International Journal of Biological Macromolecules 167 (2021) 1575–1578

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: http://www.elsevier.com/locate/ijbiomac

Classifying the superfamily of small heat shock proteins by using g-gap dipeptide compositions

Pengmian Feng *, Weiwei Liu, Cong Huang, Zhaohui Tang

School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China

ARTICLE INFO

Article history:
Received 7 August 2020
Received in revised form 2 November 2020
Accepted 13 November 2020
Available online 17 November 2020

Keywords:
Small heat shock proteins
Superfamily
Support vector machine
g-Gap dipeptide composition

ABSTRACT

Small heat shock protein (sHSP) is a superfamily of molecular chaperone and is found from archaea to human. Recent researches have demonstrated that sHSPs participate in a series of biological processes and are even closely associated with serious diseases. Since sHSP is a very large superfamily and members from different superfamilies exhibit distinct functions, accurate classification of the subfamily of sHSP will be helpful for unrevealing its functions. In the present work, a support vector machine-based method was proposed to classify the subfamily of sHSPs. In the 10-fold cross validation test, an overall accuracy of 93.25% was obtained for classifying the subfamily of sHSPs. The superiority of the proposed method was also demonstrated by comparing it with the other methods. It is anticipated that the proposed method will become a useful tool for classifying the subfamily of sHSPs.

1. Introduction

Small heat shock protein (sHSP) is a kind of ATP-independent molecular chaperones [1]. sHSPs are found to be expressed throughout the three kingdoms of life from archaea to human [2,3] and even in the viruses [4]. sHSP was regarded as the defense system of conditions that endanger the cellular proteome [5]. By preventing the aggregation of proteins, sHSPs play important roles in biological and cellular processes, such as proteasomal degradation, cell signaling, cell differentiation and cell apoptosis [6].

Recent researches have demonstrated that both abnormal expression of sHSPs and mutations in sHSPs are associated the pathological conditions in human [7]. For example, the up-regulation of sHSPs can promote the cancer development, while its down-regulation leads to beneficial outcomes [6]. sHSPs are also linked to the development of diseases, such as neurological disease [8], metabolic diseases [9] and cancers [10]. For example, several members of the sHSPs were reported to be associated with human neurodegenerative disorders, such as Alzheimer's disease, Charcot-Marie-Tooth disease (CMT) and Parkinson's disease [11]. More details about the links between sHSPs and human diseases were discussed in a recent review [12]. On the contrary, the benefits of sHSPs were also observed. sHSPs can not only promote longevity and healthy aging in vivo [13], but also become potential drug targets for new therapeutic for aging diseases and cancers [10]. For example, the ubiquitous expression of sHSPs is closely associated with the progression of numerous of cancers. Therefore, the sHSPs have been regarded as new targets of cancer therapy [14]. In addition, sHSPs can stimulate macrophages to suppress inflammation and hold the potential to be therapeutic agents for inflammatory disorders [15]. However, our understanding about its molecular mechanism is still at the infant stage.

In fact, sHSP is a very large superfamily of molecular chaperones. Based on their amino acid compositions and domains, the sHSPs can be classified into 21 superfamilies at least [16]. Since their amino acid composition and domains are distinct, the members from different superfamilies of sHSPs exhibit different functions. Hence, it is necessary to develop automated methods for demining which superfamily of a query sHSP belonging to.

Keeping this in mind, in the present work, a support vector machine-based method was proposed to classify the subfamilies of sHSPs, in which the protein sequences were encoded by using the g-gap dipeptide. Comparative results among the methods based on different kinds of commonly used features demonstrate that the superiority of the proposed method for classifying the subfamilies of sHSPs. It is anticipated that the proposed method will become a useful tool for researches on sHSPs.

2. Materials and methods

2.1. Datasets

The sHSPs were obtained from the small Heat Shock Proteins database (sHSPdb, http://forge.info.univangers.fr/~gh/Shspdb/index.php). By gathering data from Uniprot, PFAM, InterPro, etc., the sHSPdb has deposited more than 6200 sHSPs from nearly all kingdoms of life [16].

* Corresponding author.
E-mail address: fengpengmian@gmail.com (P. Feng).

https://doi.org/10.1016/j.ijbiomac.2020.11.111
0141-8130/

Based on the unique sequence motif, the sHSPs are classified into 21 superfamilies. To build the dataset for sHSP superfamily classification, we harvested all the sequences deposited in sHSPdb.

In order to obtain a high quality dataset, the CD-HIT tool was used to remove sequences with the sequence similarity greater than 40% [17,18]. Accordingly, we obtained 1683 sequences belong to the 21 superfamilies of sHSP. The detail number of sequences in each superfamily was shown in Fig. 1. It was found that the number of samples in most superfamilies was too small to have statistical significance. Thus, only sHSP10, sHSP11 and sHSP12 were left in the final benchmark dataset that includes 118, 403 and 86 samples for these superfamilies, respectively. The benchmark dataset is provided in Supplementary material.

2.2. g-Gap dipeptide composition

The most straightforward method to encode protein sequences is by using the amino acid composition [19–21]. However, the global sequence information couldn't be reflected by amino acid composition [22]. In order to integrate long-range sequence order information, the *g*-gap dipeptide composition was proposed to represent protein sequences. Compared with the amino acid composition, the *g*-gap dipeptide composition could describe both local and global correlations between amino acids in a sequence [23]. The definition of *g*-gap dipeptide composition is as following,

$$\mathrm{F} = \left[f_1^{\,g}\ f_2^{\,g} \ldots f_i^{\,g} \ldots f_{400}^{\,g} \right] \tag{1}$$

where $f_i^{\,g}$ is the frequency of the *i*-th ($i = 1, 2, \ldots, 400$) dipeptide with *g*-gap interval in the sequence. *g* is an integer and was set in the range of [0, 5] in the present work. $g = 0$ represents the correlation between two adjacent amino acids, and $g = 1$ represents the correlation of two amino acids with one amino acid interval, and so forth.

2.3. Support vector machine

Support vector machine (SVM) is a supervised learning method and has been successfully used in the realm of bioinformatics [24–30]. In the present work, the LIBSVM package 3.20 downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/ was used to perform the prediction. The radial basis kernel function (RBF) was used to obtain the classification hyperplane. In this work, the best regularization parameter *C* and kernel width parameter *g* were determined by using the grid search method in the ranges $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{-5}]$ with the steps of 2 and 2^{-1}, respectively.

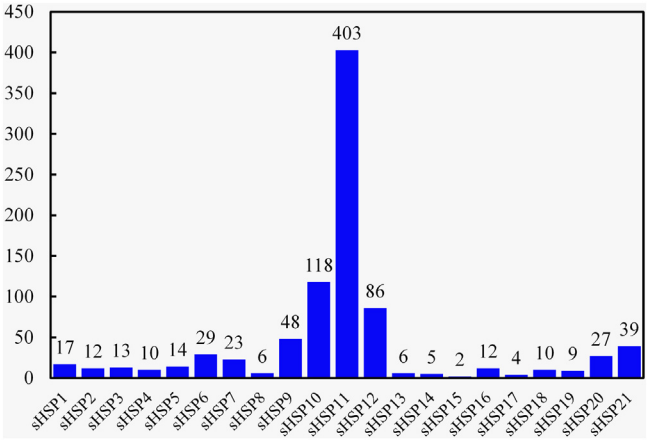

Fig. 1. The number of samples in each superfamily of sHSP.

2.4. Performance evaluation

The performance of the proposed method was evaluated by the commonly used metrics [31–37], namely sensitivity (Sn), specificity (Sp) and overall accuracy (OA), which are defined as following,

$$\begin{cases} Sn(\mathrm{i}) = \dfrac{TP(\mathrm{i})}{TP(\mathrm{i}) + FN(\mathrm{i})} \times 100\% \\ Sp(\mathrm{i}) = \dfrac{TN(\mathrm{i})}{TN(\mathrm{i}) + FP(\mathrm{i})} \times 100\% \\ OA = \dfrac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{3} TP(\mathrm{i}) \end{cases} \tag{2}$$

where TP(i) is the number of the correctly identified positive samples in the *i*-th family, FN(i) is the number of the samples in the *i*-th family that are incorrectly predicted to be of other families, TN(i) is the number of the correctly identified samples that are not belong to the *i*-th family, FP(i) is the number of the samples in the other family that are incorrectly predicted to be of the *i*-th family. *N* is the total number of samples in the dataset.

3. Results

3.1. Amino acids composition analysis

In order to demonstrate the rationality of using the sequence based information to describe the proteins, the Composition Profiler [38] was employed to analyze the relative amino acid preference in the three superfamilies of sHSP. To find out the amino acid bias in a specific superfamily, the sequences from it was set as the query samples, the sequences from the remaining superfamilies were used as the background samples.

By setting *p*-value≤0.01 and *Bootstrap* = 1000 in Composition Profiler, we analyzed the amino acid preference in the three superfamilies of sHSP. The results were shown in Fig. 2. It was found that His (H) and Ser (S) were enriched in sHSP10 and sHSP12, Trp (W) was enriched in sHSP11 and sHSP 12. Moreover, the superfamily specific amino acid preference was also observed. Ala(A), Phe(F), Asn(N), Thr (T) and Tyr(Y) were significantly enriched in sHSP10; Glu(E), Ile(I), Lys(K), Met(M) and Val(V) were significantly enriched in sHSP11; Cys (C), Pro(P) and Gln(Q) were significantly enriched in sHSP12. These results demonstrate that the amino acid preference are different among the three superfamilies of sHSP. Therefore, it is reasonable to classify sHSP superfamilies by using sequence based information.

3.2. Performance of classifying the subfamilies of sHSP

Based on the above analysis, we encode the sHSPs by using the *g*-gap dipeptide composition method. As indicated in Eq. (1), the greater the *g* is, the longer range sequence-order information will be included. However, the robustness of the signal-to-noise ratio might also be affected at the meantime. Therefore, it's necessary to determine the optimal *g*.

In the present work, our searching for the optimal value of *g* was carried out in the range of [0, 5] with a step of 1. Accordingly, six models ($g = 0, 1, \ldots, 5$) were built. The 10-fold cross validation test results for identifying the superfamilies of sHSPs in the datasets were listed in Table 1. It was found that the best predictive accuracy (93.25%) was obtained when $g = 0$ and $g = 1$. This result indicates that the accuracy wasn't improved with the increment of *g*. It should also be mentioned that, although the model based on $g = 0$ and $g = 1$ yielded the same accuracy, the sensitivity for identifying sHSP12 based on $g = 1$ is significantly higher than that based on $g = 0$. Therefore, the optimal value of *g* was used to build the model for classifying the superfamilies of sHSP.

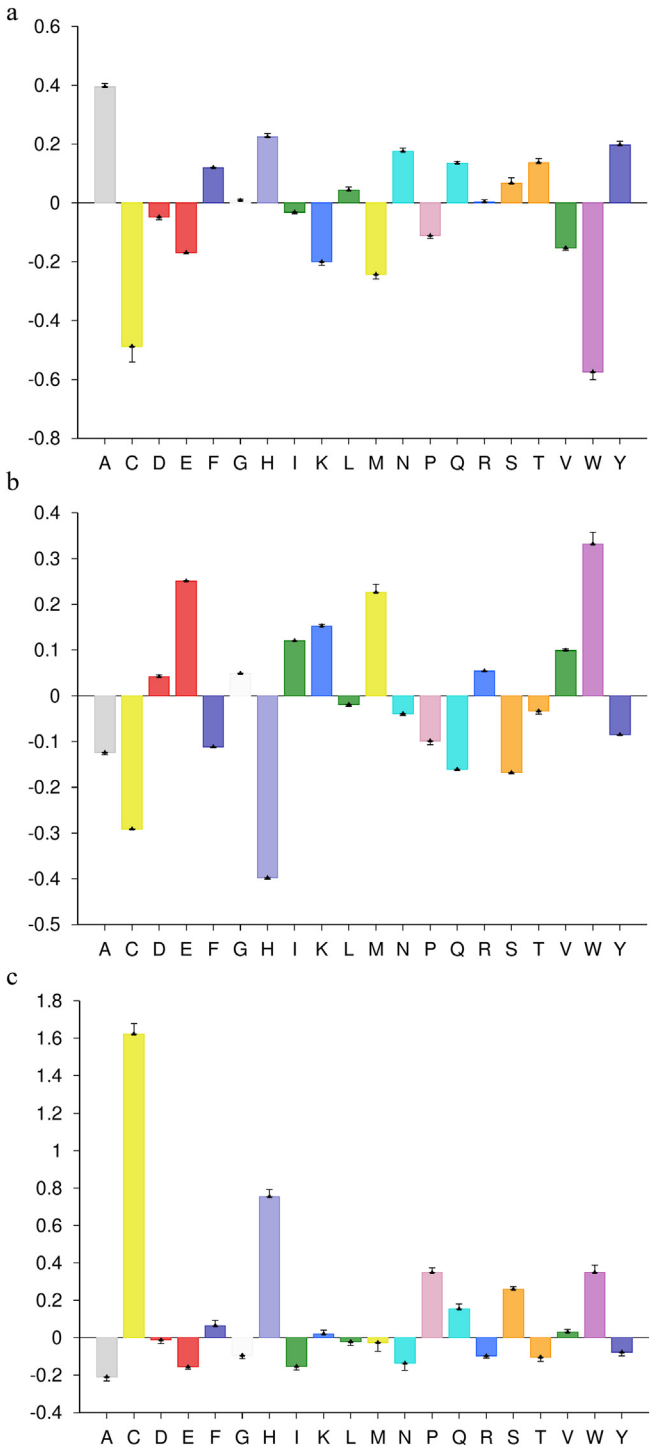

Fig. 2. The amino acid composition bias in the three superfamilies of sHSP. (a) the relative enrichment and depletion of amino acid in sHSP10; (b) The relative enrichment and depletion of amino acid in sHSP11; (c) The relative enrichment and depletion of amino acid in sHSP12. The x-axis is the 20 native amino acid, and the y-axis is the relative enrichment ratio.

3.3. Comparison with other methods

To demonstrate the performance of the proposed method, it's necessary to compare the proposed method with the other methods. Since there is no methods available for this aim, we compared the performance of the proposed method with that based on commonly used features, namely composition transition distribution and reduced amino acid composition (RAAC).

Table 1
Performance for classifying subfamilies of sHSPs by using g-gap dipeptide composition.

Sub family	Metrics	Features					
		g = 0	g = 1	g = 2	g = 3	g = 4	g = 5
sHSP_10	Sn (%)	94.07	**92.37**	92.37	52.54	85.59	84.75
	Sp (%)	98.48	**98.28**	97.20	99.76	98.29	97.41
sHSP_11	Sn (%)	97.52	**97.02**	96.03	98.77	97.52	95.78
	Sp (%)	84.80	**85.78**	85.22	38.24	82.35	82.09
sHSP_12	Sn (%)	72.09	**76.74**	74.42	18.06	77.91	75.58
	Sp (%)	97.52	**96.72**	96.72	96.72	98.33	93.55
	OA(%)	93.25	**93.25**	92.26	79.08	92.42	90.77

Table 2
Results for classifying subfamilies of sHSPs by using composition transition distribution.

Sub family	Metrics	Features		
		CTDC	CTDT	CTDD
sHSP_10	Sn (%)	72.88	88.13	73.72
	Sp (%)	93.43	95.52	96.92
sHSP_11	Sn (%)	89.82	91.56	95.03
	Sp (%)	68.84	80.09	72.50
sHSP_12	Sn (%)	59.30	66.27	67.44
	Sp (%)	87.12	87.41	92.06
	OA (%)	82.21	87.31	86.99

The composition transition distribution is widely used in computational proteomics, which describes the sequences in terms of the structural or physicochemical property of amino acid [39]. By using iFeature [40], we calculated the composition (CTDC), transition (CTDT) and distribution (CTDD), based on the normalized Van der Waals Volume, polarity, polarizability, charge, secondary structures, solvent accessibility and 7 types of hydrophobicity of the twenty amino acids. The predictive results based on composition transition distribution were reported in Table 2. The best accuracy of 87.31% was obtained based on CTDD, which is lower than that based on the 1-gap dipeptide composition.

Compared with the amino acid composition, RAAC can extract the structural similarity information of the sequence and was also widely used for protein family classifications [41]. According to different optimization procedures [42], the twenty amino acids can be clustered into five different clusters with the number of reduced amino acid alphabet of 5, 8, 9, 11 and 13, respectively. For clarity, we called these five profiles as CP(5), CP(8), CP(9), CP(11), CP(13), respectively. The predictive results of the dipeptide composition based on RAAC were listed in Table 3. The modeled based on CP(13) yielded the best accuracy of 92.42%, which is also lower than that based on the 1-gap dipeptide composition.

4. Conclusions

Based on the experimental data deposited in sHSPdb, a high quality dataset was built for classifying the subfamilies of sHSPs. Consistent with reported results, the amino acid composition preference was also

Table 3
Results for classifying subfamilies of sHSPs by using the dipeptide composition based on RAAC.

Sub family	Metrics	Features				
		CP(5)	CP(8)	CP(9)	CP(11)	CP(13)
sHSP_10	Sn (%)	75.42	87.29	88.96	90.68	93.22
	Sp (%)	94.24	97.82	96.49	97.39	97.40
sHSP_11	Sn (%)	92.06	96.77	95.04	96.77	96.53
	Sp (%)	72.96	79.70	81.47	82.27	84.24
sHSP_12	Sn (%)	62.79	67.44	66.28	69.77	70.93
	Sp (%)	89.31	95.87	92.80	95.94	97.52
	OA (%)	84.68	90.77	89.79	91.27	92.42

observed in the different superfamilies of sHSPs. Accordingly, a support vector machine based method was proposed to classify the three families of sHSP (i.e. sHSP10, sHSP11 and sHSP12), in which the 1-gap dipeptide composition was used to encode the sequences. Comparative results demonstrated that the proposed method is promising for the subfamilies of sHSPs.

It has not escaped our notice that, due to the limited number of samples in the other superfamilies, the current method is limited to classify sHSP10, sHSP11 and sHSP12. Therefore, in future works, we will enlarge the samples of sHSPs by collecting the data from literatures to develop a new model able to classify more superfamilies of sHSPs by using ensemble classifiers [43] and deep learning methods [44,45].

Funding

This work was supported by the Xinglin Scholar Research Premotion Project of Chengdu University of TCM (no. ZRQN2020003).

Declaration of competing interest

The authors declare that there is no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2020.11.111.

References

[1] M. Haslbeck, S. Weinkauf, J. Buchner, Small heat shock proteins: simplicity meets complexity, J. Biol. Chem. 294 (6) (2019) 2121–2132.
[2] S. Carra, S. Alberti, J.L.P. Benesch, W. Boelens, J. Buchner, J.A. Carver, C. Cecconi, H. Ecroyd, N. Gusev, L.E. Hightower, R.E. Klevit, H.O. Lee, K. Liberek, B. Lockwood, A. Poletti, V. Timmerman, M.E. Toth, E. Vierling, T. Wu, R.M. Tanguay, Small heat shock proteins: multifaceted proteins with important implications for life, Cell Stress Chaperones 24 (2) (2019) 295–308.
[3] S.J. Eyles, L.M. Gierasch, Nature's molecular sponges: small heat shock proteins grow into their chaperone roles, Proc. Natl. Acad. Sci. U. S. A. 107 (7) (2010) 2727–2728.
[4] H. Maaroufi, R.M. Tanguay, Analysis and phylogeny of small heat shock proteins from marine viruses and their cyanobacteria host, PLoS One 8 (11) (2013), e81207.
[5] M. Haslbeck, E. Vierling, A first line of stress defense: small heat shock proteins and their function in protein homeostasis, J. Mol. Biol. 427 (7) (2015) 1537–1548.
[6] R. Bakthisaran, R. Tangirala, M. Rao Ch, Small heat shock proteins: role in cellular functions and pathology, Biochim. Biophys. Acta 1854 (4) (2015) 291–319.
[7] A.J. Macario, E. Conway de Macario, Sick chaperones, cellular stress, and disease, N. Engl. J. Med. 353 (14) (2005) 1489–1501.
[8] E. Gaggelli, H. Kozlowski, D. Valensin, G. Valensin, Copper homeostasis and neurodegenerative disorders (Alzheimer's, prion, and Parkinson's diseases and amyotrophic lateral sclerosis), Chem. Rev. 106 (6) (2006) 1995–2044.
[9] Y. Sun, T.H. MacRae, The small heat shock proteins and their role in human disease, FEBS J. 272 (11) (2005) 2613–2627.
[10] A. Zoubeidi, M. Gleave, Small heat shock proteins in cancer therapy and prognosis, Int. J. Biochem. Cell Biol. 44 (10) (2012) 1646–1656.
[11] M.M. Wilhelmus, W.C. Boelens, I. Otte-Holler, B. Kamps, B. Kusters, M.L. Maat-Schieman, R.M. de Waal, M.M. Verbeek, Small heat shock protein HspB8: its distribution in Alzheimer's disease brains and its inhibition of amyloid-beta protein aggregation and cerebrovascular amyloid-beta toxicity, Acta Neuropathol. 111 (2) (2006) 139–149.
[12] H.H. Kampinga, C. Garrido, HSPBs: small proteins with big implications in human disease, Int. J. Biochem. Cell Biol. 44 (10) (2012) 1706–1710.
[13] M.J. Vos, S. Carra, B. Kanon, F. Bosveld, K. Klauke, O.C. Sibon, H.H. Kampinga, Specific protein homeostatic functions of small heat-shock proteins increase lifespan, Aging Cell 15 (2) (2016) 217–226.
[14] J. Xiong, Y. Li, X. Tan, L. Fu, Small heat shock proteins in cancers: functions and therapeutic potential for cancer therapy, Int. J. Mol. Sci. 21 (18) (2020).
[15] J.M. van Noort, M. Bsibsi, P. Nacken, W.H. Gerritsen, S. Amor, The link between small heat shock proteins and the immune system, Int. J. Biochem. Cell Biol. 44 (10) (2012) 1670–1679.
[16] E. Jaspard, G. Hunault, sHSPdb: a database for the analysis of small heat shock proteins, BMC Plant Biol. 16 (1) (2016) 135.
[17] Q. Zou, G. Lin, X. Jiang, X. Liu, X. Zeng, Sequence clustering in bioinformatics: an empirical study, Brief. Bioinform. 21 (1) (2020) 1–10.
[18] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152.
[19] Z. Lv, S. Jin, H. Ding, Q. Zou, A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features, Front. Bioeng. Biotechnol. 7 (2019) 215.
[20] J. Zhang, B. Liu, A review on the recent developments of sequence-based protein feature extraction methods, Curr. Bioinforma. 14 (3) (2019) 190–199.
[21] W. Chen, P. Feng, T. Liu, D. Jin, Recent advances in machine learning methods for predicting heat shock proteins, Curr. Drug Metab. 20 (3) (2019) 224–228.
[22] W. Chen, P. Feng, F. Nie, iATP: a sequence based method for identifying anti-tubercular peptides, Med. Chem. 16 (5) (2019) 620–625.
[23] P.M. Feng, H. Ding, W. Chen, H. Lin, Naive Bayes classifier with feature selection to identify phage virion proteins, Comput. Math. Methods Med. 2013 (2013), 530696.
[24] J.X. Tan, S.H. Li, Z.M. Zhang, C.X. Chen, W. Chen, H. Tang, H. Lin, Identification of hormone binding proteins based on machine learning methods, Math. Biosci. Eng. 16 (4) (2019) 2466–2480.
[25] W. Yang, X.J. Zhu, J. Huang, H. Ding, H. Lin, A brief survey of machine learning methods in protein sub-Golgi localization, Curr. Bioinforma. 14 (2019) 234–240.
[26] H. Yang, H. Lv, H. Ding, W. Chen, H. Lin, iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens, J. Comput. Biol. 25 (11) (2018) 1266–1277.
[27] C. Meng, F. Guo, Q. Zou, CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes, Comput. Biol. Chem. 87 (2020), 107304, .
[28] L. Chao, L. Wei, Q. Zou, SecProMTB: a SVM-based classifier for secretory proteins of Mycobacterium tuberculosis with imbalanced data set, Proteomics 19 (2019), e1900007.
[29] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (22) (2017) 3518–3523.
[30] B. Liu, C. Li, K. Yan, DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks, Brief. Bioinform. 21 (5) (2020) 1733–1741.
[31] H. Yang, W. Yang, F.Y. Dao, H. Lv, H. Ding, W. Chen, H. Lin, A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae, Brief. Bioinform. 21 (5) (2020) 1568–1580.
[32] K. Liu, W. Chen, iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications, Bioinformatics 36 (11) (2020) 3336–3342.
[33] W. Chen, P. Feng, X. Song, H. Lv, H. Lin, iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features, Mol. Ther. Nucleic Acids 18 (2019) 269–274.
[34] B. Liu, X. Gao, H. Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches, Nucleic Acids Res. 47 (20) (2019) e127.
[35] R. Su, X. Liu, L. Wei, MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy, Brief. Bioinform. 21 (2) (2020) 687–698.
[36] R. Su, X. Liu, G. Xiao, L. Wei, Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction, Brief. Bioinform. 21 (3) (2019) 996–1005.
[37] R. Su, H. Wu, X. Liu, L. Wei, Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies, Brief. Bioinform. (2020) https://doi.org/10.1093/bib/bbz165.
[38] V. Vacic, V.N. Uversky, A.K. Dunker, S. Lonardi, Composition profiler: a tool for discovery and visualization of amino acid composition differences, BMC Bioinform. 8 (2007) 211.
[39] P. Feng, L. Feng, Sequence based prediction of pattern recognition receptors by using feature selection technique, Int. J. Biol. Macromol. 162 (2020) 931–934.
[40] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C. Chou, J. Song, iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.
[41] L. Zheng, D. Liu, W. Yang, L. Yang, Y. Zuo, RaacLogo: a new sequence logo generator by using reduced amino acid clusters, Brief. Bioinform. (2020)https://doi.org/10.1093/bib/bbaa096.
[42] C. Etchebest, C. Benros, A. Bornot, A.C. Camproux, A.G. de Brevern, A reduced amino acid alphabet for understanding and designing protein adaptation to mutation, Eur. Biophys. J. 36 (8) (2007) 1059–1069.
[43] J. Li, L. Wei, F. Guo, Q. Zou, EP3: an ensemble predictor that accurately identifies type III secreted effectors, Brief. Bioinform. (2020)https://doi.org/10.1093/bib/bbaa008.
[44] Z. Lv, C. Ao, Q. Zou, Protein function prediction: from traditional classifier to deep learning, Proteomics 19 (14) (2019), e1900119.
[45] K.W. Liu, L. Cao, P.F. Du, W. Chen, im6A-TS-CNN: identifying N6-methyladenine site in multiple tissues by using convolutional neural network, Mol. Ther. Nucleic Acid 21 (2020) 1044–1049.