International Journal of Biological Macromolecules 162 (2020) 931–934

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

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

Sequence based prediction of pattern recognition receptors by using feature selection technique

Pengmian Feng [a,*], Lijing Feng [b]

[a] School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China
[b] School of Sciences, North China University of Science and Technology, Tangshan 063000, China

ARTICLE INFO

Article history:
Received 22 April 2020
Received in revised form 23 June 2020
Accepted 24 June 2020
Available online 26 June 2020

Keywords:
Pattern recognition receptor
Random forest
Feature selection
Amino acid composition
CTD

ABSTRACT

Pattern recognition receptors (PRRs) play crucial roles in the innate immune system, and are able to identify pathogen-associated molecular patterns and damage-associated molecular patterns. Accurate identification of PRRs is essential for understanding their functions. In the present work, a random forest based method was proposed to identify PRRs, in which the sequences were formulated by using the optimal features. In the 10-fold cross validation test, an accuracy of 80.95% was obtained in identifying PRRs. We wish that the proposed method will become a useful tool, or at least play a complementary role to the existing predictors for identifying PRRs.

1. Introduction

Pattern recognition receptors (PRRs), mainly expressed by cells of the innate immune system, are able to identify pathogen-associated molecular patterns and damage-associated molecular patterns [1]. According to their ligand specificity and function, PRRs can be classified into four families, namely Toll-like receptors (TLRs), NOD-like receptors (NLRs), RIG-1-like receptors (RLRs), and C-type lectin receptors (CLRs), each of them has their specific functions [2].

PRRs play crucial roles in the innate immune system, such as activation of complement, phagocytosis, initiation of cell activation [3] and inflammatory signal transduction [4], induction of apoptosis [5]. Recent studies have also demonstrated that PRRs are associated with a series of diseases, such as rheumatic disease [6], cancer, Alzheimer's disease [7], acute myeloid leukemia [8], asthma [9], etc. However, our knowledge about the functions of PRRs and their mechanisms associated with diseases are far from satisfactory.

Identifying PRRs is beneficial to reveal their functions and mechanisms. Although a series of experimental methods have been proposed to identify PRRs, such as immunofluorescence [10] and PAMP binding assay [11], they are still time consuming and costly. It should be pointed out that the experimental methods provided invaluable data for building computational methods. Benefiting from these experimental data, a database called PRRDB 2.0 was constructed by Kaur et al., which specially deposits pattern-recognition receptors and their ligands [12]. Recently, based on PRRDB 2.0, Kaur et al. developed a computational method, called PRRpred, to predict PRRs [13]. However, sequences in the dataset used to train their models share similarities greater than 40%, which might lead to the overestimation of the prediction performance. Therefore, it's necessary to develop a new model based on a high quality dataset.

In this work, we proposed a random forest based method to identify PRRs based on the optimal features obtained by using the feature selection method. The rest of the work is organized as following: (1) construct a benchmark dataset; (2) formulate the sequences in the dataset by using sequence encoding scheme; (3) determine the optimal features by using feature selection method; (4) evaluate the performance of the proposed method.

2. Materials and methods

2.1. Dataset

Based on the PRRDB 2.0 database [12] and the curated protein sequence database Swiss-Prot [14], Kaur et al. built a dataset including 179 PRRs and 274 non-PRRs, which was used to train and test their method [13]. However, by performing the sequence clustering program CD-HIT [15] on this dataset, it was found that there exist sequences with identify >40%. As indicated in the previous work [16], if a predictor was trained on a dataset containing redundant sequences, it might yield misleading results with overestimated accuracy [17]. Therefore, we

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

https://doi.org/10.1016/j.ijbiomac.2020.06.234
0141-8130/

deleted those sequences with identify >40% and obtained a new dataset containing 109 PRRs and 185 non-PRRs, which was used to train and test the method proposed in the present work.

2.2. Amino acids composition

Considering the distinct amino acids composition between PRRs and non-PRRS, the amino acids composition (AAC) [18] was employed to encode the sequences in the dataset, which is defined in Eq. (1).

$$\mathbf{F} = [f_1 f_2 \ldots f_i \ldots f_{20}]^T \tag{1}$$

where T is the transpose operator, f_i is the frequency of the i-th ($i = 1, 2, 3, \ldots, 20$) amino acid in a sequence.

2.3. Composition transition distribution

The Composition Transition Distribution (CTD) describes the amino acid distribution in terms of the structural or physicochemical property in the sequences [19]. At present, 13 kinds of physicochemical properties have been used for computing CTD, namely normalized Van der Waals Volume, polarity, polarizability, charge, secondary structures, solvent accessibility and 7 types of hydrophobicity.

For each kind of the property, the 20 amino acids can be divided into three groups, i.e. hydrophobic, neutral and polar. The composition (CTD-C) is the frequency value of hydrophobic, neutral and polar groups in a sequence [20]. The transition (CTD-T) is the frequency of an amino acid of one specific group followed by an amino acid of another kind of group. The distribution (CTD-D) represents the distribution of each property in a sequence. Five values can be yielded for each of the three groups, namely the first residue, 25% residues, 50% residues, 75% residues, and 100% residues in a sequence of a given specific property. Accordingly, we can obtain a 39, 39 and 195 dimensional feature vector based on CTD-C, CTD-T and CTD-D, respectively. In the present study, the iFeature [20] was to calculate the CTD.

2.4. Random Forest

The Random Forest (RF) is a meta-learning algorithm. Owing to its merits, namely easy training and fast prediction, RF is widely used for classification in bioinformatics [21–25]. RF consists an ensemble of separately decision trees trained on a subset of randomly selected instances from the given training set. The prediction results of RF are based on the ensemble of those decision trees. In the present work, the RF was performed by using WEKA [26].

2.5. Feature selection

Although we can include more information through a high-dimension feature vector by using different kind of features, such a vector often includes redundant or irrelevant information which might lead to over-fitting problems and reduce the generalization capacity of the model [27–29]. In order to solve such problems, a series of feature selection methods have been proposed to select optimal features, such as analysis of variance (ANOVA) [30–32], diffusion maps, maximum relevance maximum distance (MRMD), and so on. Compared with the other feature selection methods, MRMD can not only provide the contributions of each feature, but also reported the predictive accuracy based on the optimal feature set. Therefore, in the present work, the MRMD was used to select the optimal features. The main idea of MRMD is to measure the feature redundancy and determine the relevance between the features and target class. More detail about the MRMD algorithm can be found in Zou et al.'s work [33,34].

2.6. Evaluation matrix

The proposed method was evaluated by the commonly used four metrics [35–37], namely Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), and Mathew's correlation coefficient (MCC), which are defined as following,

$$\begin{cases} S_n = \dfrac{TP}{TP+FN} \times 100\% \\ S_p = \dfrac{TN}{TN+FP} \times 100\% \\ Acc = \dfrac{TP+TN}{TP+FN+TN+FP} \times 100\% \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \end{cases} \tag{2}$$

where TP, TN, FP and FN represent true positive, true negative, false positive, and false negative, respectively.

Furthermore, the area under the ROC curve (auROC) was also used to evaluate the performance quality of the proposed method [38]. Its value of 0.5 is a random prediction, while a value of 1 represents a perfect prediction.

3. Result and discussion

3.1. Comparison of different features for identifying PRRs

In order to demonstrate the effectiveness of different features for identifying PRRS, we firstly compared the prediction performance of AAC, CTD-C, CTD-T, and CTD-D by using RF. The 10-fold cross validation test results were listed in Table 1. It was found that the model based on AAC yielded the best accuracy of 78.91%. The model based on CTD-D obtained the lowest accuracy of 69.73% for identifying PRRs. Considering the dimension of the CTD-D derived feature is higher than that of the other kinds of features, we omitted it the following analysis.

3.2. Performance of fusing multiple features

To demonstrate whether the feature fusion strategy could improve the performance for identifying PRRs, we built another RF based model by combining the three kinds of features, namely AAC, CTD-C and CTD-T. In order to reduce the feature dimension and to build a robust and efficient predictive model, we used the MRMD method together with the Incremental Feature Selection (IFS) strategy to select optimal features.

The 98 features were firstly ranked by using the MRMD method. The 98 ranked features were then added one by one from higher to lower rank. We repeated this procedure 98 times. For each time, a RF model was built and was evaluated by using the 5-fold cross-validation test. A IFS was plotted, in which the abscissa is the number of features and the ordinate is the corresponding accuracy. The optimal features were obtained when the accuracy reaches its maximum. As shown in Fig. 1, when the top ranked 33 features were used to encode the samples, the accuracy reaches its maximum of 81.63%. Therefore, a computational model was built based on these 33 optimal features. In the 10-fold cross validation test, the proposed method obtained an accuracy

Table 1
Predictive results for identifying PRRs by using different features.

Features	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
AAC	55.96	92.43	78.91	0.54	0.82
CTD-C	55.05	89.73	76.87	0.49	0.80
CTD-T	44.04	87.03	71.09	0.35	0.77
CTD-D	43.12	85.41	69.73	0.32	0.73

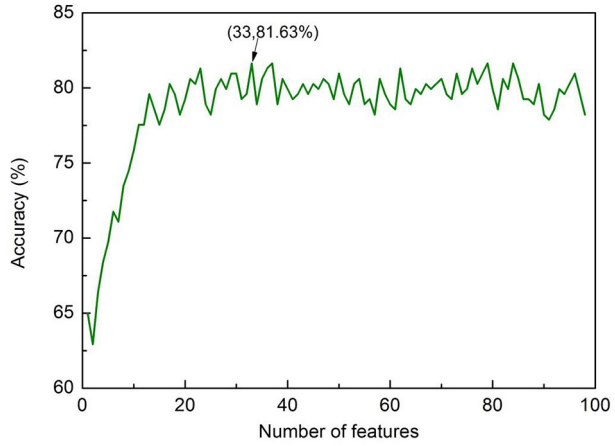

Fig. 1. The IFS curve for determining optimal features for identifying PRRs. The abscissa is the number of features and the ordinate is the corresponding accuracy.

of 80.95% with the sensitivity of 62.39%, specificity of 91.89%, MCC of 0.58, and AUC of 0.84. The result thus obtained is better than that of the model based on a single kind of feature. However, the sensitivity is still unsatisfactory. The lower sensitivity might be due to the limited number of PRRs in the dataset.

3.3. Feature contribution analysis

In order to provide an overview of the contributions of the features for identifying PRRs, the features together with their scores obtained by MRMD method was illustrated in Fig. 2. It was found that 17 of the 20 AAC, 15 of the CTD-C, and 1 of the CTD-D are the optimal features for the model, which were highlighted in red, orange and green, respectively. These results indicate that amino acid composition and their physicochemical properties make great contributions for identifying PRRs.

3.4. Comparison with other classifiers

Since the method proposed by Kaur et al. is trained and tested on a dataset containing redundant sequence, it's not fair to compare the proposed method with Kaur et al.'s method. To demonstrate its superiority,

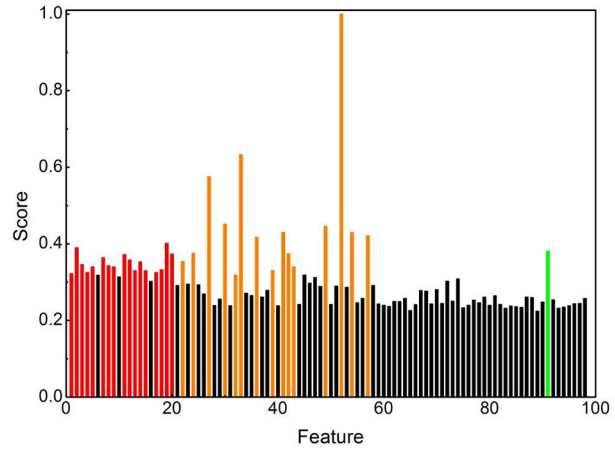

Fig. 2. Illustration of the three kinds of features used to identify PRRs. The abscissa is the features, from 1 to 20 is AAC, from 21 to 59 is CTD-C, and from 60 to 98 is CTD-T. The ordinate is the score obtained by MRMD. The optimal features are highlighted in red, orange and green for AAC, CTD-C, and CTD-T, respectively.

Table 2
Comparison of different methods for identifying PRRs.

Algorithm	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
BayesNet	66.06	76.76	72.79	0.42	0.78
Native Bayes	67.89	72.97	71.09	0.40	0.77
J48	55.05	77.84	69.39	0.33	0.66
LogitBoost	60.55	82.70	74.49	0.44	0.78
Logistic	64.22	85.41	77.55	0.51	0.81
RF	62.39	91.89	80.95	0.58	0.84

a comparison was performed between the proposed RF based method and the commonly used machine learning algorithms, namely BayesNet, Native Bayes, J48, LogitBoost, Logistic. All these algorithms were implemented in WEKA [39] with the default parameters. The 10-fold cross validation test results of these methods for identifying PRRs in the benchmark dataset was reported in Table 2. Although the sensitivity obtained by BayesNet, Native Bayes and Logistic is higher than that of RF, the other metrics defined in Eq. (2) yielded by those algorithms are all lower than that of RF for identifying PRRs. These results indicate that our proposed method is promising for identifying PRRs.

4. Conclusion

In this work, we proposed a machine learning based method to identify PRRs, in which the amino acid composition and the composition transition distribution (CTD) were used to encode the sequences in the dataset. In order to remove redundant and noise features, the MRMD method [33] was performed to select the optimal features. Comparative results demonstrated that the proposed RF based method is superior to the other machine learning methods for PRRs. However, it should be point out that the sensitivity of the current method is still unsatisfactory. In the future, we will try to improve the performance by collecting the samples from different resources to enlarge the data size and also adopt the deep learning method [40–44] to develop computational models.

Author statement

Pengmian Feng conceived and designed the experiments; Pengmian Feng and Lijing Feng performed the experiments; Pengmian Feng wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 31771471).

References

[1] T. Kawai, S. Akira, The roles of TLRs, RLRs and NLRs in pathogen recognition, Int. Immunol. 21 (4) (2009) 317–337.
[2] O. Takeuchi, S. Akira, Pattern recognition receptors and inflammation, Cell 140 (6) (2010) 805–820.
[3] G.P. Amarante-Mendes, S. Adjemian, L.M. Branco, L.C. Zanetti, R. Weinlich, K.R. Bortoluci, Pattern recognition receptors and the host cell death molecular machinery, Front. Immunol. 9 (2018) 2379.
[4] K. Newton, V.M. Dixit, Signaling in innate immunity and inflammation, Cold Spring Harb. Perspect. Biol. 4 (3) (2012).
[5] I. Tennant, J.D. Pound, L.A. Marr, J.J. Willems, S. Petrova, C.A. Ford, M. Paterson, A. Devitt, C.D. Gregory, Innate recognition of apoptotic cells: novel apoptotic cell-associated molecular patterns revealed by crossreactivity of anti-LPS antibodies, Cell Death Differ. 20 (5) (2013) 698–708.
[6] L.M. Mullen, G. Chamberlain, S. Sacre, Pattern recognition receptors as potential therapeutic targets in inflammatory rheumatic disease, Arthritis research & therapy 17 (2015) 122.
[7] M.M. Wang, D. Miao, X.P. Cao, L. Tan, L. Tan, Innate immune activation in Alzheimer's disease, Annals of translational medicine 6 (10) (2018) 177.
[8] N.J. Buteyn, R. Santhanam, G. Merchand-Reyes, R.A. Murugesan, G.M. Dettorre, J.C. Byrd, A. Sarkar, S. Vasu, B.L. Mundy-Bosse, J.P. Butchar, S. Tridandapani, Activation of the intracellular pattern recognition receptor NOD2 promotes acute myeloid

leukemia (AML) cell apoptosis and provides a survival advantage in an animal model of AML, J. Immunol. 204 (7) (2020) 1988–1997.
[9] T.H. Lee, H.J. Song, C.S. Park, Role of inflammasome activation in development and exacerbation of asthma, Asia Pacific Allergy 4 (4) (2014) 187–196.
[10] R.S. D'Souza, K.G. Bhat, D. Sailaja, D.V. Babji, T.K. Bandiwadekar, R.M. Katgalkar, Analysis of expression and localization of TLR-2 by immunofluorescent technique in healthy and inflammed oral tissues, Journal of Clinical and Diagnostic Research: JCDR 7 (12) (2013) 2780-683.
[11] S. Jiang, L. Wang, M. Huang, Z. Jia, T. Weinert, E. Warkentin, C. Liu, X. Song, H. Zhang, J. Witt, L. Qiu, G. Peng, L. Song, DM9 domain containing protein functions as a pattern recognition receptor with broad microbial recognition spectrum, Front. Immunol. 8 (2017) 1607.
[12] D. Kaur, S. Patiyal, N. Sharma, S.S. Usmani, G.P.S. Raghava, PRRDB 2.0: a comprehensive database of pattern-recognition receptors and their ligands, Database: the journal of biological databases and curation 2019 (2019).
[13] D. Kaur, C. Arora, G.P.S. Raghava, A hybrid model for predicting pattern recognition receptors using evolutionary information, Front. Immunol. 11 (2020) 71.
[14] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999, Nucleic Acids Res. 27 (1) (1999) 49–54.
[15] 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.
[16] W. Chen, P. Feng, F. Nie, iATP: a sequence based method for identifying anti-tubercular peptides, Med. Chem. (2019)https://doi.org/10.2174/1573406415666191002152441.
[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] 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.
[19] J. Zhang, B. Liu, A review on the recent developments of sequence-based protein feature extraction methods, Curr. Bioinforma. 14 (3) (2019) 190–199.
[20] 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.
[21] H. Lv, F.Y. Dao, D. Zhang, Z.X. Guan, H. Yang, W. Su, M.L. Liu, H. Ding, W. Chen, H. Lin, iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes, iScience 23 (4) (2020), 100991. .
[22] X.Q. Ru, L.H. Li, Q. Zou, Incorporating distance-based top-n-gram and random forest to identify electron transport proteins, J. Proteome Res. 18 (7) (2019) 2931–2939.
[23] Z. Lv, S. Jin, H. Ding, Q. Zou, A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features, Frontiers in Bioengineering and Biotechnology 7 (2019) 215.
[24] W. Chen, P. Feng, H. Ding, H. Lin, Classifying included and excluded exons in exon skipping event using histone modifications, Front. Genet. 9 (2018) 433.
[25] 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.
[26] J.E. Gewehr, M. Szugat, R. Zimmer, BioWeka–extending the Weka framework for bioinformatics, Bioinformatics 23 (5) (2007) 651–653.
[27] Z.M. Zhang, J.X. Tan, F. Wang, F.Y. Dao, Z.Y. Zhang, H. Lin, Early diagnosis of hepatocellular carcinoma using machine learning method, Frontiers in Bioengineering and Biotechnology 8 (2020) 254.
[28] K. Liu, W. Chen, iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications, Bioinformatics 36 (11) (2020) 3336–3342.
[29] B. Liu, BioSeq-analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches, Brief. Bioinform. 20 (4) (2019) 1280–1294.
[30] H. Tang, Y.W. Zhao, P. Zou, C.M. Zhang, R. Chen, P. Huang, H. Lin, HBPred: a tool to identify growth hormone-binding proteins, Int. J. Biol. Sci. 14 (8) (2018) 957–964.
[31] H. Ding, D. Li, Identification of mitochondrial proteins of malaria parasite using analysis of variance, Amino Acids 47 (2) (2015) 329–333.
[32] P.M. Feng, H. Ding, W. Chen, H. Lin, Naive Bayes classifier with feature selection to identify phage virion proteins, Computational and mathematical methods in medicine 2013 (2013), 530696. .
[33] Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.
[34] Q. Zou, S. Wan, Y. Ju, J. Tang, X. Zeng, Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, BMC Syst. Biol. 10 (4) (2016) 114.
[35] F.Y. Dao, H. Lv, H. Zulfiqar, H. Yang, W. Su, H. Gao, H. Ding, H. Lin, A computational platform to identify origins of replication sites in eukaryotes, Brief. Bioinform. (2020)https://doi.org/10.1093/bib/bbaa017.
[36] 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.
[37] W. Chen, P. Feng, X. Song, H. Lv, H. Lin, iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features, Molecular Therapy. Nucleic Acids 18 (2019) 269–274.
[38] W. Chen, H. Lv, F. Nie, H. Lin, i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome, Bioinformatics 35 (16) (2019) 2796–2800.
[39] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (15) (2004) 2479–2481.
[40] Z.B. Lv, C.Y. Ao, Q. Zou, Protein function prediction: from traditional classifier to deep learning, Proteomics 19 (14) (2019) 2.
[41] B. Wu, H. Zhang, L. Lin, H. Wang, Y. Gao, L. Zhao, Y.-P.P. Chen, R. Chen, L. Gu, A similarity searching system for biological phenotype images using deep convolutional encoder-decoder architecture, Curr. Bioinforma. 14 (7) (2019) 628–639.
[42] L. Wei, Y. Ding, R. Su, J. Tang, Q. Zou, Prediction of human protein subcellular localization using deep learning, Journal of Parallel & Distributed Computing 117 (2018) 212–217.
[43] L. Peng, M.M. Peng, B. Liao, G.H. Huang, W.B. Li, D.F. Xie, The advances and challenges of deep learning application in biological big data processing, Curr. Bioinforma. 13 (4) (2018) 352–359.
[44] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, X. Gao, Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites, Neurocomputing 324 (2019) 3–9.