Prediction of Thermophilic Proteins Using Voting Algorithm

Jing Li[1], Pengfei Zhu[1(✉)], and Quan Zou[2(✉)]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
lijingtju@foxmail.com, zhupengfei@tju.edu.cn
[2] Institute of Fundamental and Frontier Sciences,
University of Electronic Science and Technology of China, Chengdu, China
zouquan@nclab.net

Abstract. Thermophilic proteins have widely used in food, medicine, tanning, and oil drilling. By analyzing the protein sequence, the superior structure and properties of the protein sequence are obtained, which is used to efficiently predict the protein species. In this paper, a voting algorithm was designed independently. Protein features and dimensions were extracted and reduced, respectively. Data was predicted by WEKA. Next, the voting algorithm was applied to the data obtained by the above processing. In this experiment, the highest accuracy rate of 93.03% was achieved. This experiment has at least two advantages: First, the voting algorithm was developed independently. Second, any optimization method was not used for this experiment, which prevents over-fitting. Therefore, voting is a very effective strategy for the thermal stability of proteins. The prediction data set used in this paper can be freely downloaded from http://lab.malab.cn/~lijing/thermo_data.html.

Keywords: Thermophilic proteins · Voting algorithm · Feature selection · Machine learning

1 Introduction

Since the extreme thermophilic microbe genome (the Methanococus jannaschii) has been published, the method of comparing genomes (proteome) has been widely used for the research of protein thermostability.

By mining the charged residues and hydrophobic residues, Bayesian rules, logic functions, neural networks, support vector machines, decision trees are used to distinguish between thermophilic proteins and non-thermophilic proteins. For data of 4684 and 653 protein sequences, 85% and 91% were obtained by neural network and 5-fold cross-validation [11]. By analyzing the distribution of neighbouring amino acids, there are dramatic differences in thermophilic and non-thermophilic proteins. A statistical method was designed for the detection of dipeptide data. 86.3%, 85.5% and 89.7% were displayed, including comparative experiments [30]. Structural information is applied to the logitboost classifiers by

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 195–203, 2019.
https://doi.org/10.1007/978-3-030-17938-0_18

recognition of the first-class protein structure, and the principle of 5-fold cross-validation is set. Experiments show that 97% and 86.6% accuracy are captured separately. It is found that the logitboost classifier has strong generalization capacity and low demanding on the length of the protein sequence [32]. Experimental material is used in a variety of protein identification patterns, which has high degree of confidence. Among these methods, the credibility of the back propagation neural network is up to 98%. The experimental results show that the accuracy of 75% and 85% of thermophilic and non-thermophilic protein, respectively [31]. Potential models and sealed information were mined and found by Chaos game representation (CGR). The pseudo-amino acid information was calculated and extended into protein sequences, which were visualized by the CGR model. Features were extracted via CGR section and 87.92% was captured [17]. Considering the problem of mutations caused by the growth or shortening of protein sequences, this article claims that protein stability can be promoted by Support Vector Machine (SVM). Test results show that the classification accuracy rate reaches 88% [18]. In order to distinguish thermophilic proteins from non-thermophilic proteins and to deal with the stability changes of protein mutations, this paper invented a new type scoring function. Feature weights were taken into account by rewriting the random forest classifier. In the end, 97.3% accuracy was completed [13].

In this paper, a new voting program was developed. By extracting 13 features and integrating 24 classifiers, the better integrated combination was selected for voting, and relatively high accuracy was captured. The extracted features were CKSAAGP, AAC, CKSAAP, CTPC, GAAC, GTPC, GDPC, CTDC, DDE, DPC, CTDT, KSCTRIAD and TPC. Because there are too many classifiers, only voting classifiers will be explained in the following sections. Next, the dimensions of all features are cut, appropriately. WEKA was applied to preliminarily predict, and the results of preliminary prediction were used in the voting program. Ultimately, the accuracy of 93.62% and 92.8% was achieved, separately. The experiment found that data without dimension reduction has better performance.

Compared with published schemes to distinguish between thermophilic and non-thermophilic proteins, the strengths of this study are obvious.

(1) The accuracy is higher.
The result of the vote was 93.03%
(2) The voting program was developed, independent.
Without engineering contribution to support theory, many published papers merely describe a general method for identifying thermophilic and non-thermophilic proteins in the field of bioinformatics. In contrast, this research has corresponding engineering as the theoretical basis. In other words, professional ability of the operator is less demanding. This is crucial for the development bioinformatics [4].
(3) The data has not been optimized to prevent over-fitting.
Sometimes, in order to get better results, optimizer will be applied to the experimental process in the field of data mining. Most of the time, data

optimization does more disadvantages than advantages. Optimization will cause many problems that cannot be ignored and the prediction effect of the model is poor [33].

2 Material and Method

2.1 Data Sources

The data source is http://lab.malab.cn/~lijing/thermo_data.html, including 915 thermophilic proteins and 793 non-thermophilic proteins. The labels of the data are positive and negative.

2.2 Feature Extraction

The features extracted are significant, which will largely affect the experimental results. The theoretical basis of the amino acids features extracted is that location information and structural composition. In the key step, 13 features were extracted, namely CKSAAGP, AAC, CKSAAP, CTPC, GAAC, GTPC, GDPC, CTDC, DDE, DPC, CTDT, KSCTRIAD. Given the limited space, feature extraction algorithms will be overly generalized and will not delve into the details.

The features of AAC algorithm are extracted based on the number of appearance. 20 different amino acids were found, respectively [3]. The DDE algorithm is based on the formation of dipeptides. After a series of reversals, the ideal mean and the ideal variance are calculated, which are used to obtain the final indicator [12]. The design theory of the CKSAAGP algorithm is the frequentness of amino acid, and the homologous eigenvalues are captured by reasoning [7]. The number of protein species is a major consideration in the TPC algorithm [9]. Due to space constraints, only feature descriptors for voting are introduced.

2.3 Max Relevance Max Distance (MRMD)

After feature extracted, the MRMD [42] is used for feature selection. Cutting the less relevant features is the primary task of MRMD [25].

2.4 Classifier Selection and Tools

In the preliminary classification of amino acids, WEKA is the main operating environment for data before and after feature selection, which is fast and efficient [20]. Besides, a large number of classifiers are built into WEKA, and 24 classifiers are screened out. The classifier for voting is discribed in the following content.

LIBSVM is widely used in machine learning and data mining, whose software packages can be used across platforms [22,24]. The goal of Simple Logistic classifier is to achieve the fitting regression effect through Logistic Boost. Through

multiple iterations, the models are updated constantly. When the deviation value of the logistic regression model reduces, the update ends [23]. The random committee classifier is an extension of the random tree classifier, which is mostly used for the formation of low-level classifiers for different data sources [39]. The classification rule of the Logistic classifier is a function, which is derived from the maximum likelihood function, the activation function and the gradient descent algorithm [16]. The principle of PART is the matching of data and "decision lists". When the match reports an error [10], the default category will be called [15].

3 Experiment

In order to confirm the effectiveness of the voting algorithm, other experiments were compared. In Experiment 1, 188D was used for feature extraction of raw data (188D means 188 features were extracted from raw data, which includes 11 extraction principles of amino acid content, hydrophilicity, van der Waals force and polarity, etc.). In Experiment 2, the features were extracted utilizing IFEATURE [5] algorithm, and the WEKA and voting algorithms were used in subsequent experimental procedures. In Experiment 3, MRMD was used to select the extracted features to retain necessary features. WEKA and voting procedures were used to expect better experimental results.

3.1 Performance of Evaluation Standards

$$SN = TP/(TP + FN) \tag{1}$$

$$SP = TN/(TN + FN) \tag{2}$$

$$ACC = (TP + TN)/(TP + TN + FP + FN) \tag{3}$$

3.2 Performance of Experiments

Experiment 1. The raw data includes 915 thermophilic proteins and 793 non-thermophilic proteins. The 188D was used for feature extraction of raw data. After a series of conversions, the data results were processed into the ARFF format, which was run on WEKA (cross-validation was set to 10-fold, and 8 classifiers were selected, namely Bayesian network, Naive Bayes, Decision tree J4.8, Bagging meta learning, Logistic function, Multiclass classifier, Classification via Regression and random forest). Experiment l finds that the multi-class classifier and Logistic function classifier have the highest accuracy. The details are demonstrated in Table 1.

Table 1. The different classifiers performance of 188D.

Methods	AAC
Bays Net	82.50%
Random Forest	88.64%
Decision tree J4.8	81.85%
Bagging meta learning	88.06%
Logistic function	88.93%
Multiclass classifier	88.93%
Classification via Regression	86.71%
Naïve Bayes	83.43%

Data of Experiment 2. Affected by the design principle of IFEATURE, 13 features were extracted from 1708 protein sequences, which are AAC, CKSAAGP, CKSAAP, GTPC, GDPC, CTDC, DDE, DPC, CTDT, KSTRIAD, TPC, GAAC, and CTDD. Besides, many classifiers were tested on WEKA, and only Random Forest results were shown in the Table 2. The highest accuracy rate is 90.57%.

Table 2. The different features accuracy of RF.

Feature	Dimension	ACC
AAC	20	90.57%
CKSAAGP	150	79.22%
CKSAAP	2400	88.23%
CTPC	125	79.04%
GDPC	25	79.63%
CTDC	39	88.06%
CTDT	39	83.49%
DDE	400	88.47%
TPC	8000	84.66%
KSCTRIAD	343	80.91%
CTDD	195	69.67%
GAAC	5	77.22%
DPC	400	88.0%

Data of Experiment 3. The extracted features is selected by MRMD. For comparison, Table 3 shows that the accuracy after dimension reduction with Random Forest classifier on WEKA. For the purposes of comparison, the dimension information is displayed in the Table 3.

Table 3. The different features accuracy of RF after dimension reduction.

Feature	Dimension	ACC
AAC	19	90.93%
CKSAAGP	123	78.98%
CKSAAP	1501	88.23%
CTPC	113	79.04%
GDPC	23	79.63%
CTDC	35	87.7%
CTDT	38	83.49%
DDE	44	85.77%
TPC	25	79.74%
KSCTRIAD	343	80.91%
CTDD	136	68.27%
GAAC	4	76.93%
DPC	398	88.29%

3.3 Data of Voting

Lin's experiment was recurrence. Since the Jackknife took a long time, the experiment switched to 10-fold cross-validation and 92.15% accuracy was achieved. The data of Experiment 2 and Experiment 3 were used for preliminary prediction on WEKA, and a total of 24 classifiers were utilized. In this process, the information of accuracy below 80% is deleted. After all the steps are completed, a matrix of $1702 * 264$ was obtained. For the comparison experiment, the data before and after the feature selection were operated like above.

3.4 Performance of the Algorithm

The voting-based program was developed independently, whose design ideas are as follows:

(1) BASE
After careful consideration, AAC's LIBSVM information is used as a benchmark. The data source is Lin's paper, and it is general accepted to use Lin's results as a voting benchmark.
(2) Based on the information of BASE, the data that is least relevant to BASE is selected.
(3) The algorithm can directly calculate the voting composition, and the accuracy, confusion matrix, F-score and other indicators.
(4) Repeat steps (2) and (3) to achieve higher voting accuracy with fewer data as far as possible.

The data test results of Experiment 1 show that 93.03% is the best result. Not only is higher accuracy achieved, but less information is utilized. The voting combination are LIBSVM (c = 2, g = 2), Random Committee and PART of AAC, LIBSVM (default parameters) and Logistic of DDE, Simple Logistic of TPC and Multi-class classifier of CKSAAGP.

Compared with Experiment 1, the data results of Experiment 2 were relatively poor. After comprehensive consideration, 92.8% was regarded as the best performance. This result integrates information of LIBSVM (c = 2, g = 2) of AAC, LIBSVM (c = 2, g = 2), Naïve Bayes and Logistic of CKSAAP, Multi-class classifier and Simple Logistic of DPC, Logistic of CKSAAGP. It deserves special explanation that the cross-validation of all experiments was set to 10-fold.

4 Conclusion

Amino acid classification is a major problem in bioinformatics. Since the development of bioinformatics, many theories and algorithms based on amino acid classification have been proposed. Due to the limitation of generalization ability, the classification has not reached the ideal accuracy. In this paper, various factors are considered and a voting algorithm is proposed, whose execution result is the integration of LIBSVM (c = 2, g = 2), Random Committee and PART of AAC, LIBSVM (default parameters) and Logistic of DDE, Simple Logistic of TPC and Multi-class classifier of CKSAAGP. The final accuracy rate was 93.03.

As a new interdisciplinary technology in the bioinformatics field, thermophilic proteins play very important role in the study of human health. To systematically present the experimental results and improve ease of use, a server for predicting thermophilic proteins has been developed. The user only needs to input protein sequence, and the highest accuracy of voting and corresponding protein data can be obtained, automatically. On the other hand, Link prediction paradigms [40] have been applied in the prediction of disease genes [27], circular RNAs [29], miRNAs [6,8,21,37], drug side effects [35] and LncRNAs [1,34,36,38]. Also, computational intelligence such as neural networks [2,19], evolutionary algorithms [26,41] and unsupervised learning [14,28] can be applied to predict health related thermophilic proteins.

References

1. Alshahrani, M., Khan, M.A., Maddouri, O., Kinjo, A.R., Queralt-Rosinach, N., Hoehndorf, R.: Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics **33**(17), 2723–2730 (2017)
2. Cabarle, F.G.C., Adorna, H.N., Jiang, M., Zeng, X.: Spiking neural P systems with scheduled synapses. IEEE Trans. Nanobiosci. **16**(8), 792–801 (2017)
3. Chen, W., Ding, H., Zhou, X., Lin, H., Chou, K.-C.: iRNA(m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition. Anal. Biochem. **561**, 59–65 (2018)

4. Chen, W., Yang, H., Feng, P., Ding, H., Lin, H.: iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics **33**(22), 3518–3523 (2017)
5. Chen, Z., et al.: iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics **34**(14), 2499–2502 (2018)
6. Cheng, L., Hu, Y., Sun, J., Zhou, M., Jiang, Q.: DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics **34**(11), 1953–1956 (2018)
7. Cheng, L., et al.: InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. BMC Genom. **19**(1), 919 (2018)
8. Cheng, L., et al.: LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. **47**(D1), D140–D144 (2018)
9. Cheng, L., et al.: MetSigDis: a manually curated resource for the metabolic signatures of diseases. Briefings Bioinform. **20**(1), 203–209 (2017)
10. Feng, C.-Q., et al.: iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics (2018)
11. Michael Gromiha, M., Xavier Suresh, M.: Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins: Struct. Funct. Bioinform. **70**(4), 1274–1279 (2008)
12. Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., Cheng, L.: Identifying diseases-related metabolites using random walk. BMC Bioinform. **19**(5), 116 (2018)
13. Li, Y., Russell Middaugh, C., Fang, J.: A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. BMC Bioinform. **11**(1), 62 (2010)
14. Liao, Z., Li, D., Wang, X., Li, L., Zou, Q.: Cancer diagnosis through isomiR expression with machine learning method. Curr. Bioinform. **13**(1), 57–63 (2018)
15. Liu, B., Yang, F., Chou, K.-C.: 2L-piRNA: a two-layer ensemble classifier for identifying Piwi-interacting RNAs and their function. Mol. Ther.-Nucleic Acids **7**, 267–277 (2017)
16. Liu, B., Yang, F., Huang, D.-S., Chou, K.-C.: iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics **34**(1), 33–40 (2017)
17. Liu, X.-L., Lu, J.-L., Hu, X.-H.: Predicting thermophilic proteins with pseudo amino acid composition: approached from chaos game representation and principal component analysis. Protein Peptide Lett. **18**(12), 1244–1250 (2011)
18. Montanucci, L., Fariselli, P., Martelli, P.L., Casadio, R.: Predicting protein thermostability changes from sequence upon multiple mutations. Bioinformatics **24**(13), i190–i195 (2008)
19. Song, T., Rodríguez-Patón, A., Zheng, P., Zeng, X.: Spiking neural P systems with colored spikes. IEEE Trans. Cogn. Dev. Syst. **10**(4), 1106–1115 (2018)
20. Su, R., Wu, H., Xu, B., Liu, X., Wei, L.: Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. IEEE/ACM Trans. Comput. Biol. Bioinform. (2018)
21. Tang, Y., Liu, D., Wang, Z., Wen, T., Deng, L.: A boosting approach for prediction of protein-RNA binding residues. BMC Bioinform. **18**(13), 465 (2017)
22. Wei, L., Chen, H., Su, R.: M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. Mol. Ther.-Nucleic Acids **12**, 635–644 (2018)
23. Wei, L., Wan, S., Guo, J., Wong, K.K.L.: A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. **83**, 82–90 (2017)

24. Wei, L., Xing, P., Zeng, J., Chen, J.X., Su, R., Guo, F.: Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. **83**, 67–74 (2017)
25. Wei, L., Zhou, C., Chen, H., Song, J., Su, R.: ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics **34**(23), 4007–4016 (2018)
26. Xu, H., Zeng, W., Zeng, X., Yen, G.G.: An evolutionary algorithm based on Minkowski distance for many-objective optimization. IEEE Trans. Cybern. (99), 1–12 (2018)
27. Zeng, X., Ding, N., Rodríguez-Patón, A., Zou, Q.: Probability-based collaborative filtering model for predicting gene-disease associations. BMC Med. Genom. **10**(5), 76 (2017)
28. Zeng, X., Liao, Y., Liu, Y., Zou, Q.: Prediction and validation of disease genes using hetesim scores. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **14**(3), 687–695 (2017)
29. Zeng, X., Lin, W., Guo, M., Zou, Q.: A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput. Biol. **13**(6), e1005420 (2017)
30. Zhang, G., Fang, B.: Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process Biochem. **41**(8), 1792–1798 (2006)
31. Zhang, G., Fang, B.: Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. Process Biochem. **41**(3), 552–556 (2006)
32. Zhang, G., Fang, B.: Logitboost classifier for discriminating thermophilic and mesophilic proteins. J. Biotechnol. **127**(3), 417–424 (2007)
33. Zhang, J., Feng, P., Lin, H., Chen, W.: Identifying RNA N6-methyladenosine sites in escherichia coli genome. Front. Microbiol. **9**, 955 (2018)
34. Zhang, J., Zhang, Z., Chen, Z., Deng, L.: Integrating multiple heterogeneous networks for novel LncRNA-disease association inference. IEEE/ACM Trans. Comput. Biol. Bioinform. (2017)
35. Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., Li, X.: Feature-derived graph regularized matrix factorization for predicting drug side effects. Neurocomputing **287**, 154–162 (2018)
36. Zhang, W., Qu, Q., Zhang, Y., Wang, W.: The linear neighborhood propagation method for predicting long non-coding RNA-protein interactions. Neurocomputing **273**, 526–534 (2018)
37. Zhang, X., Zou, Q., Rodriguez-Paton, A., et al.: Meta-path methods for prioritizing candidate disease miRNAs. IEEE/ACM Trans. Comput. Biol. Bioinform (2017)
38. Zhang, Z., Zhang, J., Fan, C., Tang, Y., Deng, L.: KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. IEEE/ACM Trans. Comput. Biol. Bioinform (2017)
39. Zhu, X.-J., Feng, C.-Q., Lai, H.-Y., Chen, W., Hao, L.: Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl.-Based Syst. **163**, 787–793 (2019)
40. Zou, Q., Li, J., Song, L., Zeng, X., Wang, G.: Similarity computation strategies in the microrna-disease network: a survey. Briefings Func. Genom. **15**(1), 55–64 (2015)
41. Zou, Q., Wan, S., Zeng, X., Ma, Z.S.: Reconstructing evolutionary trees in parallel for massive sequences. BMC Syst. Biol. **11**(6), 100 (2017)
42. Zou, Q., Zeng, J., Cao, L., Ji, R.: A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing **173**, 346–354 (2016)