Building Random Forest QSAR Models for Affinity Identification of 14-3-3 ζ with Optimized Parameters

Ying Fan
Shenzhen Institute of Information Technology, Shenzhen 518172, China

Xiaojun Wang
Shenzhen Institute of Information Technology, Shenzhen 518172, China

Chao Wang *
HKBU Institute for Research and Continuing Education, Shenzhen 518057, China; School of Chinese Medicine, Hong Kong Baptist University, Hong Kong 999077, China

ABSTRACT

14-3-3s present in multiple isoforms in human cells and mediate signal transduction by binding to phosphoserine-containing proteins. Previous studies demonstrate that the isoform 14-3-3 ζ acts as a key factor in promoting chemoresistance of cancer. Here, our work is devoted to developing the predictive models that can determine the binding affinity of phosphopeptide fragments against 14-3-3 ζ by the random forest approach. Based on the variable matrix built by the simple descriptors DPPS and statistical methods coupled with optimized hyperparameters, the robust models are obtained by a combinatorial peptide microarray dataset (n = 385 for N-terminal sublibrary, n = 384 for C-terminal sublibrary). For the test set, the R^2 and RMSE are 0.8532 and 0.4516 at the N-terminal sublibrary (n = 96) and are 0.7998 and 0.5929 at the C-terminal sublibrary (n = 94), respectively. We also find that the distinct physiochemical properties function on the 14-3-3 ζ interaction. Overall, our results demonstrate that the computational methods based on QSAR analysis can be used for building the predictive models on the binding affinity of phosphopeptide against 14-3-3 ζ and contribute to the further research on clinical research.

CCS CONCEPTS

• **Applied computing** → Bioinformatics; Computational biology.

KEYWORDS

Computational peptidology, 14-3-3 proteins, Radom forest, QSAR study, Peptide microarray

ACM Reference Format:
Ying Fan, Xiaojun Wang, and Chao Wang . 2020. Building Random Forest QSAR Models for Affinity Identification of 14-3-3 ζ with Optimized Parameters. In *2020 9th International Conference on Bioinformatics and Biomedical Science (ICBBS '20), October 16–18, 2020, Xiamen, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3431943.3431951

*To whom the correspondence should be addressed. Email: wangchao@hkbu.edu.hk.

ICBBS '20, October 16–18, 2020, Xiamen, China

ACM ISBN 978-1-4503-8865-8/20/10…$15.00
https://doi.org/10.1145/3431943.3431951

1 INTRODUCTION

The human 14-3-3s protein family consists of seven distinct but related isoforms, which are named $\beta, \gamma, \epsilon, \eta, \sigma, \tau$, and ζ. These isoforms exist in a wide range of tissues and express in high abundance in human cells [9]. 14-3-3s binding to specific phosphoserine-containing motifs results in the assembly of important signaling complexes [22]. Through interaction with their ligands, 14-3-3s participate in the regulation of various cellular processes [7]. Furthermore, recent findings demonstrate that 14-3-3s as a key factor in promoting chemoresistance of cancer cells [10], [23], [31], [36], especially 14-3-3 ζ.

Although the sequence and structure of the proteins in 14-3-3s are highly conserved, many reports provide evidences that 13-3-3s may have distinct binding targets/ligands and specific functions for each isoform. For 14-3-3 ζ, a lot of confirmed binding ligands are the products of proto-oncogene/oncogene, e.g. c-Raf-1 [8], polyomavirus middle T antigen [34], BCR/ab1 [26], [44], hTERT [29], IGF-IR [15], [33], β-catenin [6]. These binding ligands have direct connections with the development of human cancer. These findings suggest that 14-3-3 ζ plays an important role in oncogenesis and metastasis. Except for cancer-related binding targets, 14-3-3 ζ presents high abundance at the transcriptional level in multiple types of tumor cells. In the lung cancer cell, the oncogenic activity during tumorigenesis is suppressed without 14-3-3 ζ [21]. Overexpression of 14-3-3 ζ occurs in breast disease and promote the disease transformation. The overexpression persists and can be found in the advanced stage of breast cancer [21]. In the mouse model of breast cancer, the relevance of 14-3-3 ζ overexpression to tumorigenesis is confirmed. It is found that MAPK/c-Jun signaling pathway is induced and p27 CDKI translation is inhibited by 14-3-3 ζ overexpression, which leads to the proliferation of cancer cells [25]. Additionally, high expression level of 14-3-3 ζ is found to be associated with disease recurrence and poor survival in the cancer patients received chemotherapy [13], [18], [41]. The protein structure of 14-3-3s presents a well-defined phosphopeptide binding pocket with a relatively small number of positively charged residues required for binding, which makes them a seemingly attainable PPI target. Although we have yet seen a 14-3-3 ζ inhibitor in the clinic application, current efforts are underway to develop small molecule disruptors of 14-3-3s interactions. Since the quantitative structure-activity relationships (QSAR) are widely used for predicting high activity peptides, we can utilize this method in the study of the nature of 14-3-3 binding interactions.

Here, a computational QSAR model to identify the optimized phosphopeptide sequence in 14-3-3 ζ is reported. For a given phosphopeptide sequence, the property of binding affinity can be seen as a kind of bioactivity against 14-3-3 ζ. The divide physicochemical property scores (DPPS) coupled with peptide microarray data were applied to generate a feature matrix for phosphopeptide sequence prediction. Based on the feature matrix, the binding affinity of 14-3-3 ζ against phosphopeptide is predicted by the random forest approach. The descriptors that have high impacts on the models of 14-3-3 ζ are estimated by statistical methods.

2 MATERIALS AND METHODS

2.1 Dataset

The dataset was extracted from a fingerprint peptide library coupled with seven mammalian 14-3-3 isoforms [17]. The different source of isoforms was used in this peptide library. The 14-3-3s binding affinity data in the library was generated by the fragment-based combinatorial peptide microarray and consisted of 1000-member phosphopeptide fragments with seven 14-3-3 isoforms, respectively. The data comprise two sublibraries, which have 500 members of N-terminal ($P_{-3}P_{-2}P_{-1}$-*p*S-$X_{+1}X_{+2}X_{+3}$) and 500 members of C-terminal ($X_{-3}X_{-2}X_{-1}$-*p*S-$P_{+1}P_{+2}P_{+3}$) sublibraries each. Peptide residues in the phosphopeptide fragments are designated as P and X, then numbered at subscript according to their proximity to the phosphoserine (*p*S) and negative if they are located N-terminal of it. Ten representative amino acids (Ala, Glu, Phe, Gly, Lys, Leu, Pro, Gln, Arg, Val) are used at the position of $P_{\pm 1/2}$, and among them five amino acids (Glu, Phe, Leu, Pro, Arg) at the position of $P_{\pm 3}$. X represents an isokinetic mixture of the rest 14 amino acids. The relative fluorescence values are generated by the fluorescence signals of Cy3-labeled 14-3-3s, which represent the affinity between phosphopeptide and 14-3-3s. To improve the data discrimination, raw relative fluorescence values were transformed by taking the logarithm to the base of two before analysis [40]. The processed data represents the relative binding affinity of 14-3-3 isoforms with each phosphopeptide fragment.

2.2 Feature Selection

The tripeptide sequences in the phosphopeptide fragments were extracted and used for building the feature matrix. The relative binding affinity of each tripeptide against 14-3-3 ζ was characterized by describing the residues with the descriptor of DPPS. Nonbonding effects, like electrostatic, van der Waals, hydrophobic interactions, and hydrogen bond play central roles in phosphopeptide interaction with 14-3-3s [42], [43]. DPPS descriptor is a kind of suitable way to be used as the amino acid descriptor here [35]. The DPPS descriptor is made up of ten variables (V1-V10) for describing the electronic property (V1-V4), steric property (V5-V6), hydrophobicity (V7-V8), and hydrogen bond (V9-V10) of all 20 amino acids. For tripeptide sequences in both N- and C-terminal sublibraries, physicochemical parameters of amino acid residue at each position were characterized by all ten DPPS descriptors. Thus, a total of 30 features were generated.

2.3 QSAR Modeling by Random Forest

The random forest (RF) algorithm was implemented according to the previous reports [2] by R packages of *randomForest* (version 4.6-14) and *caret* (version 6.0-86). The details of model construction by RF are demonstrated as follows. For the training set, several trees (*ntree*) were built randomly at the initial tree construction process. These trees worked as decision trees, namely "forest". The final prediction of the forest was made by the majority prediction from each of the trees. Each tree was used to construct the tree partition of feature space recursively and to make the tree prediction finally. In the given tree, some descriptors (*mtry*) were randomly selected among the feature space. The descriptors that were not selected in the tree were placed in the out-of-bag set. The best split in the selected descriptors by the CART criterion was chosen. This process was repeated until the given tree had less than a predefined number of features (*nodesize*). After tree partition was completed, the prediction at the new tree was computed. Finally, the predicted values of these trees were averaged to get the forest prediction. The ten-fold cross-validation method was used to estimating the performance of the predictive model on the different subset of the training set.

2.4 Measurements of Descriptors Contributions from Models

The contribution of each descriptor was estimated by the following methods [1], [2]. For a given tree, the prediction error (mean square error) on the out-of-bag set was computed and recorded. Then the same computed after permuting each descriptor. The value of the difference between the two of prediction error was then averaged over all trees and normalized by the standard error. Also, the normalization was denied if the standard error was equal to 0 for a descriptor.

2.5 Measurement of Prediction Quality

Two metrics, R^2 (coefficient of determination) and RMSE (root mean square error) were used to evaluate whether the model can predict the binding affinity precisely. Most data analysis steps were performed using the R statistical computing environment (R-project.org). Some analysis was also performed using MATLAB and Microsoft Excel.

3 RESULTS AND DISCUSSIONS

The study aims to build the RF models that can determine the binding affinity of phosphopeptide fragments against 14-3-3 ζ isoforms. It helps us to decode the ligand interaction at the residue level. The data contained the raw affinity values and peptide sequences were transformed and processed. The variables matrixes were established by DPPS. RF method was utilized to perform QSAR modeling and the predictive models for 14-3-3 ζ were created and compared with other approaches. The key hyperparameters in RF were explored by the training set to get the optimal performance. Based on the established RF model, the contributions of residues were analyzed.

3.1 Performance of the Predictive Models

The number of the relative binding affinity data was 481, which was split into the training set (80%) and the test set (20%). The training set had 385 peptide sequences and was used for building the RF model. The test set had 96 peptide sequences and was used for evaluating the power of the model in predicting the binding affinity.

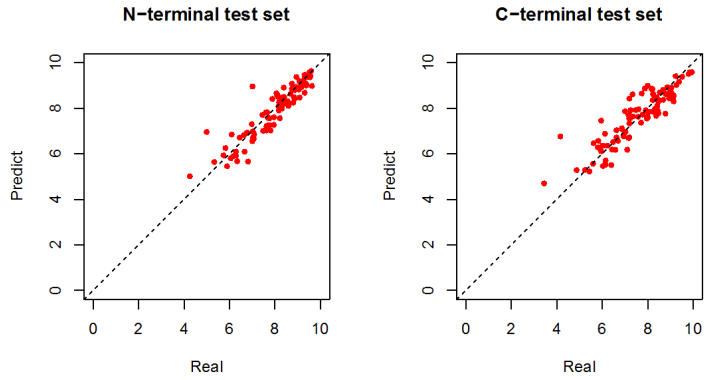

Figure 1: Scatter plot of the relative binding affinity values (Real) *versus* the prediction by the RF model (Predict). Left panel shows a comparison in the test set between the experimental affinity data from N-terminal sublibrary and the predicted one by RF approach (R^2 = 0.8458813, RMSE = 0.4789412). Right Panel shows a comparison in the test set from C-terminal sublibrary (R^2 = 0.7536284, RMSE = 0.6931992)

We first built the model by the default settings. To evaluate the predictive power of the model, the prediction was performed on the test set. The prediction error of RMSE and R^2 was 0.4789412 and 0.8458813 at N-terminal sublibrary (Figure 1). The same analysis was also applied to the dataset of C-terminal library, which had RMSE and R^2 value of 0.6931992 and 0.7536284. Low prediction errors are yielded by our method, and the prediction results indicate the robustness of the RF model.

Previously, we reported that the linear regression method is used for building the predictive models of 14-3-3s binding [5]. In those models, the relative low prediction power is obtained with optimized parameters for 14-3-3 ζ compared with current results. By using RF in this study, the prediction errors are decreased, and the quality of models is improved obviously. RF is an ensemble of decision trees, which is a non-linear approach like the neural network [4]. Unlike linear regression, there is no clear relationship or equation between the peptide activity and descriptors. However, the results indicate that the models developed by RF outperformed in predicting 14-3-3s binding, which suggests that there is no linear relationship exited between the binding affinity and descriptor matrix we constructed here.

3.2 Comparison with Other Machine Learning Based Approaches

Besides, different regression approaches, included MLR (multiple linear regression), GRNN (general regression neural network), and LASSO (least absolute shrinkage and selection operator), were applied to the training set for comparison purposes. To build the predictive model by GRNN, the cross-validation method was introduced to find the optimal value of the smooth parameter σ, which ranges from 1 to 6 with the step size 0.01. The parameter σ was determined by RMSE for each subset of cross-validation and averaged for all subsets, which was regarded as the optimal value to build the models. Finally, the parameter σ was set to 3.864 for N-terminal library and 3.034 for C-terminal library. To build the predictive model by LASSO, the cross-validation method was used to find the optimal value of λ. The parameter λ was set to 0.000433 for N-terminal library and 0.000632 for C-terminal library. Comparing the prediction results on the test set, it can be found that a better result and high performance are obtained by using RF to build the predictive model (Table 1). Especially, the absolute fit to the test set by random forest is better than the models developed by the other methods. This indicated that our analysis can outperform in building the regression models for the prediction of the binding affinity against 14-3-3 ζ.

Artificial neural network (ANN) is widely used for building QSAR models. A feedforward neural network (FNN) is firstly proposed as one of the simplest ANN methods [27]. During building the neural network, the quality of the model depends on the many iterations in the selection of parameters and network topology by the backpropagating method to reduce the errors, which make it a time-consuming process [11]. To overcome these drawbacks, one pass neural network learning algorithm is introduced as an alternative to speed up the training process [11]. For optimizing the network, the efforts mainly focus on the adjustment of a new parameter θ, which depends on the estimating kernel σ [11]. Thus, the rest of the customized parameters have been reduced to minimal. Single pass algorithm is implemented in the probabilistic neuronal network for solving regression problems, which is the general regression neural network (GRNN) [28], [32]. Generally speaking, GRNN has a shorter running time and a simpler optimization process. GRNN is taken in our considerations at the beginning of this study. However, the performance of GRNN is worse than the other approaches in the results. The reasons for the lower performance by GRNN are under investigation.

3.3 Comparison with Other Descriptors

For the development of QSAR models, it is important to screen the amino acid descriptors. To date, there are diverse types of amino acid descriptors are available from published literature or database.

Table 1: Comparison of the prediction by different approaches

Sublibrary	Statistics	RF	MLR	GRNN	LASSO
N-terminal sublibrary	RMSE	0.4650	0.6082	0.7788	0.6080
	R^2	0.8511	0.8244	0.5582	0.8246
C-terminal sublibrary	RMSE	0.5894	0.7324	0.8174	0.6132
	R^2	0.8017	0.7150	0.6430	0.7328

Table 2: Comparison of the prediction by multiple descriptors

Sublibrary	Statistics	DPPS	HESH	VHSE	G8
N-terminal sublibrary	RMSE	0.4650	0.4879	0.4736	0.4740
	R^2	0.8511	0.8328	0.8444	0.8395
C-terminal sublibrary	RMSE	0.5894	0.6260	0.5944	0.6882
	R^2	0.8017	0.7993	0.7791	0.7531

To estimate the influence of the variable matrix on the quality of the predictive models and improve the performance of modeling, several types of descriptors commonly used in the QSAR modeling were collected and introduced in the process of model building to obtain a suitable descriptor. These descriptors are VHSE [20], HESH [30], G8 [37] and DPPS [39]. These models were built by the RF approach in the same settings and were applied on the prediction of the test set to estimate the prediction power. The final results show that high prediction power and low errors are obtained by using DPPS as the amino acid predictors (Table 2). The rest of the other three descriptors also have comparable performances but lower than DPPS still.

It is commonly known that every type of descriptor has its advantages and limitations. VHSE has three types of variables (hydrophobicity, steric property, and electronic property). In the activity prediction of angiotensin-converting enzyme inhibitors, bradykinin-potentiating pentapeptides, bitter-tasting dipeptides, the application of VHSE descriptors could acquire satisfactory results [20], [38]. This suggests the advantageous use of VHSE for describing the structural variability in the protein-protein interactions. DPPS and HESH are newly published descriptors. More physicochemical parameters are considered both by DPPS (n = 171) and HESH (n = 119) than VHSE (n = 50). The property of the hydrogen bond is added into DPPS and HESH to describe the peptides. These two types of amino acid descriptors have been applied to build some good models with different regression methods [30], [35], [45]. The performance of prediction with DPPS is slightly better than the rest of the descriptors in this study. Thus, DPPS is chosen as the amino acid descriptor here.

3.4 The Optimization of Hyperparameters

To improve the performance of RF, the key hyperparameters were explored by the training set to get the optimal performance for the RF models, namely *mtry* (the number of candidate variables considered at each split), *ntree* (the number of trees), and *nodesize* (the minimal size of a node should have to be split). These hyperparameters are involved in controlling the structure of each RF tree. The overall task of optimization was to evaluate the influence of these key hyperparameters on the RF models with high prediction performance that can be defined by R^2 and RMSE. The ten-fold cross-validation method was used for every specific hyperparameter to consolidate our understanding of the influence of hyperparameter to the model disturbance.

Experiments for the hyperparameter *nodesize* (1-10), *ntree* (10-500), *mtry* (2-30) were implemented on the N- and C-terminal training set with fixing the other environments, and then choose the optimal value for getting the best performance.

3.4.1 Ntree. By default, *ntree* is set to 1000. The values of *ntree* that ranged from 10 to 500 were explored here. The values of RMSE are decline sharply following *ntree* increases (Figure 2A). Theoretically, the number of trees should be set as high as possible since more trees are capable of getting clear predictions [24]. Following the value of *ntree* increased, especially larger than 200, it can be found that the prediction errors of RMSE are not decreased apparently and numerically. Considering that the computation time increased linearly with *ntree*, the default value of *ntree* was not used here. The value of *ntree* was set to 90 for N-terminal sublibrary and 170 for C-terminal sublibrary.

3.4.2 Nodesize. By default, *nodesize* is set to 5. It can be found that the values of RMSE vary from 1 to 10 but get the lowest value at 3 for N-terminal sublibrary (Figure 2B). For C-terminal training set, the optimal value is 6. The *nodesize* parameter defines the minimum size of a terminal node. The lower value leads to trees with a larger depth, which means that more splits are performed until the end. It is believed to generally provide good results by the default setting [3], [12]. But the prediction power of the model can be improved by the optimization process [16]. It is reported that increasing the number of noise variables leads to a higher optimal *nodesize* [19]. The low value of *nodesize* suggests that most of the data can be used in model development.

3.4.3 Mtry. After defining the value of *nodesize* and *ntree*, the number of descriptors for splitting *mtry* was optimized. In default settings, the value of *mtry* is set to one-third of the total number of descriptors. Our results indicated that the optimized value of *mtry* is 9 for N-terminal sublibrary and 25 for C-terminal sublibrary (Figure 2C). Smaller *mtry* value leads to less correlated trees, which are

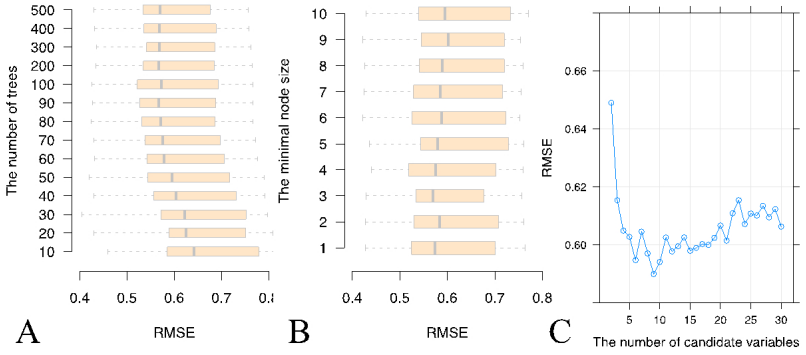

Figure 2: Performance comparison of models with different hyperparameters by cross-validation. (A) The RMSE distributions with different values of *ntree* computed by the 10-fold cross validation are illustrated. The lines inside of the boxes indicate the median value. (B) The RMSE distributions with different values of *nodesize* are illustrated (C) The changes of the RMSE following the increased *mtry* values

more different from each other. On some occasions, the important variables might be masked during forest construction.

Finally, the optimized hyperparameters of *ntree*, *nodesize*, and *mtry* were defined as 170, 3, 9 for N-terminal sublibrary, and 90, 6, 25 for C-terminal sublibrary. These hyperparameters were used for model development and applied to the test set. The prediction errors of R^2 were 0.8511 for N-terminal sublibrary and 0.8017 for C-terminal sublibrary, both of which are higher than the results with the default hyperparameters. Similarly, the values of RMSE were accordingly decreased to 0.4650 for N-terminal sublibrary and 0.5894 for C-terminal sublibrary, respectively.

3.5 The Contributions of the Descriptors

After building the models, the descriptors that contribute more to the predictive power were estimated by statistical approaches. Descriptors with high contributions are drivers of the predictions and their values have a significant impact on the modeling process. All descriptors at residues are illustrated (Figure 3). For the C-terminal sublibrary, the physicochemical property of the hydrogen bond at position +1 is important for predicting that the binding affinity of phosphopeptide sequence against 14-3-3 ζ. For the N-terminal sublibrary, the electronic property at position -2 is the key factor. It can be found that the weak connections of some residues exist, e.g. position -3, +2, and +3, which suggests that the amino acid at position -1, -2 and +2 functions as conserved portions in the 14-3-3 ζ interactions. This is consistent with our previously developed models by linear methods and indicates that the contribution of position -2 and -1 is associated with the electronic property of residues, and the contribution of position +1 comes from the part of electronic property and hydrophobicity [5].

The sequence analysis of 14-3-3s binding sites with Leu/Arg at position -5 in mammalian indicates that the frequently occurred types of the amino acid at position +1 are Glu, Leu, Phe, Ser, Val, Ala, and Trp [14]. Except for Ser and Trp, the rest of the side chains of amino acids are incapable to form hydrogen bond and have low value in the descriptor space V10 of DPPS. And the frequently

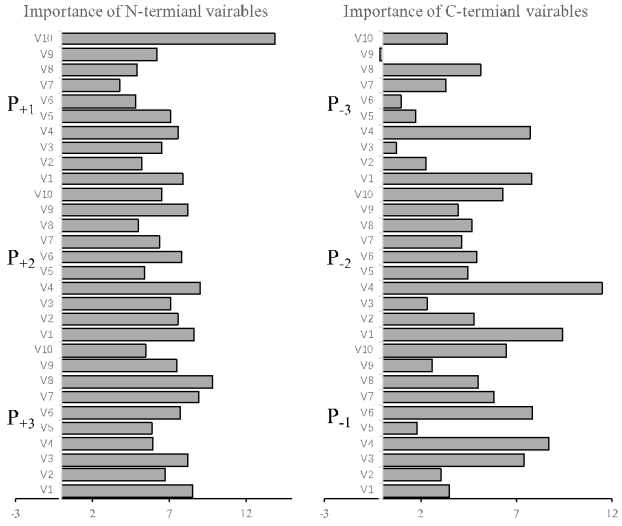

Figure 3: The contributions of the descriptors in the RF models. Left panel presents that the distribution of the contribution of all the descriptors (V1~V10) to the predictive models for the residues (P_{+1}, P_{+2}, P_{+3}) in N-terminal sublibrary. Right panel presents that the distribution of the contribution of the descriptors for the residues (P_{-1}, P_{-2}, P_{-3}) in C-terminal sublibrary

occurred types of the amino acid at position -2 are Ser, Arg, and His. Both of these are polar residues and have negative values in the descriptor space V4 of DPPS. The contribution of descriptor V4 is the highest in all N-terminal residues. It is plausible that the electronic property is essential in 14-3-3 ζ interaction.

4 CONCLUSIONS

In this study, the development of the predictive QSAR models that can be used for predicting the binding affinity of phosphopeptide

fragments against 14-3-3 ζ is described. By introducing the RF approach with the optimized hyperparameters in the QSAR analysis, excellent results were obtained on the test set. The physicochemical properties in residues on the contribution of binding were identified. Based on these solid results, the predictive models on other 14-3-3 isoforms will be also developed. And the binding affinity of phosphopeptide sequences against 14-3-3s can be predicted and analyzed. The predictions will help us elucidate the mechanism of 14-3-3s interaction and could be applied in further studies.

ACKNOWLEDGMENTS

This research was funded by Shenzhen Science and Technology Innovation Commission of China (grant number: JCYJ20180307124010740, JCYJ20170817115152903) and Natural Science Foundation of Guangdong Province, China (grant number: 2018A030310693).

REFERENCES

[1] Leo Breiman. OUT-OF-BAG ESTIMATION. Technical Report. University of California at Berkeley. Retrieved Jul 27, 2020 from https://www.stat.berkeley.edu/pub/users/breiman/OOBestimation.pdf

[2] Leo Breiman. 2001. Random forests. Mach. Learn. 45, 1 (October 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 1 (2006), 3. https://doi.org/10.1186/1471-2105-7-3

[4] Frank Emmert-Streib. 2012. Statistical modelling of molecular descriptors in QSAR/QSPR. John Wiley & Sons.

[5] Ying Fan, Xiaojun Wang, and Chao Wang. 2020. DEVELOPMENT OF PREDICTIVE QSAR MODELS ON THE PHOSPHOPEPTIDE BINDING AFFINITY AGAINST 14-3-3 ISOFORMS A PREPRINT. bioRxiv (July 2020), 2020.07.24.217752. https://doi.org/10.1101/2020.07.24.217752

[6] Dexing Fang, David Hawke, Yanhua Zheng, Yan Xia, Jill Meisenhelder, Heinz Nika, Gordon B Mills, Ryuji Kobayashi, Tony Hunter, and Zhimin Lu. 2007. Phosphorylation of β-catenin by AKT promotes β-catenin transcriptional activity. J. Biol. Chem. 282, 15 (2007), 11221–11229. https://doi.org/10.1074/jbc.M611871200

[7] Robert J. Ferl, Michael S. Manak, and Matthew F. Reyes. 2002. The 14-3-3s. Genome Biology 3, reviews3010.1. https://doi.org/10.1186/gb-2002-3-7-reviews3010

[8] E Freed, F McCormick, and R Ruggieri. 1994. Proteins of the 14-3-3 family associate with Raf and contribute to its activation. In Cold Spring Harbor Symposia on Quantitative Biology, 187–193. https://doi.org/10.1101/SQB.1994.059.01.023

[9] Haian Fu, Romesh R Subramanian, and Shane C Masters. 2000. 14-3-3 proteins: structure, function, and regulation. Annu. Rev. Pharmacol. Toxicol. 40, 1 (2000), 617–647. https://doi.org/https://doi.org/10.1146/annurev.pharmtox.40.1.617

[10] Cathie Garnis, Bradley P Coe, Adrian Ishkanian, Lewei Zhang, Miriam P Rosin, and Wan L Lam. 2004. Novel Regions of Amplification on 8q Distinct from the MYC Locus and Frequently Altered in Oral Dysplasia and Cancer. Genes Chromosom. Cancer 39, 1 (2004), 93–98. https://doi.org/10.1002/gcc.10294

[11] Mohammad Gholamrezaei and Kaveh Ghorbanian. 2007. Rotated general regression neural network. In IEEE International Conference on Neural Networks - Conference Proceedings, 1959–1964. https://doi.org/10.1109/IJCNN.2007.4371258

[12] Benjamin A Goldstein, Eric C Polley, and Farren B S Briggs. 2011. Random forests for genetic association studies. Stat. Appl. Genet. Mol. Biol. 10, 1 (2011).

[13] Y Gu, K Xu, C Torre, M Samur, B G Barwick, M Rupji, J Arora, P Neri, J Kaufman, A Nooka, L. Bernal-Mizrachi, P. Vertino, S. Y. Sun, J. Chen, N. Munshi, H. Fu, J. Kowalski, L. H. Boise, and S. Lonial. 2018. 14-3-3σ binds the proteasome, limits proteolytic function and enhances sensitivity to proteasome inhibitors. Leukemia 32, 3 (2018), 744–751. https://doi.org/10.1038/leu.2017.288

[14] Catherine Johnson, Sandra Crowther, Margaret J Stafford, David G Campbell, Rachel Toth, and Carol MacKintosh. 2010. Bioinformatic and experimental survey of 14-3-3-binding sites. Biochem. J. 427, 1 (March 2010), 69–78. https://doi.org/10.1042/BJ20091834

[15] Shiwei Li, Mariana Resnicoff, and Renato Baserga. 1996. Effect of mutations at serines 1280-1283 on the mitogenic and transforming activities of the insulin-like growth factor I receptor. J. Biol. Chem. 271, 21 (1996), 12254–12260. https://doi.org/10.1074/jbc.271.21.12254

[16] Yi Lin and Yongho Jeon. 2006. Random forests and adaptive nearest neighbors. J. Am. Stat. Assoc. 101, 474 (2006), 578–590. https://doi.org/10.1198/016214505000001230

[17] Candy H S Lu, Hongyan Sun, Farhana B. Abu Bakar, Mahesh Uttamchandani, Wei Zhou, Yih Cherng Liou, and Shao Q. Yao. 2008. Rapid affinity-based fingerprinting of 14-3-3 isoforms using a combinatorial peptide microarray. Angew. Chemie - Int. Ed. 47, (2008), 7438–7441. https://doi.org/10.1002/anie.200801395

[18] Jing Lu, Hua Guo, Warapen Treekitkarnmongkol, Ping Li, Jian Zhang, Bin Shi, Chen Ling, Xiaoyan Zhou, Tongzhen Chen, Paul J Chiao, and others. 2009. 14-3-3 ζ cooperates with ErbB2 to promote ductal carcinoma in situ progression to invasive breast cancer by inducing epithelial-mesenchymal transition. Cancer Cell 16, 3 (2009), 195–207.

[19] Kathryn L Lunetta, L Brooke Hayward, Jonathan Segal, and Paul van Eerdewegh. 2004. Screening large-scale association study data: Exploiting interactions using random forests. BMC Genet. 5, 1 (2004), 32. https://doi.org/10.1186/1471-2156-5-32

[20] Hu Mei, Yuan Zhou, Li Li Sun, and Zhi Liang Li. 2004. A new descriptor of amino acids and its application in peptide QSAR. Acta Phys. - Chim. Sin. 20, 8 (2004), 821–825. https://doi.org/10.3866/PKU.WHXB20040808

[21] Maarten Niemantsverdriet, Koen Wagner, Mijke Visser, and Claude Backendorf. 2008. Cellular functions of 14-3-3ζ in apoptosis and cell adhesion emphasize its oncogenic character. Oncogene 27, 9 (2008), 1315–1319. https://doi.org/10.1038/sj.onc.1210742

[22] Kl Pennington, Ty Chan, Mp Torres, and Jl Andersen. 2018. The dynamic and stress-adaptive signaling hub of 14-3-3: emerging mechanisms of regulation and context-dependent protein–protein interactions. Oncogene 37, 42 (2018), 5587–5604. https://doi.org/10.1038/s41388-018-0348-3

[23] Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne Lise Børresen-Dale, and Patrick O Brown. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc. Natl. Acad. Sci. U. S. A. 99, 20 (2002), 12963–12968. https://doi.org/10.1073/pnas.162471999

[24] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9, 3 (May 2019). https://doi.org/10.1002/widm.1301

[25] Sumaiyah K Rehman, Shau Hsuan Li, Shannon L Wyszomierski, Qingfei Wang, Ping Li, Ozgur Sahin, Yi Xiao, Siyuan Zhang, Yan Xiong, Jun Yang, Hai Wang, Hua Guo, Jitao D. Zhang, Daniel Medina, William J. Muller, and Dihua Yu. 2014. 14-3-3ζ orchestrates mammary tumor onset and progression via miR-221-mediated cell proliferation. Cancer Res. 74, 1 (2014), 363–373. https://doi.org/10.1158/0008-5472.CAN-13-2016

[26] Gary W Reuther, Haian Fu, Larry D Cripe, R John Collier, and Ann Marie Pendergast. 1994. Association of the protein kinases c-Bcr and Bcr-Abl with proteins of the 14-3-3 family. Science (80-.). 266, 5182 (1994), 129–133. https://doi.org/10.1126/science.7939633

[27] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. Nature 323, 6088 (1986), 533–536. https://doi.org/10.1038/323533a0

[28] Henrik Schløler and Uwe Hartmann. 1992. Mapping neural network derived from the parzen window estimator. Neural Networks 5, 6 (1992), 903–909. https://doi.org/10.1016/S0893-6080(05)80086-3

[29] Hiroyuki Seimiya, Hiroko Sawada, Yukiko Muramatsu, Mayuko Shimizu, Kumiko Ohko, Kazuhiko Yamane, and Takashi Tsuruo. 2000. Involvement of 14-3-3 proteins in nuclear localization of telomerase. EMBO J. 19, 11 (2000), 2652–2661. https://doi.org/10.1093/emboj/19.11.2652

[30] Mao Shu, Hu Mei, Shanbin Yang, Limin Liao, and Zhiliang Li. 2009. Structural parameter characterization and bioactivity simulation based on peptide sequence. QSAR Comb. Sci. 28, 1 (2009), 27–35. https://doi.org/10.1002/qsar.200710169

[31] Pranav Sinha, Sandra Kohl, Jochen Fischer, Gero Hütter, Monika Kern, Eckard Köttgen, Manfred Dietel, Hermann Lage, Martina Schnölzer, and Dirk Schadendorf. 2000. Identification of novel proteins associated with the development of chemoresistance in malignant melanoma using two-dimensional electrophoresis. Electrophoresis 21, 14 (2000), 3048–3057. https://doi.org/10.1002/1522-2683(20000801)21:14<3048::AID-ELPS3048>3.0.CO;2-W

[32] Donald F. Specht. 1991. A General Regression Neural Network. IEEE Trans. Neural Networks 2, 6 (1991), 568–576. https://doi.org/10.1109/72.97934

[33] Susan L Spence, Bhakta R Dey, Cheryl Terry, Paul Albert, Peter Nissley, and Richard W Furlanetto. 2003. Interaction of 14-3-3 proteins with the insulin-like growth factor I receptor (IGFIR): Evidence for a role of 14-3-3 proteins in IGFIR signaling. Biochem. Biophys. Res. Commun. 312, 4 (2003), 1060–1066. https://doi.org/10.1016/j.bbrc.2003.11.043

[34] W. Su, W. Liu, B. S. Schaffhausen, and Thomas M Roberts. 1995. Association of Polyomavirus middle tumor antigen with phospholipase C-γ1. J. Biol. Chem. 270, 21 (1995), 12331–12334. https://doi.org/10.1074/jbc.270.21.12331

[35] F. Tian, L. Yang, F. Lv, Q. Yang, and P. Zhou. 2009. In silico quantitative prediction of peptides binding affinity to human MHC molecule: An intuitive quantitative structure-activity relationship approach. Amino Acids 36, 3 (2009), 535–554. https://doi.org/10.1007/s00726-008-0116-8

[36] Alexei Vazquez, Lukasz F Grochola, Elisabeth E. Bond, Arnold J Levine, Helge Taubert, Thomas H Müller, Peter Würl, and Gareth L Bond. 2010. Chemosensitivity profiles identify polymorphisms in the p53 network genes 14-3-3τ and CD44

that affect sarcoma incidence and survival. Cancer Res. 70, 1 (2010), 172–180. https://doi.org/10.1158/0008-5472.CAN-09-2218

[37] Xiao Yu Wang, Juan Wang, Yong Hu, Yong Lin, Mao Shu, Li Wang, Xiao Ming Cheng, and Zhi Hua Lin. 2011. Predicting the activity of peptides based on amino acid information. J. Chinese Chem. Soc. 58, 7 (2011), 877–883. https://doi.org/10.1002/jccs.201190139

[38] Xiaoyu Wang, Juan Wang, Yong Lin, Yuan Ding, Yuanqiang Wang, Xiaoming Cheng, and Zhihua Lin. 2011. QSAR study on angiotensin-converting enzyme inhibitor oligopeptides based on a novel set of sequence information descriptors. J. Mol. Model. 17, 7 (2011), 1599–1606. https://doi.org/10.1007/s00894-010-0862-x

[39] S. Wold, J. Jonsson, M. Sjörström, M. Sandberg, and S. Rännar. 1993. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal. Chim. Acta 277, 2 (1993), 239–253. https://doi.org/10.1016/0003-2670(93)80437-P

[40] Jin Xiong. 2006. Essential bioinformatics. Cambridge University Press.

[41] Jia Xu, Sunil Acharya, Ozgur Sahin, Qingling Zhang, Yohei Saito, Jun Yao, Hai Wang, Ping Li, Lin Zhang, Frank J Lowery, Wen Ling Kuo, Yi Xiao, Joe Ensor, Aysegul A. Sahin, Xiang H.F. Zhang, Mien Chie Hung, Jitao David Zhang, and Dihua Yu. 2015. 14-3-3ζ Turns TGF-β's Function from Tumor Suppressor to Metastasis Promoter in Breast Cancer by Contextual Changes of Smad Partners from p53 to Gli2. Cancer Cell 27, 2 (2015), 177–192. https://doi.org/10.1016/j.ccell.2014.11.025

[42] Michael B. Yaffe, Katrin Rittinger, Stefano Volinia, Paul R. Caron, Alastair Aitken, Henrik Leffers, Steven J. Gamblin, Stephen J. Smerdon, and Lewis C. Cantley. 1997. The structural basis for 14-3-3: phosphopeptide binding specificity. Cell 91, 7 (December 1997), 961–971. https://doi.org/10.1016/s0092-8674(00)80487-0

[43] Xiaowen Yang, Wen Hwa Lee, Frank Sobott, Evangelos Papagrigoriou, Carol V Robinson, J Günter Grossmann, Michael Sundström, Declan a Doyle, and Jonathan M Elkins. 2006. Structural basis for protein-protein interactions in the 14-3-3 protein family. Proc. Natl. Acad. Sci. U. S. A. 103, 35 (2006), 17237–17242. https://doi.org/10.1073/pnas.0605779103

[44] Kiyotsugu Yoshida, Tomoko Yamaguchi, Tohru Natsume, Donald Kufe, and Yoshio Miki. 2005. JNK phosphorylation of 14-3-3 proteins regulates nuclear targeting of c-Abl in the apoptotic response to DNA damage. Nat. Cell Biol. 7, 3 (2005), 278–285. https://doi.org/10.1038/ncb1228

[45] Peng Zhou, Xiang Chen, Yuqian Wu, and Zhicai Shang. 2010. Gaussian process: An alternative approach for QSAM modeling of peptides. Amino Acids 38, 1 (2010), 199–212. https://doi.org/10.1007/s00726-008-0228-1.