionary, sequence composition, and structural features, are used for predicting microbial phosphorylation sites (Fig. **1**).

4.1. Evolutionary Features

The evolutionary profile is generated from the position-specific scoring matrix (PSSM) by using PSI-BLAST with different constraints including e-value and iteration times [60, 61]. Recently, different potential evolutionary schemes have been generated, including the amino acid composition of PSSM, tri-gram PSSM, dipeptide composition of PSSM, [62-65]. The MPSite predictor has introduced different evolutionary features for bacterial phosphorylation site prediction [16].

4.2. Sequence Composition-based Features

Different types of sequence composition encoding approaches were used, including amino acid frequency composition, amino acid composition (AAC), amino acid index properties (AIP), and binary encoding, for bacterial phosphorylation site prediction (Table **2**). Amino acid location encoding is widely used in the field of PTM research [16, 25], where a sequence fragment is encoded into a feature vector by replacing any of the 20 native amino acids with numerical indexes. The dimension of the feature vector depends on the length of the sequence fragment. The composition of k-spaced amino acid pairs (CKSAAP) is widely used in PTMs research [24, 66]. Binary encoding is another common feature [16, 67-69].

Table 1. List of currently available protein phosphorylation site databases in mycobacteria.

Database	Number of Phosphorylation Sites/ Total Proteins	Year	Database URL	References
dbPSP	7391/3750	2015	**http://dbpsp.biocuckoo.org/**	[15]
SysPTM 2.0	348/213	2014	**http://lifecenter.sgst.cn/SysPTM/**	[15, 33]
UniProt	176/135	2014	**http://www.uniprot.org/**	[15]
PHOSIDA	305/382	2010	**http://www.phosida.com**	[15, 31]
dbPTM 3.0	186/138	2006	**http://dbPTM.mbc.nctu.edu.tw/**	[15, 30]
Phosphorylation Site Database	1400/960	2004	**http://vigen.biochem.vt.edu/xpd/xpd.htm** (Not available)	[15, 29]

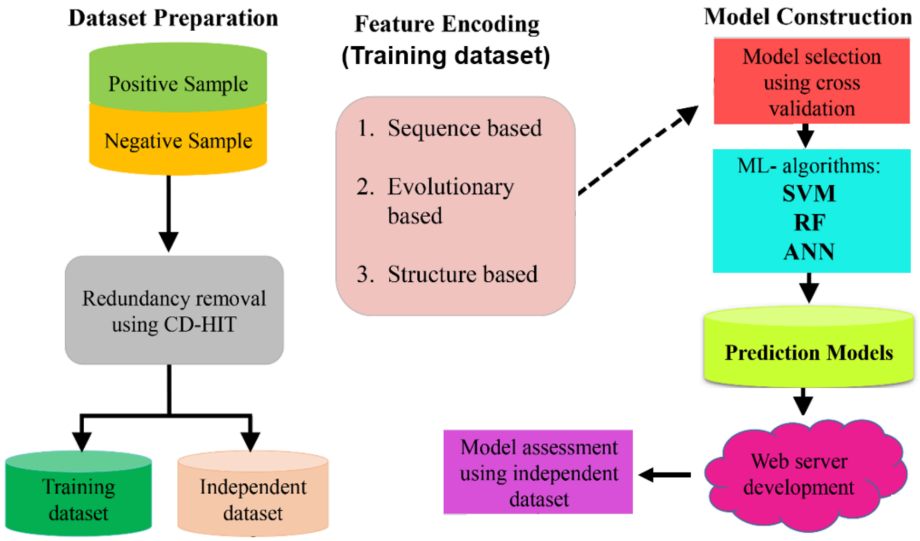

Fig. (1). An overview of the general framework of machine learning (ML) based computational approaches for phosphorylation sites prediction. Generally, the construction of ML-approaches roughly consists of the following 5 steps: (i) dataset preparation; (ii) selection encoding methods; (iii) building prediction models; (iv) performance evaluation; and (v) development of a web-server. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

4.3. Structure Features

The function of proteins critically depends on their tertiary structures and secondary structures. The α-helix, β-strand, and coil were used to build the native protein structures. It plays an important role in the interaction of the residues inside the proteins [65, 70, 71]. For further investigation of bacterial phosphorylation sites, secondary and tertiary structure information could be integrated [59, 72].

5. MACHINE LEARNING ALGORITHMS

We reviewed existing ML-based bacterial phosphorylation site predictors including our MPSite [16], prkC-PSP [25], cPhosBac [24], and NetPhosBac [26] and compared their key aspects. As mentioned in section 4, most of the developed computational tools (Tables **2** and **3**) have been constructed by using the ML algorithms. Based on our survey, the ML classifiers that predict bacterial phosphorylation sites are described below.

5.1. Support Vector Machine

The SVM is a popular supervised classification algorithm and widely used in protein bioinformatics research for classifying biological data. It aims to find the optimal hyperplane with the largest margin to accurately classify samples based on the feature dimensionality of the training dataset [73-75]. For the computational purpose, provided sequences are filtered, converted into the representative fixed-length feature vector and subjected to an objective function (class labels: phosphorylation site: 1 and non-phosphorylation site: 0). The generating mapping formula for SVM learns a function in the form of,

$$y = \text{sign}\left(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b\right) \quad (1)$$

where y stands for the predicted class relative to the input feature vector of x; a_i is the modifiable weight coefficients associated with the sample x_i; b is the bias term which we target to maximize; K represents the suitable kernel function. So once a test dataset is given, features associated with the data are mapped to a high-dimensional space. Their class is predicted based on the sign by applying equation 1, such that if the sign is positive (+) y belongs to class 1; if the sign is negative (-) y belongs to the other class. It is worth mentioning that based on different computational scenarios several kernel functions are available for SVM, including Gaussian radial-basis function (RBF), linear/polynomial functions, and sigmoid functions. SVM is widely used in many bioinformatics studies [43, 76, 77]. The RBF kernel function was commonly employed, but it is important to make a better choice of kernel approaches according to needs and questions of interest [43]. Another important point is to choose the controlling parameters. In the SVM algorithm, two critical parameters are C (the penalty factor that controls the trade-off between the training error and margin) and γ (the parameter that configures the kernel function) [25, 77]. Since variation in parameter configurations could significantly change prediction accuracy, the parameters should be optimized by the cross-validation test using a grid search approach to obtain the best performance.

To date, only four ML-based predictors for bacterial phosphorylation sites have been available. Interestingly, 2 out of 4 methods used SVM [24, 25] (Tables **2** and **3**). In 2015, Li *et al.* [24] retrieved a phosphorylation dataset from NetPhosBac [26], containing 152 of pS/pT sites in 199 substrates (while 90% sequence identity were confirmed by CD-HIT) [78], and proposed a predictor of cPhosbac employing a SVM-based ML algorithm [24]. The method generated 2205 dimensional feature vector based on the CKSAAP methods. They have shown that the cPhosBac achieved high prediction accuracy compared to the NetPhosBac [26]. In 2018, Zhang *et al.* developed a new online prediction tool

Table 2. List of currently available machine-learning tools for bacterial phosphorylation sites prediction.

Tool (Year)	PTM Residue	ML Algorithm	Feature Encoding	Dataset Size (Positive/Negative)		Homolog and Redundancy	Window Size	References
				Training dataset	*Independent dataset*			
MPSite (2019)	Serine Threonine	RF	Evolutionary, sequence composition, and structure features	S: 1704/ 3408 T: 1401/ 2802	S: 341/ 682 T: 254/ 508	30% 30%	21	[16]
prkC-PSP (2018)	Serine Threonine	SVM	Amino acid location	36/512		-	31	[25]
cPhosBac (2015)	Serine Threonine	SVM	CKSAAP	152/5761		-	21	[24]
NetPhosBac (2008)	Serine Threonine	Neural network	sequence composition features	152/841		90%	13	[26]

'- ', means not available.

Table 3. Detail of the available web server for bacterial phosphorylation sites prediction.

Predictor	Description	Predictor URL	References
MPSite	Web-based machine learning predictor for identifying bacterial phosphorylation sites using the sequence features. This is a non-organism specific or general phosphorylation site predictor.	http://kurata14.bio.kyutech.ac.jp/MPSite/	[16]
prkC-PSP	Web application for identification of prkC-specific phosphorylation sites in bacteria based on sequence information.	http://free.cancerbio.info/prkc/	[25]
cPhosBac	Web application to predict phosphorylation sites in bacteria. It includes protein and motif length scan to optimize the prediction.	http://netalign.ustc.edu.cn/cphosbac/	[24]
NetPhosBac	First web-based bacterial phosphorylation site predictor based on the sequence information. It is a taxa-specific predictor.	http://www.cbs.dtu.dk/services/NetPhosBac	[26]

employing the SVM methods, called prkC-PSP [25]. Basically, this is the first kinase (prkC)-specific phosphorylation site predictor in bacterial organisms. Since the kinase-specific dataset for bacterial phosphorylation is not available, they constructed a prkC kinase-related dataset of 36 phosphorylation sites in 14 experimentally validated protein substrates by curating the published literature. The prkC-PSP predictor used the amino acid location feature extraction method for encoding input features and achieved 94.89% accuracy.

5.2. Random Forest

Random forest (RF) [79] is one of the well-known and widely applied ML-algorithms to address various bioinformatics applications [16, 43, 67, 80-90]. RF essentially consists of a large number of N individual decision trees to operate as an ensemble learning algorithm [79]. For a given training sample of size T with Q features, RF builds Q subsets of training datasets by utilizing the bootstrap sampling, and then at each node randomly T features are selected to find the best split according to Gini impurity. Usually, the average prediction performance is reported to avoid overfitting problems. In RF, each decision tree consists of a number of 'if then rules' which are fairly simple to provide potential insight and knowledge to biologists. It is worth mentioning that the RF algorithm has three most important parameters: the number of decision trees; the number of variables that are randomly selected in each node partition; and optimization is necessary to minimize the number of samples to split internal nodes.

In 2019 [16], we developed a web-based bacterial phosphorylation site predictor called 'MPsite' (Table **2**) using the enhanced characteristics of sequence features. To date, this has been only one general or non-species specific method for predicting microbial phosphorylation sites. In the MPsite [16], 2709 pS sites in 1940 proteins, and 2174 pT sites in 1534 proteins were collected from the dbPSP database [15]. From different feature encoding methods, the Wilcoxon rank-sum test was employed to select the key features. Finally, the optimized features were used to train the RF classifi-

er. The MPsite achieved promising performance compared to the existing predictor NetphosBac.

5.3. Artificial Neural Network

Artificial neural network (ANN or NN) is well established and widely used in bioinformatics research [91-95]. ANN is a machine learning approach inspired by biological neural networks (the central nervous system of the animal at brain). Generally, ANN receives multiple input nodes, connect these inputs with their multiple internal states to generate the outputs using an output function. Each connection is assigned a weight coefficient which indicates its relative importance [96]. Generally, there are three important layers of ANN: the input layer; hidden layers; and the output layer [97]. Computationally, ANN is complex in addressing the problem of multiple hidden layers [98, 99]. Among the four reviewed predictors (Table **2**) the NetPhosBac applied the ANN algorithm as their core method [26].

Besides the proposed operational framework, it is important to discuss their pros and cons, as noted below: (i) since overestimation is a major issue in ML-based methods and benchmark dataset-based performance is often subject to high risks of overfitting, hence the evaluation of the developed models by using independent dataset other than the benchmark dataset is essential; (ii) in general, web-based predictors are useful to detect putative phosphorylation sites and to develop further advanced tools. In this case, the details of the publicly available tools are listed in Table **3**.

6. CURRENT MACHINE LEARNING TOOLS FOR PHOSPHORYLATION SITE PREDICTION

The development and proliferation of ML-based computation approaches to the prediction of phosphorylation PTM have been witnessed in recent decades. ML methods can be selected or designed with respect to training-test datasets, sequence/structural descriptors, targeted phosphorylation types, applied physiochemical properties, feature encoding techniques, *etc*. In this study, we explored the latest ML-based tools for predicting novel phosphorylation sites in bacterial organisms. In general, the development of the ML-based computation methodology consists of roughly five steps: (i) preparation of high-quality training dataset and independent dataset; (ii) extraction of informative features by suitable encoding schemes; (iii) construction of prediction models using ML-algorithms; (iv) performance evaluation of the models, and (v) web-server development (Fig. **1**). This methodology is well established and widely used in computational protein biology and bioinformatics fields [16, 25, 34-43]. In computational biology, identifying phosphorylation sites accurately on a given protein is still a major challenge. From our review, the existing ML-based phosphorylation site predictors can be described in three categories: organism-specific, general, and kinase-specific modes [27].

7. GENERAL OR ORGANISM SPECIFIC PHOSPHORYLATION SITE PREDICTOR

Protein phosphorylation analysis in eukaryotes has almost matured over the past few decades [3], but it is still countable in bacterial organisms [16]. To predict non-specific or general phosphorylation sites in microbes, Hasan *et al.* in 2019 [16] developed the first online-based ML predictor, namely MPSite with a random forest (RF) classifier, which predicts phosphorylated serine (pS) and phosphorylated threonine (pT) residues on the targeted protein sequences (Tables **2** and **3**). It is well known that the proteins of each species have a distinct substrate structure for the attachment of different protein kinases (PKs). Thus, the prediction accuracy could be improved by designing the ML-based predictors in an organism-specific manner. In 2008, Miller *et al.* prepared the training dataset consisting of 103 phosphorylated serine sites (pS) and 37 phosphorylated threonine sites (pT), and developed the first bacterial-specific online predictor of NetPhosBac 1.0 [26] (Tables **2** and **3**) by implementing an artificial neural network (ANN) algorithm. Li *et al.* [24] in 2015 retrieved the same dataset of pS and pT from NetPhosBac, and constructed a predictor of cPhosbac using a support vector machine (SVM) algorithm. The cPhosBac achieved higher prediction performance than the NetPhosBac predictor (Tables **2** and **3**).

8. KINASE-SPECIFIC PHOSPHORYLATION SITE PREDICTOR

Currently, a number of studies have reported that kinase-specific phosphorylation plays an important role in various cellular activities that are inherently responsible for bacterial pathogenicity [1, 20, 44]. In bacteria, several recent studies have identified that the threonine/serine protein kinase, known as prkC, shows homology in catalytic domains. [45-47]. Further studies found that prkC is often involved in the various cellular process including bacterial resuscitation [48], gliding motility [49], and antimicrobial resistance [47]. Kinase has become one of the largest 'druggable groups' in cancer therapeutics in recent years [12]. Although numerous ML-based predictors, such as GPS, Scansite, PKIS, and PPSP, have been proposed to detect the kinase-specific phosphorylation sites in eukaryotes [7, 25, 27, 50-52], predictors for bacteria remain to be developed. In this regard, Zhang *et al.* recently have constructed a prkC kinase-related dataset of 36 phosphorylation sites in 14 experimentally validated proteins. They developed the first kinase (prkC)-specific web-application, prkC-PSP in 2018 (Table **2**), using the SVM algorithm [25].

In recent years, protein kinases have become an important group of 'druggable' targets [10, 12]. To connect protein kinases or phosphorylated proteins to drug design and potential biomarker identification, several computational approaches were developed. In particular, unsupervised cluster analysis was used for phosphoproteomics profiling of kinases. The Wilcoxon rank-sum test was used to select important features and a linear kernel-based SVM algorithm was employed to build the final classifier [53, 54]. Recently, Leung *et al.* [12, 55] developed a command-line-tool called HyperModules to detect clinically and phenotypically related network modules for the discovery of disease mutations biomarkers.

9. CAVEATS OF THE STATE-OF-THE-ART ML APPROACHES

Even though great progress has been made in the development of phosphorylation site prediction tools, several chal-

lenges and limitations need to be addressed. Firstly, the prediction accuracy evaluated by cross-validation test is difficult to reproduce, unless the ML parameters and source codes regarding feature encodings are provided. On the other hand, the prediction performances will be reproduced on independent datasets, if a developer provides a standalone program or web application. Unfortunately, many published methods neither open their assigned source codes nor datasets, which delays the development of next-generation methods. Therefore, it is highly recommended to provide datasets and source codes while publishing a new methodology [100]. Secondly, most of the publicly available methods used their own independent dataset to evaluate the prediction performance in comparison with existing methods. For fair comparisons, the construction of a unique or independent dataset is essential. While constructing distinct datasets, care should be taken that none of the sequences overlap with the benchmark dataset.

CONCLUSION AND PERSPECTIVES

Recently, the field of bacterial phosphorylation site detection has made noticeable progress in recruiting the ML-approaches, as mentioned in Tables **2** and **3**. Owing to high-throughput sequencing techniques, automated computational approaches are required to enable rapid and accurate prediction phosphorylation sites related to kinases from a large number of candidate proteins. In this regard, several ML-based approaches have been developed in both the sequence-based and structure-based classes; many predictors were built in the kinase-specific and organism-specific/general manners by using a variety of training and test dataset resources [3, 27]. In order to develop the next-generation methodology, the following challenges could be explored [27]. First, a reliable, high-quality benchmark dataset is constructed by carefully searching existing phosphorylation site databases and through rigorous literature inquiry. Second, most of the existing feature descriptors are extracted from primary sequences. On the other hand, many functional sites were found based on evolutional and structural information [101-103]. The addition of structure-based and evolutionary information of protein kinases proves valuable to improve the predictors [104-108]. Third, predictors available for a wider variety of organisms are required, because protein kinases are disparate in different organisms [3, 109, 110]. Forth, different ML classifiers are explored to increase prediction performance. It is important not only to integrate different feature encodings [111-115], such as K-nearest neighbors, multivariate information, biochemical properties, and pseudo residues composition, but also to investigate different ML classifiers [116-122], such as an extremely randomized tree, extreme gradient boosting, light gradient boosting, and deep learning. The feature selection technique should remove redundant information to improve performance. In this regard, mRMR [123], ANOVA [124], and MRMD [125, 126] can be considered. Rapid development in structural bioinformatics and sequential bioinformatics have driven the medicinal chemistry undergoing an unprecedented revolution of proteins [127-129], in which the recently proposed encoding methods [130-135] may play an important role in discovering new microbial phosphorylation sites.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This study was supported by Japan Society for the Promotion of Science by Grant-in-Aid for Scientific Research (B) (19H04208) and by the developing key technologies for discovering and manufacturing pharmaceuticals used for next-generation treatments and diagnoses both from the Ministry of Economy, Trade and Industry, Japan (METI) and from Japan Agency for Medical Research and Development (AMED).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We thank the reviewers for their great comments in helping to improve this manuscript.

REFERENCES

[1] Lai, S.J.; Tu, I.F.; Wu, W.L.; Yang, J.T.; Luk, L.Y.P.; Lai, M.C.; Tsai, Y.H.; Wu, S.H. Site-specific His/Asp phosphoproteomic analysis of prokaryotes reveals putative targets for drug resistance. *BMC Microbiol.,* **2017**, *17*(1), 123. http://dx.doi.org/10.1186/s12866-017-1034-2 PMID: 28545444

[2] Chao, J.D.; Wong, D.; Av-Gay, Y. Microbial protein-tyrosine kinases. *J. Biol. Chem.,* **2014**, *289*(14), 9463-9472. http://dx.doi.org/10.1074/jbc.R113.520015 PMID: 24554699

[3] Trost, B.; Kusalik, A. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics,* **2011**, *27*(21), 2927-2935. http://dx.doi.org/10.1093/bioinformatics/btr525 PMID: 21926126

[4] Cohen, P. The role of protein phosphorylation in neural and hormonal control of cellular activity. *Nature,* **1982**, *296*(5858), 613-620. http://dx.doi.org/10.1038/296613a0 PMID: 6280056

[5] Wood, C.D.; Thornton, T.M.; Sabio, G.; Davis, R.A.; Rincon, M. Nuclear localization of p38 MAPK in response to DNA damage. *Int. J. Biol. Sci.,* **2009**, *5*(5), 428-437. http://dx.doi.org/10.7150/ijbs.5.428 PMID: 19564926

[6] Uddin, S.; Lekmine, F.; Sassano, A.; Rui, H.; Fish, E.N.; Platanias, L.C. Role of Stat5 in type I interferon-signaling and transcriptional regulation. *Biochem. Biophys. Res. Commun.,* **2003**, *308*(2), 325-330. http://dx.doi.org/10.1016/S0006-291X(03)01382-2 PMID: 12901872

[7] Obenauer, J.C.; Cantley, L.C.; Yaffe, M.B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.,* **2003**, *31*(13), 3635-3641. http://dx.doi.org/10.1093/nar/gkg584 PMID: 12824383

[8] Lian, I.; Kim, J.; Okazawa, H.; Zhao, J.; Zhao, B.; Yu, J.; Chinnaiyan, A.; Israel, M.A.; Goldstein, L.S.; Abujarour, R.; Ding, S.; Guan, K.L. The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Genes Dev.,* **2010**, *24*(11), 1106-1118. http://dx.doi.org/10.1101/gad.1903310 PMID: 20516196

[9] Bu, Y.-H.; He, Y.-L.; Zhou, H.-D.; Liu, W.; Peng, D.; Tang, A.-G.; Tang, L.-L.; Xie, H.; Huang, Q.-X.; Luo, X.-H.; Liao, E.Y. Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallopeptidase expression of preosteoblastic cells. *J. Endocrinol.,* **2010**, *206*(3), 271-277. http://dx.doi.org/10.1677/JOE-10-0064 PMID: 20525764

[10] Cohen, P. Protein kinases--the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.,* **2002**, *1*(4), 309-315. http://dx.doi.org/10.1038/nrd773 PMID: 12120282

[11] Roskoski, R., Jr A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacol. Res.,* **2015**, *100*, 1-23. http://dx.doi.org/10.1016/j.phrs.2015.07.010 PMID: 26207888

[12] Chen, Y.A.; Eschrich, S.A. Computational methods and opportunities for phosphorylation network medicine. *Transl. Cancer Res.,* **2014**, *3*(3), 266-278. PMID: 25530950

[13] Loughery, J.; Meek, D. Switching on p53: an essential role for protein phosphorylation? *Biodiscovery,* **2013**, *8,* e8946. http://dx.doi.org/10.7750/BioDiscovery.2013.8.1

[14] Pawson, T.; Scott, J.D. Protein phosphorylation in signaling--50 years and counting. *Trends Biochem. Sci.,* **2005**, *30*(6), 286-290. http://dx.doi.org/10.1016/j.tibs.2005.04.013 PMID: 15950870

[15] Pan, Z; Wang, B; Zhang, Y; Wang, Y; Ullah, S; Jian, R; Liu, Z; Xue, Y. dbPSP: a curated database for protein phosphorylation sites in prokaryotes. *Database,* **2015**, *2015,* bav031.

[16] Hasan, M.M.; Rashid, M.M.; Khatun, M.S.; Kurata, H. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. *Sci. Rep.,* **2019**, *9*(1), 8258. http://dx.doi.org/10.1038/s41598-019-44548-x PMID: 31164681

[17] Dworkin, J. Ser/Thr phosphorylation as a regulatory mechanism in bacteria. *Curr. Opin. Microbiol.,* **2015**, *24*, 47-52. http://dx.doi.org/10.1016/j.mib.2015.01.005 PMID: 25625314

[18] Mijakovic, I.; Macek, B. Impact of phosphoproteomics on studies of bacterial physiology. *FEMS Microbiol. Rev.,* **2012**, *36*(4), 877-892. http://dx.doi.org/10.1111/j.1574-6976.2011.00314.x PMID: 22091997

[19] Hutchings, M.I.; Hong, H.J.; Buttner, M.J. The vancomycin resistance VanRS two-component signal transduction system of *Streptomyces coelicolor*. *Mol. Microbiol.,* **2006**, *59*(3), 923-935. http://dx.doi.org/10.1111/j.1365-2958.2005.04953.x PMID: 16420361

[20] Ohlsen, K.; Donat, S. The impact of serine/threonine phosphorylation in *Staphylococcus aureus*. *Int. J. Med. Microbiol.,* **2010**, *300*(2-3), 137-141. http://dx.doi.org/10.1016/j.ijmm.2009.08.016 PMID: 19783479

[21] Meier, R.; Alessi, D.R.; Cron, P.; Andjelković, M.; Hemmings, B.A. Mitogenic activation, phosphorylation, and nuclear translocation of protein kinase Bbeta. *J. Biol. Chem.,* **1997**, *272*(48), 30491-30497. http://dx.doi.org/10.1074/jbc.272.48.30491 PMID: 9374542

[22] Huttlin, E.L.; Jedrychowski, M.P.; Elias, J.E.; Goswami, T.; Rad, R.; Beausoleil, S.A.; Villén, J.; Haas, W.; Sowa, M.E.; Gygi, S.P. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell,* **2010**, *143*(7), 1174-1189. http://dx.doi.org/10.1016/j.cell.2010.12.001 PMID: 21183079

[23] Boersema, P.J.; Mohammed, S.; Heck, A.J. Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.,* **2009**, *44*(6), 861-878. http://dx.doi.org/10.1002/jms.1599 PMID: 19504542

[24] Li, Z.; Wu, P.; Zhao, Y.; Liu, Z.; Zhao, W. Prediction of serine/threonine phosphorylation sites in bacteria proteins. *Advance in Structural Bioinformatics*; Springer, **2015**, pp. 275-285. http://dx.doi.org/10.1007/978-94-017-9245-5_16

[25] Zhang, Q.B.; Yu, K.; Liu, Z.; Wang, D.; Zhao, Y.; Yin, S.; Liu, Z. Prediction of prkC-mediated protein serine/threonine phosphorylation sites for bacteria. *PLoS One,* **2018**, *13*(10), e0203840. http://dx.doi.org/10.1371/journal.pone.0203840 PMID: 30278050

[26] Miller, M.L.; Soufi, B.; Jers, C.; Blom, N.; Macek, B.; Mijakovic, I. NetPhosBac - a predictor for Ser/Thr phosphorylation sites in bacterial proteins. *Proteomics,* **2009**, *9*(1), 116-125. http://dx.doi.org/10.1002/pmic.200800285 PMID: 19053140

[27] Xue, Y.; Gao, X.; Cao, J.; Liu, Z.; Jin, C.; Wen, L.; Yao, X.; Ren, J. A summary of computational resources for protein phosphorylation. *Curr. Protein Pept. Sci.,* **2010**, *11*(6), 485-496. http://dx.doi.org/10.2174/138920310791824138 PMID: 20491621

[28] Chen, X.; Shi, S.P.; Suo, S.B.; Xu, H.D.; Qiu, J.D. Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity. *Bioinformatics,* **2015**, *31*(2), 194-200. http://dx.doi.org/10.1093/bioinformatics/btu598 PMID: 25236462

[29] Wurgler-Murphy, S.M.; King, D.M.; Kennelly, P.J. The Phosphorylation Site Database: a guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms. *Proteomics,* **2004**, *4*(6), 1562-1570. http://dx.doi.org/10.1002/pmic.200300711 PMID: 15174126

[30] Lee, T.-Y; Huang, H.-D; Hung, J.-H; Huang, H.-Y; Yang, Y.-S; Wang, T.-H. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.,* **2006**, *34*(suppl_1), D622-D627.

[31] Gnad, F; Gunawardena, J; Mann, M. PHOSIDA 2011: the post-translational modification database. *Nucleic Acids Res.,* **2010**, *39*(suppl_1), D253-D260.

[32] Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics,* **2010**, *26*(5), 680-682. http://dx.doi.org/10.1093/bioinformatics/btq003 PMID: 20053844

[33] Li, J.; Jia, J.; Li, H.; Yu, J.; Sun, H.; He, Y.; Lv, D.; Yang, X.; Glocker, M.O.; Ma, L. SysPTM 2.0: an updated systematic resource for post-translational modification. *Database ,* **2014**, *2014, bau025.* http://dx.doi.org/10.1093/database/bau025. Print 2014

[34] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247. http://dx.doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

[35] Liu, Y.; Wang, M.; Xi, J.; Luo, F.; Li, A. PTM-ssMP: a web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *Int. J. Biol. Sci.,* **2018**, *14*(8), 946-956. http://dx.doi.org/10.7150/ijbs.24121 PMID: 29989096

[36] Hasan, M.M.; Khatun, M.S. Recent progress and challenges for protein pupylation sites prediction. *EC Proteom. Bioinformatics,* **2017**, *2*(1), 36-45.

[37] Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med. Res. Rev.,* **2020**. http://dx.doi.org/10.1002/med.21658 PMID: 31922268

[38] Song, J.; Wang, H.; Wang, J.; Leier, A.; Marquez-Lago, T.; Yang, B.; Zhang, Z.; Akutsu, T.; Webb, G.I.; Daly, R.J. PhosphoPredict: a bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci. Rep.,* **2017**, *7*(1), 6862. http://dx.doi.org/10.1038/s41598-017-07199-4 PMID: 28761071

[39] Hasan, M.M.; Khatun, M.S.; Kurata, H. A comprehensive review of *in silico* analysis for protein s-sulfenylation sites. *Protein Pept. Lett.,* **2018**, *25*(9), 815-821. http://dx.doi.org/10.2174/0929866525666180905110619 PMID: 30182830

[40] Hasan, M.M.; Zhou, Y.; Lu, X.; Li, J.; Song, J.; Zhang, Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. *PLoS One,* **2015**, *10*(6), e0129635. http://dx.doi.org/10.1371/journal.pone.0129635 PMID: 26080082

[41] Hasan, M.M.; Khatun, M.S. Prediction of protein Post-Translational Modification sites: an overview. *Ann. Proteom. Bioinform.,* **2018**, *2*, 049-055.

[42] Xu, Z.-C.; Feng, P.-M.; Yang, H.; Qiu, W.-R.; Chen, W.; Lin, H. iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics,* **2019**, *35*(23), 4922-4929. http://dx.doi.org/10.1093/bioinformatics/btz358 PMID: 31077296

[43] Chen, Z.; Liu, X.; Li, F.; Li, C.; Marquez-Lago, T.; Leier, A.; Akutsu, T.; Webb, G.I.; Xu, D.; Smith, A.I. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief. Bioinform.,* **2018**, *20*(6), 2267-2290. PMID: 30285084

[44] Cousin, C.; Derouiche, A.; Shi, L.; Pagot, Y.; Poncet, S.; Mijakovic, I. Protein-serine/threonine/tyrosine kinases in bacterial signaling and regulation. *FEMS Microbiol. Lett.,* **2013**, *346*(1), 11-19. http://dx.doi.org/10.1111/1574-6968.12189 PMID: 23731382

[45] Madec, E.; Laszkiewicz, A.; Iwanicki, A.; Obuchowski, M.; Séror, S. Characterization of a membrane-linked Ser/Thr protein kinase in *Bacillus subtilis*, implicated in developmental processes. *Mol. Microbiol.,* **2002**, *46*(2), 571-586. http://dx.doi.org/10.1046/j.1365-2958.2002.03178.x PMID: 12406230

[46] Pereira, S.F.; Goss, L.; Dworkin, J. Eukaryote-like serine/threonine kinases and phosphatases in bacteria. *Microbiol. Mol. Biol. Rev.,* **2011**, *75*(1), 192-212. http://dx.doi.org/10.1128/MMBR.00042-10 PMID: 21372323

[47] Kristich, C.J.; Wells, C.L.; Dunny, G.M. A eukaryotic-type Ser/Thr kinase in *Enterococcus faecalis* mediates antimicrobial resistance and intestinal persistence. *Proc. Natl. Acad. Sci. USA,* **2007**, *104*(9), 3508-3513. http://dx.doi.org/10.1073/pnas.0608742104 PMID: 17360674

[48] Squeglia, F.; Marchetti, R.; Ruggiero, A.; Lanzetta, R.; Marasco, D.; Dworkin, J.; Petoukhov, M.; Molinaro, A.; Berisio, R.; Silipo, A. Chemical basis of peptidoglycan discrimination by PrkC, a key kinase involved in bacterial resuscitation from dormancy. *J. Am. Chem. Soc.,* **2011**, *133*(51), 20676-20679. http://dx.doi.org/10.1021/ja208080r PMID: 22111897

[49] Page, C.A.; Krause, D.C. Protein kinase/phosphatase function correlates with gliding motility in *Mycoplasma pneumoniae*. *J. Bacteriol.,* **2013**, *195*(8), 1750-1757. http://dx.doi.org/10.1128/JB.02277-12 PMID: 23396910

[50] Xue, Y.; Li, A.; Wang, L.; Feng, H.; Yao, X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics,* **2006**, *7*, 163. http://dx.doi.org/10.1186/1471-2105-7-163 PMID: 16549034

[51] Zou, L.; Wang, M.; Shen, Y.; Liao, J.; Li, A.; Wang, M. PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC Bioinformatics,* **2013**, *14*, 247. http://dx.doi.org/10.1186/1471-2105-14-247 PMID: 23941207

[52] Xue, Y.; Ren, J.; Gao, X.; Jin, C.; Wen, L.; Yao, X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics,* **2008**, *7*(9), 1598-1608. http://dx.doi.org/10.1074/mcp.M700574-MCP200 PMID: 18463090

[53] Khatun, M.S.; Hasan, M.M.; Mollah, M.N.H.; Kurata, H. sipma: a systematic identification of protein-protein interactions in zea mays using autocorrelation features in a machine-learning framework. *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan,* **2018**, pp. 122-125.

[54] Cawley, G.C.; Talbot, N.L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.,* **2010**, *11*(Jul), 2079-2107.

[55] Leung, A.; Bader, G.D.; Reimand, J. HyperModules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. *Bioinformatics,* **2014**, *30*(15), 2230-2232. http://dx.doi.org/10.1093/bioinformatics/btu172 PMID: 24713437

[56] Xu, Y.; Wen, X.; Wen, L.-S.; Wu, L.-Y.; Deng, N.-Y.; Chou, K.-C. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One,* **2014**, *9*(8), e105018. http://dx.doi.org/10.1371/journal.pone.0105018 PMID: 25121969

[57] Su, R.; Hu, J.; Zou, Q.; Manavalan, B.; Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.,* **2019**, *21*(2), 408-420. PMID: 30649170

[58] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids,* **2019**, *16*, 733-744. http://dx.doi.org/10.1016/j.omtn.2019.04.019 PMID: 31146255

[59] Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D.C. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.,* **2019**, *20*(8), E1964. http://dx.doi.org/10.3390/ijms20081964 PMID: 31013619

[60] Hasan, M.M.; Kurata, H. Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem.,* **2018**, *30*(1), pp. 163-172.

[61] Chen, K.; Jiang, Y.; Du, L.; Kurgan, L. Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem.,* **2009**, *30*(1), 163-172. http://dx.doi.org/10.1002/jcc.21053 PMID: 18567007

[62] Wang, J.; Yang, B.; Revote, J.; Leier, A.; Marquez-Lago, T.T.; Webb, G.; Song, J.; Chou, K.C.; Lithgow, T. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics,* **2017**, *33*(17), 2756-2758. http://dx.doi.org/10.1093/bioinformatics/btx302 PMID: 28903538

[63] Hasan, M.M.; Khatun, M.S.; Kurata, H. Computational modeling of lysine post-translational modification: an overview. *Curr. Synthetic Systems Biol.,* **2018**, *6*, 137. http://dx.doi.org/10.4172/2332-0737.1000137

[64] Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol. Omics,* **2019**, *15*(6), 451-458. http://dx.doi.org/10.1039/C9MO00098D

[65] Shatabda, S.; Saha, S.; Sharma, A.; Dehzangi, A. iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features. *J. Theor. Biol.,* **2017**, *435*, 229-237. http://dx.doi.org/10.1016/j.jtbi.2017.09.022 PMID: 28943403

[66] Fu, H.; Yang, Y.; Wang, X.; Wang, H.; Xu, Y. DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins. *BMC Bioinformatics,* **2019**, *20*(1), 86. http://dx.doi.org/10.1186/s12859-019-2677-9 PMID: 30777029

[67] Hasan, M.M.; Kurata, H. gpsuc: global prediction of generic and species-specific succinylation sites by aggregating multiple sequence features. *PLoS One,* **2018**, *13*(10), e0200283. http://dx.doi.org/10.1371/journal.pone.0200283 PMID: 30312302

[68] Khatun, S.; Hasan, M.; Kurata, H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. *FEBS Lett.,* **2019**, *593*(21), 3029-3039. http://dx.doi.org/10.1002/1873-3468.13536 PMID: 31297788

[69] Mosharaf, M.P.; Hassan, M.M.; Ahmed, F.F.; Khatun, M.S.; Moni, M.A.; Mollah, M.N.H. Computational prediction of protein ubiquitination sites mapping on *Arabidopsis thaliana*. *Comput. Biol. Chem.,* **2020**, *85,* 107238. http://dx.doi.org/10.1016/j.compbiolchem.2020.107238 PMID: 32114285

[70] López, Y.; Sharma, A.; Dehzangi, A.; Lal, S.P.; Taherzadeh, G.; Sattar, A.; Tsunoda, T. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC Genomics,* **2018**, *19*(Suppl. 1), 923. http://dx.doi.org/10.1186/s12864-017-4336-8 PMID: 29363424

[71] Chowdhury, S.Y.; Shatabda, S.; Dehzangi, A. iDNAProt-ES: identification of dna-binding proteins using evolutionary and structural features. *Sci. Rep.,* **2017**, *7*(1), 14938. http://dx.doi.org/10.1038/s41598-017-14945-1 PMID: 29097781

[72] Shatabda, S.; Newton, M.A.; Rashid, M.A.; Pham, D.N.; Sattar, A. The road not taken: retreat and diverge in local search for simplified protein structure prediction. *BMC Bioinformatics,* **2013**, *14*(Suppl. 2), S19. http://dx.doi.org/10.1186/1471-2105-14-S2-S19 PMID: 23368768

[73] Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.,* **2018**, *9*, 1695. http://dx.doi.org/10.3389/fimmu.2018.01695 PMID: 30100904

[74] Vapnik, V. *The nature of statistical learning theory*; Springer: Science & Business media, **2013**.

[75] Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.,* **1999**, *10*(5), 988-999. http://dx.doi.org/10.1109/72.788640 PMID: 18252602

[76] Chen, Z.; Chen, Y.-Z.; Wang, X.-F.; Wang, C.; Yan, R.-X.; Zhang, Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One,* **2011**, *6*(7), e22930. http://dx.doi.org/10.1371/journal.pone.0022930 PMID: 21829559

[77] Chen, Z.; Zhou, Y.; Song, J.; Zhang, Z. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim. Biophys. Acta,* **2013**, *1834*(8), 1461-1467. http://dx.doi.org/10.1016/j.bbapap.2013.04.006 PMID: 23603789

[78] Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **2006**, *22*(13), 1658-1659. http://dx.doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

[79] Breiman, L. Random forests. *Mach. Learn.,* **2001**, *45*(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324

[80] Qiang, X; Zhou, C; Ye, X; Du, P.-f; Su, R; Wei, L. A predictor for CPP identification. *Brief. Bioinform.,* **2018**.

[81] Manavalan, B.; Lee, J.; Lee, J. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS One,* **2014**, *9*(9), e106542. http://dx.doi.org/10.1371/journal.pone.0106542 PMID: 25222008

[82] Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.,* **2018**, *9*, 276. http://dx.doi.org/10.3389/fphar.2018.00276 PMID: 29636690

[83] Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.,* **2018**, *17*(8), 2715-2726. http://dx.doi.org/10.1021/acs.jproteome.8b00148 PMID: 29893128

[84] Hasan, M.M.; Khatun, M.S.; Mollah, M.N.H.; Yong, C.; Dianjing, G. NTyroSite: computational identification of protein nitrotyrosine sites using sequence evolutionary features. *Molecules,* **2018**, *23*(7), 1667. http://dx.doi.org/10.3390/molecules23071667 PMID: 29987232

[85] Khatun, M.S.; Hasan, M.M.; Kurata, H. PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front. Genet.,* **2019**, *10*, 129. http://dx.doi.org/10.3389/fgene.2019.00129 PMID: 30891059

[86] Hasan, M.M.; Guo, D.; Kurata, H. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. *Mol. Biosyst.,* **2017**, *13*(12), 2545-2550. http://dx.doi.org/10.1039/C7MB00491E PMID: 28990628

[87] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.,* **2016**, *497*, 48-56. http://dx.doi.org/10.1016/j.ab.2015.12.009 PMID: 26723495

[88] Shoombuatong, W.; Schaduangrat, N.; Pratiwi, R.; Nantasenamat, C. THPep: A machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.,* **2019**, *80*, 441-451. http://dx.doi.org/10.1016/j.compbiolchem.2019.05.008 PMID: 31151025

[89] Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.,* **2016**, *394*, 223-230. http://dx.doi.org/10.1016/j.jtbi.2016.01.020 PMID: 26807806

[90] Hasan, M.M.; Khatun, M.S.; Mollah, M.N.H.; Yong, C.; Guo, D. A systematic identification of species-specific protein succinylation sites using joint element features information. *Int. J. Nanomedicine,* **2017**, *12*, 6303-6315. http://dx.doi.org/10.2147/IJN.S140875 PMID: 28894368

[91] Tang, Y.-R.; Chen, Y.-Z.; Canchaya, C.A.; Zhang, Z. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng. Des. Sel.,* **2007**, *20*(8), 405-412. http://dx.doi.org/10.1093/protein/gzm035 PMID: 17652129

[92] Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics,* **2004**, *4*(6), 1633-1649. http://dx.doi.org/10.1002/pmic.200300771 PMID: 15174133

[93] Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics,* **2009**, *25*(19), 2537-2543. http://dx.doi.org/10.1093/bioinformatics/btp445 PMID: 19654118

[94] McGuffin, L.J.; Bryson, K.; Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics,* **2000**, *16*(4), 404-405. http://dx.doi.org/10.1093/bioinformatics/16.4.404 PMID: 10869041

[95] Johansen, M.B.; Kiemer, L.; Brunak, S. Analysis and prediction of mammalian protein glycation. *Glycobiology,* **2006**, *16*(9), 844-853. http://dx.doi.org/10.1093/glycob/cwl009 PMID: 16762979

[96] Zhang, J.; Zhao, X.; Sun, P.; Ma, Z. PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. *Int. J. Mol. Sci.,* **2014**, *15*(7), 11204-11219. http://dx.doi.org/10.3390/ijms150711204 PMID: 24968264

[97] Blom, N.; Gammeltoft, S.; Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.,* **1999**, *294*(5), 1351-1362. http://dx.doi.org/10.1006/jmbi.1999.3310 PMID: 10600390

[98] Kavuncuoglu, H.; Kavuncuoglu, E.; Karatas, S.M.; Benli, B.; Sagdic, O.; Yalcin, H. Prediction of the antimicrobial activity of walnut (*Juglans regia* L.) kernel aqueous extracts using artificial neural network and multiple linear regression. *J. Microbiol. Methods,* **2018**, *148*, 78-86. http://dx.doi.org/10.1016/j.mimet.2018.04.003 PMID: 29649523

[99] Wu, K.; Wei, G.-W. Quantitative toxicity prediction using topology based multitask deep neural networks. *J. Chem. Inf. Model.,* **2018**, *58*(2), 520-531. http://dx.doi.org/10.1021/acs.jcim.7b00558 PMID: 29314829

[100] Peters, B.; Brenner, S.E.; Wang, E.; Slonim, D.; Kann, M.G. *Putting benchmarks in their rightful place: The heart of computational biology*; Public Library of Science, **2018**.

[101] Berezikov, E.; Guryev, V.; Plasterk, R.H.; Cuppen, E. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.,* **2004**, *14*(1), 170-178. http://dx.doi.org/10.1101/gr.1642804 PMID: 14672977

[102] Biswas, A.K.; Noman, N.; Sikder, A.R. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics,* **2010**, *11*, 273. http://dx.doi.org/10.1186/1471-2105-11-273 PMID: 20492656

[103] Macek, B.; Gnad, F.; Soufi, B.; Kumar, C.; Olsen, J.V.; Mijakovic, I.; Mann, M. Phosphoproteome analysis of *E. coli* reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol. Cell. Proteomics,* **2008**, *7*(2), 299-307. http://dx.doi.org/10.1074/mcp.M700311-MCP200 PMID: 17938405

[104] Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.,* **2018**, *9*, 476. http://dx.doi.org/10.3389/fmicb.2018.00476 PMID: 29616000

[105] Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.,* **2018**, *16*, 412-420. http://dx.doi.org/10.1016/j.csbj.2018.10.007 PMID: 30425802

[106] Charoenkwan, P.; Nantasenamat, C.; Hasan, M.M.; Shoombuatong, W. iTTCA-Hybrid: improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. *Anal. Biochem.,* **2020**, *599,* 113747. http://dx.doi.org/10.1016/j.ab.2020.113747 PMID: 32333902

[107] Gnad, F.; Ren, S.; Cox, J.; Olsen, J.V.; Macek, B.; Oroshi, M.; Mann, M. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.,* **2007**, *8*(11), R250. http://dx.doi.org/10.1186/gb-2007-8-11-r250 PMID: 18039369

[108] Hasan, M.M.; Yang, S.; Zhou, Y.; Mollah, M.N. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Mol. Biosyst.,* **2016**, *12*(3), 786-795. http://dx.doi.org/10.1039/C5MB00853K PMID: 26739209

[109] Ward, P.; Equinet, L.; Packer, J.; Doerig, C. Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics,* **2004**, *5*(1), 79. http://dx.doi.org/10.1186/1471-2164-5-79 PMID: 15479470

[110] Charoenkwan, P.; Yana, J.; Schaduangrat, N.; Nantasenamat, C.; Hasan, M.M.; Shoombuatong, W. iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics,* **2020**, *112*(4), 2813-2822. http://dx.doi.org/10.1016/j.ygeno.2020.03.019 PMID: 32234434

[111] Hasan, M.M.; Manavalan, B.; Shoombuatong, W.; Khatun, M.S.; Kurata, H. i4mC-Mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. *Comput. Struct. Biotechnol. J.,* **2020**, *18*, 906-912. http://dx.doi.org/10.1016/j.csbj.2020.04.001 PMID: 32322372

[112] Chen, W.; Song, X.; Lv, H.; Lin, H. iRNA-m2G: identifying N^2-methylguanosine sites based on sequence-derived information. *Mol. Ther. Nucleic Acids,* **2019**, *18*, 253-258. http://dx.doi.org/10.1016/j.omtn.2019.08.023 PMID: 31581049

[113] Lai, H.-Y.; Zhang, Z.-Y.; Su, Z.-D.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids,* **2019**, *17*, 337-346. http://dx.doi.org/10.1016/j.omtn.2019.05.028 PMID: 31299595

[114] Lv, H.; Zhang, Z.-M.; Li, S.-H.; Tan, J.-X.; Chen, W.; Lin, H. Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform.,* **2019**, *21*(3), 982-995. PMID: 31157855

[115] Govindaraj, R.G.; Subramaniyam, S.; Manavalan, B. Extremely-randomized-tree-based prediction of N6-methyladenosine sites in *Saccharomyces cerevisiae*. *Curr. Genomics,* **2020**, *21*(1), 26-33.

[116] Chen, X.; Huang, L.; Xie, D.; Zhao, Q. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.,* **2018**, *9*(1), 3. http://dx.doi.org/10.1038/s41419-017-0003-x PMID: 29305594

[117] Li, F.; Chen, J.; Leier, A.; Marquez-Lago, T.; Liu, Q.; Wang, Y.; Revote, J.; Smith, A.I.; Akutsu, T.; Webb, G.I. DeepCleave: a deep

learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics,* **2019**, *36*(4), 1057-1065. http://dx.doi.org/10.1093/bioinformatics/btz721 PMID: 31566664

[118] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J.,* **2019**, *17*, 972-981. http://dx.doi.org/10.1016/j.csbj.2019.06.024 PMID: 31372196

[119] Hasan, M.M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics*, **2020**, *36*(11), 3350-3356. http://dx.doi.org/10.1093/bioinformatics/btaa160

[120] Hasan, M.M.; Manavalan, B.; Shoombuatong, W.; Khatun, M.S.; Kurata, H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol. Biol.,* **2020**, *103*(1-2), 225-234. http://dx.doi.org/10.1007/s11103-020-00988-y PMID: 32140819

[121] Hasan, MM; Manavalan, B; Khatun, MS; Kurata, H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. *Int. J. Biol. Macromol.,* **2019**, S0141-8130(19)38547-2.

[122] Hasan, M.M.; Khatun, M.S.; Kurata, H. Large-scale assessment of bioinformatics tools for lysine succinylation sites. *Cells,* **2019**, *8*(2), E95. http://dx.doi.org/10.3390/cells8020095 PMID: 30696115

[123] Radovic, M.; Ghalwash, M.; Filipovic, N.; Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics,* **2017**, *18*(1), 9. http://dx.doi.org/10.1186/s12859-016-1423-9 PMID: 28049413

[124] Gayatri, N.; Nickolas, S.; Reddy, A. anova discriminant analysis for features selected through decision tree induction method. *International Conference on Computing and Communication Systems,* **2011**, pp. 61-70.

[125] Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.,* **2016**, *10*(Suppl. 4), 114. http://dx.doi.org/10.1186/s12918-016-0353-5 PMID: 28155714

[126] Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing,* **2016**, *173*, 346-354. http://dx.doi.org/10.1016/j.neucom.2014.12.123

[127] Cheng, X.; Lin, W.Z.; Xiao, X.; Chou, K.C. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics,* **2019**, *35*(3), 398-406. http://dx.doi.org/10.1093/bioinformatics/bty628 PMID: 30010789

[128] Chou, K.C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.,* **2004**, *11*(16), 2105-2134. http://dx.doi.org/10.2174/0929867043364667 PMID: 15279552

[129] Chou, K.C.; Cai, Y.D. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.,* **2003**, *90*(6), 1250-1260. http://dx.doi.org/10.1002/jcb.10719 PMID: 14635197

[130] Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids,* **2016**, *5,* e332. PMID: 28427142

[131] Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.,* **2015**, *43*(W1), W65-W71. http://dx.doi.org/10.1093/nar/gkv458 PMID: 25958395

[132] Basith Mail, S.; Manavalan, B.; Shin, T.H.; Lee, D.; Lee, G. Evolution of machine learning algorithms in the prediction and design of anticancer peptides. *Curr. Protein Pept. Sci.,* **2020**. http://dx.doi.org/10.2174/1389203721666200117171403 PMID: 31957610

[133] Charoenkwan, P.; Kanthawong, S.; Schaduangrat, N.; Yana, J.; Shoombuatong, W. PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method. *Cells,* **2020**, *9*(2), E353. http://dx.doi.org/10.3390/cells9020353 PMID: 32028709

[134] Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int. J. Mol. Sci.,* **2019**, *20*(22), E5743. http://dx.doi.org/10.3390/ijms20225743 PMID: 31731751

[135] Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J.,* **2018**, *17*, 734-752. PMID: 30190664