International Journal of
Molecular Sciences

MDPI

Article

RFAmyloid: A Web Server for Predicting Amyloid Proteins

Mengting Niu [1], Yanjuan Li [1,*], Chunyu Wang [2] and Ke Han [3]

[1] School of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China; yunzeer@gmail.com
[2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150040, China; chunyu@hit.edu.cn
[3] School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150040, China; hanke@hrbcu.edu.cn
* Correspondence: liyanjuan@nefu.edu.cn; Tel.: +86-136-5455-0662

Received: 12 June 2018; Accepted: 12 July 2018; Published: 16 July 2018

Abstract: Amyloid is an insoluble fibrous protein and its mis-aggregation can lead to some diseases, such as Alzheimer's disease and Creutzfeldt–Jakob's disease. Therefore, the identification of amyloid is essential for the discovery and understanding of disease. We established a novel predictor called RFAmy based on random forest to identify amyloid, and it employed SVMProt 188-D feature extraction method based on protein composition and physicochemical properties and pse-in-one feature extraction method based on amino acid composition, autocorrelation pseudo acid composition, profile-based features and predicted structures features. In the ten-fold cross-validation test, RFAmy's overall accuracy was 89.19% and F-measure was 0.891. Results were obtained by comparison experiments with other feature, classifiers, and existing methods. This shows the effectiveness of RFAmy in predicting amyloid protein. The RFAmy proposed in this paper can be accessed through the URL http://server.malab.cn/RFAmyloid/.

Keywords: amyloid protein; random forest; RFAmy; protein classification; machine learning

1. Introduction

The name of the amyloid protein comes from the technique of the early immature iodine staining [1]. In many neurological diseases such as Alzheimer disease and Parkinson's disease, large amounts of amyloid accumulation in the nervous system can be observed [2]. Many scholars believe that it may lead to degeneration or dysfunction of the brain or other organs [3,4]. At the time, the scientific community debated whether it was a matter of fat deposition or carbohydrate precipitation until finally it was discovered that it was a protein substance [5]. The exact mechanism of amyloid formation is not fully understood, but the precondition for the deposition of amyloid fibrils is the excessive production of its precursor protein [6]. The prevention of this disease should be based on active treatment of the primary disease that can induce the disease [7]. Researchers have demonstrated that the immune system has similar efficacy in humans. Therefore, to understand amyloid proteins and related diseases deeply, the most research on amyloid proteins focuses on amyloidosis [8–10], amyloid region [11,12], aggregation [3,13,14], and antibody amyloid [15].

Many calculation methods on the problem of amyloid accumulation have been developed, such as AmylPred [16], Pafig [17], FoldAmyloid [12], and Waltz [18]. AmylPred method mainly uses five different and independently published methods to form a consensus prediction of amyloidogenic region. Pafig employs the support vector machine to predict the amyloid protein region through the recognition of the hexapeptides associated with the aggregation of amyloid protein. FoldAmyloid

realizes the prediction of the amyloid region by combining the method of predicting hydrogen bond formation with the expected bulk density of the residues. Waltz [19] employs the position-specific scoring matrix to predict the amyloid region. As reviewed [20], there are currently two major methods to study the aggregation of amyloid proteins and to identify the amyloidogenic regions that are most likely to form fibrils: (1) using a phenomenological model based on the physicochemical properties of amino acids to identify the amyloidogenic regions; and (2) modeling the microcrystalline structure of the peptides by simulating the short fibers of the amyloid fragment [21,22].

For the amyloid protein region, many studies and prediction methods exist. Garbuzynskiy et al. proposed and developed an online web server named FoldAmyloid [12]. It mainly uses the statistical data features of the amyloid protein and introduces two features, namely expected probability of hydrogen bonds formation and expected packing density of residues, to predicted amyloidgenic regions [12]. Wieczorek et al. proposed amyloid protein region prediction based on fuzzy grammar [23]. In their paper, the amyloid sequence is described by fuzzy context-free grammar and the amyloidogenic region is identified by fuzzy grammar. To accurately predict the amyloidogenic region, Emily et al. combined weighted merging of existing popular methods to create a meta-predictor called MetAmyloid [24]. There are also many methods for successfully predicting the amyloid region in amino acid sequences with computational techniques, such as biological mutagenesis and quantitative calculations. For the formation of antibody amyloid, Otoo et al. proposed the automatic and cross-species prediction method AbAmyloid [25], which employs random forest algorithm. The prediction has been tested on 12 datasets, and outperformed other methods. David et al. combined Naive Bayes and decision trees to predict amyloidogenesis in antibodies [15]. Although the mis-aggregation of amyloid may lead to some clinical studies, many studies have recently shown that amyloid still has positive significance in some aspects, for example bacterial and antimicrobial activity [4,26], fungal biofilm formation [21,27–29], storage of peptide hormones [3,30], the formation of zona pellucida to protect mammalian and fish oocytes [31], etc. These studies show the importance of increasing the awareness of amyloid.

Although there is a lot of research on amyloid protein, they ignored the first step of identifying amyloid protein. In this paper, we present RFAmy to identify amyloid with random forest (RF), which it based on composition and physicochemical features from protein primary sequences.

As we conclude above, machine learning frame has been employed to identify special proteins, including cytokines [32], DNA-binding proteins [33,34], RNA-binding proteins [35], lncRNA-interacting proteins [36], drug targets [37], etc. Different sequence features have been proposed to describe proteins, including Chou's Pse features, SVMProt [38], secondary structure features [39], evolutionary features [40], etc. Different machine learning classifiers have been employed, including support vector machine [41,42], random forests [43], ANN [44], etc. There are also some special classifiers for different conditions, such as ensemble classifier [42,45–54], multi-label classifier [55–58], imbalance classifier [59,60], hierarchical classifier [61–63], etc. All these previous works guide us to build a frame for amyloid protein identification.

2. Results and Discussion

2.1. Measurement

To evaluate the prediction effect of the prediction algorithm used, this article selects four commonly used indicators: sensitivity (SE) (Equation (1)), specificity (SP) (Equation (2)), accuracy (ACC) (Equation (3)) and Matthew coefficient (MCC) (Equation (4)).

$$\mathrm{SE} = \frac{TP}{TP + FN} \tag{1}$$

$$\mathrm{SP} = \frac{TN}{TN + FP} \tag{2}$$

$$\text{ACC} = \frac{TN + TP}{TN + FP + TP + FN} \tag{3}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{4}$$

where TP indicates the number of amyloid proteins predicted in the sequence of positive cases, FP indicates the number of non-amyloid proteins predicted in the counterexample sequence, TN indicates the number of non-amyloid proteins predicted in the sequence of positive cases, and FN indicates the number of non-amyloid proteins predicted in the counterexample sequence. SE denotes the ratio of being positive in the sequence and predicting positive. SP indicates the correct rate of prediction of counterexamples.

ACC denotes the proportion of correct predictions in all the positive and negative examples, and the reliability of the MCC represents the results of the algorithm. When the difference between the positive and negative examples is large, the prediction ability can be more equitably reflected.

In this paper, the positive and negative dataset are unbalanced, so we have additionally adopted a criterion F-measure (Equation (7)) which is calculated with precision (Equation (5)) and recall (Equation (6)).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F} - \text{measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

2.2. Performance of Different Features on Cross-Validation

This section presents the selection of *n*-gram features, adaptive skip-gram features (400-D), pse-in-one features and 188-D feature to verify the validity of the 188-D and pse-in-one combined feature representation method used in this paper. The results are shown in Table 1. Here, the features representation methods that we used to compare with our feature representation method are briefly listed.

The *n*-gram features are common in natural language processing and we employed this feature in protein prediction problems [64]. This is a method of checking "*n*" consecutive words or sounds from a given text or speech sequence. The *n*-gram needs to link n words together as a feature. The *n*-gram assumes that the *n*th word is only affected by the first $n - 1$ words. The probability of the entire sentence is the product of the probability of occurrence of each word. This model helps to predict the next item in the sequence.

The Adaptive Skip-Gram Features model (400-D) is a variant of the *n*-gram model. The corpora counted by jumping a certain number or position of words were used to obtain n-gram information, Adaptive Skip-Gram features more content than n-gram. The correlation between distance and sequence amino acids to a certain extent solves the problem of feature space sparsity caused by the traditional n-gram method.

The results are shown in Table 1. In Table 1, we can see that the feature representation method used in this paper performs well on all indicators compared to other methods. The accuracy, MCC, SE, SP and F-measure all reached maximum: 89.19%, 0.739, 0.781, 0.927 and 0.891, respectively. In short, the random forest based RFAmy predictor feature extraction algorithm outperforms the others. Therefore, the feature extraction method used in this paper is feasible and effective.

Table 1. The result of using different feature representation methods on cross-validation.

Method	ACC (%)	MCC	SE	SP	F-Measure
188-D+Pse-in-One	89.1941	0.739	0.781	0.927	0.891
188-D	84.8482	0.626	0.655	0.932	0.626
Pse-in-one	81.31	0.5626	0.6374	0.8989	0.792
400-D	84.1105	0.634	0.691	0.917	0.838
n-gram (n = 1)	81.3187	0.522	0.534	0.930	0.802

2.3. Performance of Different Features on External Validation

To test the robustness of the proposed method, external validation is required to evaluate the developed predictive model. Therefore, we evaluated RFAmyloid on an independent dataset and again compared its performance to the performance of different feature representation methods. We only used 80% of the data to develop the predictive model, and the remaining 20% was used for external or independent verification. The independent test results are shown in Table 2. In Table 2, we can see that the feature representation method used in this paper performs well on all indicators compared to other methods. The accuracy, MCC, SE, SP and F-measure all reached maximum: 89.71%, 0.757, 0.818, 0.932 and 0.897, respectively. Independent testing confirmed the previous test results and confirmed that our proposed predictor effectively recognizes amyloid. Since the proposed method is robust in independent testing, it should be effective in predicting new amyloids.

Table 2. The result of using different feature representation methods on external validation.

Method	ACC (%)	MCC	SE	SP	F-Measure
188-D+Pse-in-One	89.7196	0.757	0.818	0.932	0.897
188-D	73.1841	0.524	0.512	0.960	0.678
Pse-in-one	78.7037	0.679	0.676	0.880	0.782
400-D	71.2963	0.543	0.503	0.893	0.684
n-gram (n = 1)	69.4444	0.522	0.534	0.893	0.657

2.4. Comparison with Other Classifiers

In this subsection, the performance of RFAmy is compared with the performances of other classifiers, namely Naive Bayes, SGD, Nearest Neighbors, Decision Tree, LinearSVC, Logistic Regression, LibSVM, ExtraTrees, Bagging, AdaBoost, GradientBoosting, and LibD3C [65]. The experimental results are shown in Table 3. In Table 3, although the RFAmy method presented in this paper is lower than Nearest Neighbors in SP index, RFAmy is obviously superior to the four other indices. The RFAmyloid has the highest accuracy and F-measure: 89.19% and 0.891, respectively. Figure 1 shows the ROC curve (the further the curve is projected to the left, the better the effect is) of RFAmy and the comparison classifiers' experimental results. It is well verified that the random forest classifier outperforms other classifiers in predicting the accuracy of amyloid, demonstrating the validity of the proposed method.

Table 3. The result of using different classifiers based on 188-D feature.

Classifier	ACC (%)	MCC	SE	SP	F-Measure
Random Forest	89.19	0.739	0.781	0.927	0.891
Naive Bayes	75.50	0.3791	0.4606	0.8822	0.8721
SGD	77.51	0.4451	0.5515	0.8717	0.6533
Nearest Neighbors	77.70	0.4293	0.2970	0.9843	0.8818
Decision Tree	67.28	0.2567	0.5333	0.7330	0.7461
LinearSVC	77.51	0.4654	0.6242	0.8403	0.8658
Logistic Regression	79.52	0.5123	0.6545	0.8560	0.8694
LibSVM	70.02	0.0651	0.0061	1.0000	0.8239
ExtraTrees	74.95	0.4128	0.6061	0.8115	0.8087
Bagging	74.95	0.4128	0.6061	0.8115	0.7727
AdaBoost	76.78	0.4700	0.6788	0.8063	0.8763
GradientBoosting	80.26	0.5298	0.6667	0.8613	0.8668
LibD3C	86.99	0.683	0.732	0.929	0.868

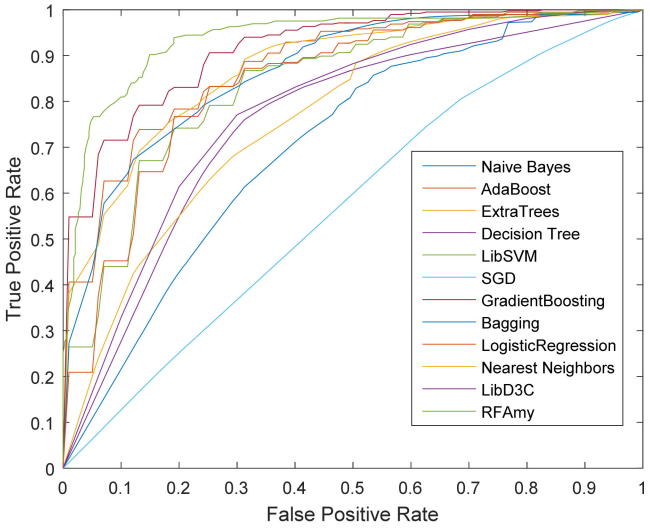

Figure 1. Receiver Operating Characteristic (ROC) curve for RFAmy and other methods.

2.5. Comparison with Other Predictors

In this section, the proposed prediction method is compared with the existing prediction method BioSeq-Analysis. The online address for this method is http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/PROTEIN/Kmer/ [66]. The SVM and random forest algorithm are used in The BioSeq-Analysis prediction method. This section compars them separately. The prediction results are shown in Table 4. Figure 2 shows the roc curve of RFAmy and comparison algorithm method experimental results. Table 4 shows that the RFAmy method proposed in this paper achieved the best results on the all evaluation indicators. In addition, the ROC curve diagram (the further the curve is projected to the left, the better the effect is) shows that the RFAmy method in this paper is obviously better than the other two methods.

Table 4. The result of using different methods.

Method	ACC (%)	MCC	SE	SP
RFAmy	89.1941	0.739	0.781	0.927
BioSeq-SVM	76.86	0.4419	0.4953	0.9006
BioSeq-RF	81.31	0.5626	0.6374	0.8989

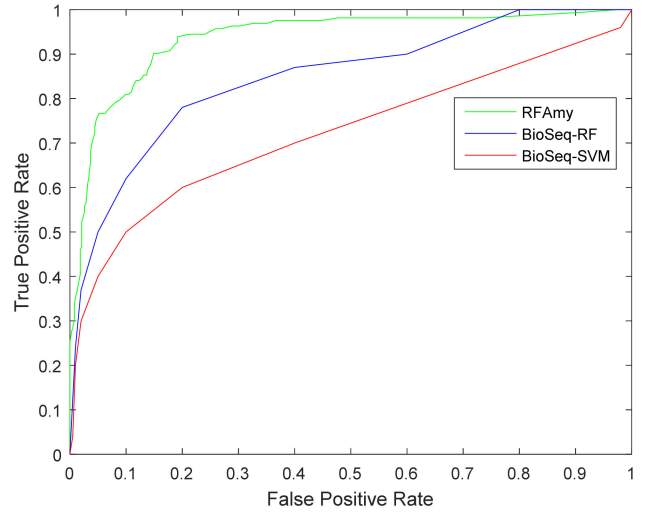

Figure 2. Receiver Operating Characteristic (ROC) curve for RFAmy and two other methods.

2.6. Comparison with Balanced Dataset

The results in Tables 1–4 show that the specificity was much higher than the sensitivity, which is the effect of the unbalanced dataset in the development of predictive models. Therefore, we used a balanced dataset to develop a predictive model and compared its performance to the performance based on an unbalanced dataset. The results are shown in Table 5. From the comparison results in Table 5, we can see that, although the accuracy under the balanced dataset and the F-measure index are slightly lower than the unbalanced dataset, the sensitivity under the balanced dataset is much higher than the specificity. This also proves the importance of the selection of datasets for model prediction.

Table 5. The result of using different feature representation methods.

Method	ACC (%)	MCC	SE	SP	F-Measure
unbalanced	89.1941	0.739	0.781	0.927	0.891
balanced	83.4962	0.757	0.847	0.823	0.865

3. Methods

Figure 3 shows the paper framework for an Amyloid classifier. We introduce the datasets, features and classifiers in detail in this section.

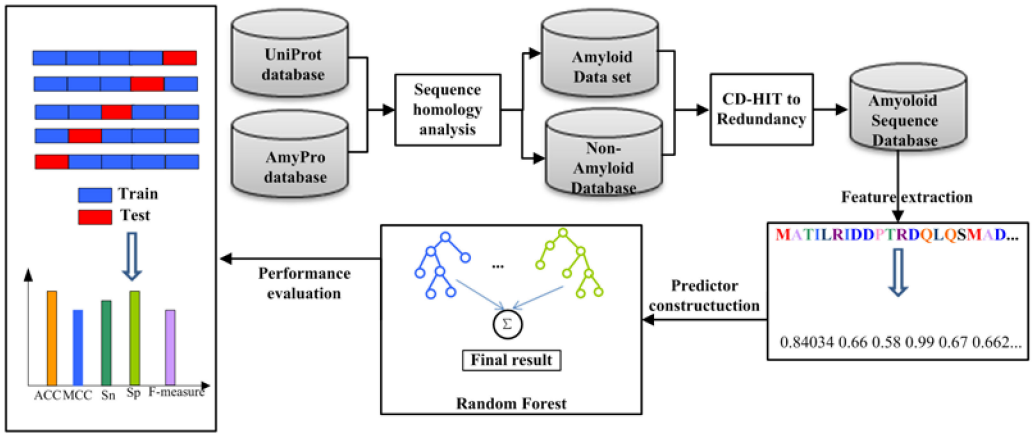

Figure 3. Overview of the paper framework for an Amyloid classifier. First, the original protein sequence was generated from the Uniprot and AmyPro datasets and then subjected to a de-redundant operation to generate the final protein sequence data called Amy. The second step is feature extraction of protein sequences. The third step is to use RF to classify protein sequences.

3.1. Dataset

This study used a self-built dataset named Amy. The dataset constructin followed the common steps of protein prediction.

The first step was to search for proteins. The source databases are the Universal Protein (UniProt, http://www.uniprot.org/) and Amyloid (AmyPro, http://www.amypro.net/) [67] database. The second was to remove the sequences which are less than 50 amino acids. In the third step, protein sequences eliminated redundancy. We used the program CD-HIT to cluster proteins that meet a similarity threshold [68] and eliminate redundancy and homology biases that could lead to overestimation of performance. In this study, through these three steps, a set of amyloid data, the Amy dataset, was formed which consists of 165 amyloid proteins and 382 non-amyloid. The Amy dataset can be downloaded from the server.

3.2. Feature Extraction

Feature extraction is the first and most important component in predictors [69]. We employed a multi-feature representation method that includes two feature representation methods, namely, 188-D feature extraction method based on protein composition and physicochemical properties and pse-in-one feature extraction method based on amino acid composition, autocorrelation pseudo acid composition, profile-based features and predicted structures features.

Different kinds of amino acids have their own special physicochemical properties, which can predict the type of protein as a feature of the amino acid sequence. In addition, the 20 compositional features of amino acids can describe the characteristics of the protein. Both methods achieved good predictive results. Dubchak first attempted to fuse the two features together and achieved better results in predicting protein folding patterns [9]. Afterwards, many scholars proposed a variety of feature fusion methods [70]. The 188-D combined feature extraction method extracts eight physical and chemical characteristics, the frequency of occurrence of 20 amino acids in the protein sequence, the frequency of bipartite subsequences, and the distribution of amino acids with different physical properties in the sequence. The 188-D feature is mainly obtained by the following four steps.

The first step is to extract the proportional characteristics of the amino acid components in the sequence, a total of 20 dimensions. In the second step, using hydrophilicity and hydrophobicity as an example, the compositional content of amino acids with hydrophilicity, hydrophobicity, and neutrality can be calculated to extract 3D features. In the third step, if there is a total of n hydrophilic, hydrophobic, and neutral amino acids in the sequence, calculate the proportions of the first, 25% * n, 50% * n, 75% * n, and the last such amino acids in the sequence of the protein in which they are located. Each category has five dimensions, so there are 15 dimensional features in all three categories. Finally, in accordance with the "hydrophilic", "hydrophobic" and "neutral" properties, two or two combinations are constructed. The 3-D characteristics of "hydrophilic, hydrophobic," "hydrophilic, neutral," and "hydrophobic, neutral" are calculated, and the ratio is calculated in the sequence of the bisimplex, which is also 3D. Table 6 describes the 188-D function.

Table 6. Structure of 188-D Feature.

Physical-Chemical Property	Dimensions
Amino acid composition	20
Hydrophobicity	21
Normalized van der Waals volume	21
Polarity	21
Polarizability	21
Charge	21
Surface tension	21
Secondary structure	21
Solvent accessibility	21

Pse-in-one has five groups of 22 features extraction methods [71]. The first group uses kmers, distance-based residue (DR), and distance pair to indicate the composition of amino acids. The second group uses auto covariance (AC), cross covariance (CC), auto-cross covariance (ACC), and physicochemical distance transformation (PDT) to represent autocorrelation features. The third group uses four indicators such as PC-PseAAC to indicate the characteristics of false amino acids; the indicators can be found in the literature [72]. The fourth group uses Top-n-gram, distance-based auto covariance, profile-based Auto-cross covariance, sequence conservation score, and other three indictors to represent Profile-based features. The fifth group uses secondary structure and solvent accessible surface area to represent predicted structure features. Table 7 shows the 22 feature extraction methods.

Table 7. Pse-in-one feature extraction method of protein sequences.

Category	Method
Amino acid composition	K-mer DR Distance Pair
Autocorrelation	AC CC ACC PDT
Pseudo amino acid composotion	PC-PseAAC SC-PseAAC PC-PseAAC-General SC-PseAAC-General
Profile-based features	Top-*n*-gram PDT-Profile DT AC-PSSM CC-PSSM ACC-PSSM PSSM-DT PSSM-RT CS
Predicted structure features	SS SASA

3.3. Classifier

For the identification of amyloid protein, random forest was selected as the classification algorithm in this study. It is popular and has been successfully used in biometrics many times [55,72–76]. Random forests are a combination of tree predictors. Algorithms are implemented by building multiple decision trees and using voting mechanisms to improve decision trees. Random forests are generated in the following four steps.

The first step is to generate n samples from the sample set by resampling. The second step is to assume that the number of sample features is q, and select k features from q for n samples, and then obtain the best segmentation point by building a decision tree. The third step is to repeat m times, and then generate m decision trees. The fourth step is to predict by a majority voting mechanism. It should be noted that, where m represents the number of cycles, and n represents the number of samples, then n samples constitute the sample set for training, and m such samples are generated in m cycles.

In machine learning, the algorithm model needs to be trained to update each parameter in the model. Therefore, it is necessary to provide the training set as a training sample. At the same time, to describe the generalization ability of the model, a test set is needed to test and obtain the generalization error. In practical applications, cross-checking is often used as a test method because of the limited number of datasets. There are three types of cross-validation: n-fold cross-validation, folding cross-validation and independent data testing [77–80]. In three tests, the folding knife test has been widely used in bioinformatics because it produces unique results [81–85]. However, it takes time and resources. Therefore, in this paper, we use K-fold cross-validation to examine the proposed model, where $K = 10$ is the most common. In detail, the training set is divided into K parts, and then the ith is taken as the test set, the other $K - 1$ is trained as the training set. The operation diagram of the ten-fold cross-validation is shown in Figure 4.

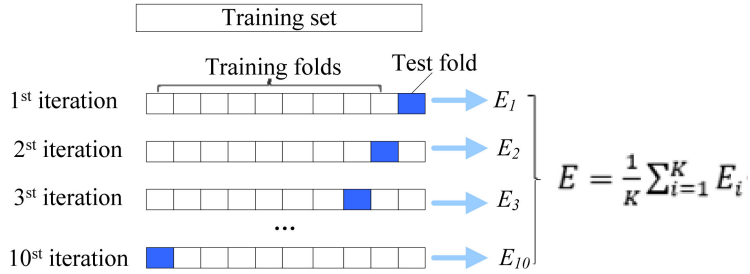

Figure 4. Ten-fold cross validation diagram. The dataset was divided into ten parts, and nine of them were taken as training data in turn, and one was used as test data for testing. The average value E of the ten-groups test results is calculated as an estimate of the model accuracy and is used as a performance indicator for the current K-fold cross-validation model. Where E_i represents the cross-validation error of the ith group.3.4. The RFAmyloid Online Prediction Server.

With the development of bioinformatics, it is important to make better use of machine learning methods to solve related biological information. As mentioned in a series of documents, the development of predictive methods and related servers is very practical and urgently needed for researchers. Therefore, based on the prediction method of the paper and the data used, we also carried out server development. The URL is: http://server.malab.cn/RFAmyloid/.

On the server, the user can paste the protein sequence or upload the file in fasta format. After submitting the protein sequence, the page will give the probability information of whether it is Amyloid protein, and query the prediction result. The dataset used in this paper can be downloaded from the web.

4. Conclusions

In this paper, we propose a new learning algorithm RFAmy for amyloid prediction. We used SVMProt 188-D feature representation, pse-in-one feature representation and random forest classifier. To verify the effect of the proposed predictor, we compared its performance with 10 other cross-validation and independent test sets with other feature representations. In the 10-fold cross-validation, we obtained ACC 89.19% and F-measure 0.891. In the independent test set, we obtained ACC 89.19% and F-measure 0.891. In addition, our models have better predictive effects than other feature extraction algorithms, classifiers and existing methods. The RFAmy proposed in this paper can be accessed through the URL http://server.malab.cn/RFAmyloid/. In future work, we will optimize RFAmy's prediction performance through the improvement of feature extraction algorithms and classification algorithms. For the improvement of the classifier, the use of an integrated classifier will be considered, combining multiple classifiers to complete the classification task and improve the classification accuracy.

Author Contributions: Conceptualization, M.N.; Data curation, C.W.; Formal analysis, C.W. and K.H.; Project administration, Y.L.; Software, M.N. and K.H.; Writing—original draft, M.N.; and Writing—review and editing, Y.L.

Funding: The work was supported by the Fundamental Research Funds for the Central Universities (No.2572017CB33), the Natural Science Foundation of China (No.91735306, 61402138, 61300098).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Beerten, J.; Van Durme, J.; Gallardo, R.; Capriotti, E.; Serpell, L.; Rousseau, F.; Schymkowitz, J. WALTZ-DB: A benchmark database of amyloidogenic hexapeptides. *Bioinformatics* **2015**, *31*, 1698–1700. [CrossRef] [PubMed]

2. Ikeda, S.-I. Localized amyloidogenic immunoglobulin light chain-derived amyloidosis in a young boy and an adolescent girl. *Amyloid* **2017**, *24*, 138–140. [CrossRef] [PubMed]
3. Louros, N.N.; Iconomidou, V.A.; Giannelou, P.; Hamodrakas, S.J. Structural analysis of peptide-analogues of human zona pellucida ZP1 protein with amyloidogenic properties: Insights into mammalian zona pellucida formation. *PLoS ONE* **2013**, *8*, e73258. [CrossRef] [PubMed]
4. Gour, S.; Kaushik, V.; Kumar, V.; Bhat, P.; Yadav Subhash, C.; Yadav Jay, K. Antimicrobial peptide (Cn-AMP2) from liquid endosperm of cocos nucifera forms amyloid-like fibrillar structure. *J. Pept. Sci.* **2016**, *22*, 201–207. [CrossRef] [PubMed]
5. Rochet, J.-C.; Lansbury, P.T. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Bio.* **2000**, *10*, 60–68. [CrossRef]
6. Kallberg, Y.; Gustafsson, M.; Persson, B.; Thyberg, J.; Johansson, J. Prediction of amyloid fibril-forming proteins. *J. Biol. Chem.* **2001**, *276*, 12945–12950. [CrossRef] [PubMed]
7. Dobson, C.M. The structural basis of protein folding and its links with human disease. *Philos. Trans. R. Soc. Lond. B* **2001**, *356*, 133–145. [CrossRef] [PubMed]
8. Sipe, J.D.; Benson, M.D.; Buxbaum, J.N.; Ikeda, S.-I.; Merlini, G.; Saraiva, M.J.M.; Westermark, P. Amyloid fibril proteins and amyloidosis: Chemical identification and clinical classification international society of amyloidosis 2016 nomenclature guidelines. *Amyloid* **2016**, *23*, 209–213. [CrossRef] [PubMed]
9. Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.-H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 8700–8704. [CrossRef] [PubMed]
10. Ahmed, A.B.; Znassi, N.; Château, M.-T.; Kajava, A.V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimers Dement.* **2015**, *11*, 681–690. [CrossRef] [PubMed]
11. De Groot, N.S.; Pallarés, I.; Avilés, F.X.; Vendrell, J.; Ventura, S. Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct. Biol.* **2005**, *5*, 18. [CrossRef] [PubMed]
12. Garbuzynskiy, S.O.; Lobanov, M.Y.; Galzitskaya, O.V. Foldamyloid: A method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* **2010**, *26*, 326–332. [CrossRef] [PubMed]
13. Paladin, L.; Piovesan, D.; Tosatto, S.C.E. Soda: Prediction of protein solubility from disorder and aggregation propensity. *Nucleic Acids Res.* **2017**, *45*, W236–W240. [CrossRef] [PubMed]
14. Makin, O.S.; Atkins, E.; Sikorski, P.; Johansson, J.; Serpell, L.C. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 315–320. [CrossRef] [PubMed]
15. David, M.P.C.; Concepcion, G.P.; Padlan, E.A. Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. *BMC Bioinform.* **2010**, *11*, 79. [CrossRef] [PubMed]
16. Frousios, K.K.; Iconomidou, V.A.; Karletidi, C.-M.; Hamodrakas, S.J. Amyloidogenic determinants are usually not buried. *BMC Struct. Biol.* **2009**, *9*, 44. [CrossRef] [PubMed]
17. Tian, J.; Wu, N.; Guo, J.; Fan, Y. Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinform.* **2009**, *10*, S45. [CrossRef] [PubMed]
18. López de la Paz, M.; Serrano, L. Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 87–92. [CrossRef] [PubMed]
19. Maurer-Stroh, S.; Debulpaep, M.; Kuemmerer, N.; de la Paz, M.L.; Martins, I.C.; Reumers, J.; Morris, K.L.; Copland, A.; Serpell, L.; Serrano, L.; et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods* **2010**, *7*, 237. [CrossRef] [PubMed]
20. Caflisch, A. Computational models for the prediction of polypeptide aggregation propensity. *Curr. Opin. Chem. Biol.* **2006**, *10*, 437–444. [CrossRef] [PubMed]
21. Thompson, M.J.; Sievers, S.A.; Karanicolas, J.; Ivanova, M.I.; Baker, D.; Eisenberg, D. The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 4074–4078. [CrossRef] [PubMed]
22. Yoon, S.; Welsh, W.J. Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci.* **2009**, *13*, 2149–2160. [CrossRef] [PubMed]
23. Wieczorek, W.; Unold, O. Use of a novel grammatical inference approach in classification of amyloidogenic hexapeptides. *Comput. Math. Methods Med.* **2016**, *2016*, 1782732. [CrossRef] [PubMed]
24. Emily, M.; Talvas, A.; Delamarche, C. Metamyl: A meta-predictor for amyloid proteins. *PLoS ONE* **2013**, *8*, e79722. [CrossRef] [PubMed]
25. Otoo, H.N.; Lee, K.G.; Qiu, W.; Lipke, P.N. Candida albicans als adhesins have conserved amyloid-forming sequences. *Eukaryot. Cell* **2008**, *7*, 776–782. [CrossRef] [PubMed]

26. Liaw, C.; Tung, C.-W.; Ho, S.-Y. Prediction and analysis of antibody amyloidogenesis from sequences. *PLoS ONE* **2013**, *8*, e53235. [CrossRef] [PubMed]
27. Lembre, P.; Vendrely, C.; Di Martino, P. Identification of an amyloidogenic peptide from the bap protein of staphylococcus epidermidis. *Protein Pept. Lett.* **2014**, *21*, 75–79. [CrossRef] [PubMed]
28. Tartaglia, G.G.; Cavalli, A.; Pellarin, R.; Caflisch, A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* **2009**, *14*, 2723–2734. [CrossRef] [PubMed]
29. Trovato, A.; Seno, F.; Tosatto, S.C.E. The pasta server for protein aggregation prediction. *Protein Eng. Des. Sel.* **2007**, *20*, 521–523. [CrossRef] [PubMed]
30. Sipe, J.D.; Benson, M.D.; Buxbaum, J.N.; Ikeda, S.-I.; Merlini, G.; Saraiva, M.J.M.; Westermark, P. Nomenclature 2014: Amyloid fibril proteins and clinical classification of the amyloidosis. *Amyloid* **2014**, *21*, 221–224. [CrossRef] [PubMed]
31. Louros, N.N.; Petronikolou, N.; Karamanos, T.; Cordopatis, P.; Iconomidou, V.A.; Hamodrakas, S.J. Structural studies of "aggregation-prone" peptide-analogues of teleostean egg chorion zpb proteins. *Pept. Sci.* **2014**, *102*, 427–436. [CrossRef] [PubMed]
32. Zeng, X.; Yuan, S.; Huang, X.; Zou, Q. Identification of cytokine via an improved genetic algorithm. *Front. Comput. Sci.* **2015**, *9*, 643–651. [CrossRef]
33. Qu, K.; Han, K.; Wu, S.; Wang, G.; Wei, L. Identification of DNA-binding proteins using mixed feature representation methods. *Molecules* **2017**, *22*, 1602. [CrossRef] [PubMed]
34. Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: Predicting tata binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* **2016**, *10*, 114. [CrossRef] [PubMed]
35. Xiao, Y.; Zhang, J.; Deng, L. Prediction of lncRNA-protein interactions using hetesim scores based on heterogeneous networks. *Sci. Rep.* **2017**, *7*, 3664. [CrossRef] [PubMed]
36. Zhang, W.; Qu, Q.; Zhang, Y.; Wang, W. The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* **2018**, *273*, 526–534. [CrossRef]
37. Zhang, W.; Chen, Y.; Li, D. Drug-target interaction prediction through label propagation with linear neighborhood information. *Molecules* **2017**, *22*, 2056. [CrossRef] [PubMed]
38. Cai, C.Z.; Han, L.Y.; Ji, Z.L.; Chen, X.; Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **2003**, *31*, 3692–3697. [CrossRef] [PubMed]
39. Wei, L.; Liao, M.; Gao, X.; Zou, Q. An improved protein structural classes prediction method by incorporating both sequence and structure information. *IEEE Trans. Nanobiosci.* **2015**, *14*, 339–349. [CrossRef] [PubMed]
40. Gao, J.; Zhang, N.; Ruan, J. Prediction of protein modification sites of gamma-carboxylation using position specific scoring matrices based evolutionary information. *Comput. Biol. Chem.* **2013**, *47*, 215–220. [CrossRef] [PubMed]
41. Zhang, W.; Yue, X.; Huang, F.; Liu, R.; Chen, Y.; Ruan, C. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods* **2018**. [CrossRef] [PubMed]
42. Zhang, W.; Chen, Y.; Liu, F.; Luo, F.; Tian, G.; Li, X. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinform.* **2017**, *18*, 18. [CrossRef] [PubMed]
43. Chen, L.; Chu, C.; Huang, T.; Kong, X.; Cai, Y.D. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino Acids* **2015**, *47*, 1485–1493. [CrossRef] [PubMed]
44. Jiang, L.; Zhang, J.; Xuan, P.; Zou, Q. Bp neural network could help improve pre-miRNA identification in various species. *Biomed. Res. Int.* **2016**, *2016*, 9565689. [CrossRef] [PubMed]
45. Zou, Q.; Guo, J.; Ju, Y.; Wu, M.; Zeng, X.; Hong, Z. Improving tRNAscan-se annotation results via ensemble classifiers. *Mol. Inform.* **2015**, *34*, 761–770. [CrossRef] [PubMed]
46. Zou, Q.; Wang, Z.; Guan, X.; Liu, B.; Wu, Y.; Lin, Z. An approach for identifying cytokines based on a novel ensemble classifier. *Biomed. Res. Int.* **2013**, *2013*, 686090. [CrossRef] [PubMed]
47. Pan, Y.; Wang, Z.; Zhan, W.; Deng, L. Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* **2017**, *34*, 1473–1480.
48. 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**, *5*. [CrossRef] [PubMed]

49. Deng, L.; Chen, Z. An integrated framework for functional annotation of protein structural domains. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 902–913. [CrossRef] [PubMed]
50. Zhang, W.; Niu, Y.; Xiong, Y.; Zhao, M.; Yu, R.; Liu, J. Computational prediction of conformational b-cell epitopes from antigen primary structures by ensemble learning. *PLoS ONE* **2012**, *7*, e43575. [CrossRef] [PubMed]
51. Zhang, W.; Niu, Y.; Zou, H.; Luo, L.; Liu, Q.; Wu, W. Accurate prediction of immunogenic t-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. *PLoS ONE* **2015**, *10*, e0128194. [CrossRef] [PubMed]
52. Li, D.; Luo, L.; Zhang, W.; Liu, F.; Luo, F. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinform.* **2016**, *17*, 329. [CrossRef] [PubMed]
53. Zhang, W.; Zou, H.; Luo, L.; Liu, Q.; Wu, W.; Xiao, W. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* **2016**, *173*, 979–987. [CrossRef]
54. Zhang, W.; Shi, J.; Tang, G.; Wu, W.; Yue, X.; Li, D. Predicting small RNAs in bacteria via sequence learning ensemble method. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 643–647.
55. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. Mlacp: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121–77136. [CrossRef] [PubMed]
56. Zou, Q.; Chen, W.; Huang, Y.; Liu, X.; Jiang, Y. Identifying multi-functional enzyme by hierarchical multi-label classifier. *J. Comput. Theor. Nanosci.* **2013**, *10*, 1038–1043. [CrossRef]
57. Zhang, W.; Zhu, X.; Fu, Y.; Tsuji, J.; Weng, Z. The prediction of human splicing branchpoints by multi-label learning. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 254–259.
58. Zhang, W.; Zhu, X.; Fu, Y.; Tsuji, J.; Weng, Z. Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinform.* **2017**, *18*, 464. [CrossRef] [PubMed]
59. Song, L.; Li, D.; Zeng, X.; Wu, Y.; Guo, L.; Zou, Q. nDNA-prot: Identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinform.* **2014**, *15*, 298. [CrossRef] [PubMed]
60. Wang, C.; Hu, L.; Guo, M.; Liu, X.; Zou, Q. Imdc: An ensemble learning method for imbalanced classification with miRNA data. *Genet. Mol. Res.* **2015**, *14*, 123–133. [CrossRef] [PubMed]
61. Li, D.; Ju, Y.; Zou, Q. Protein folds prediction with hierarchical structured SVM. *Curr. Proteom.* **2016**, *13*, 79–85. [CrossRef]
62. Lin, C.; Zou, Y.; Qin, J.; Liu, X.; Jiang, Y.; Ke, C.; Zou, Q. Hierarchical classification of protein folds using a novel ensemble classifier. *PLoS ONE* **2013**, *8*, e56499. [CrossRef] [PubMed]
63. Zhang, J.; Zhang, Z.; Wang, Z.; Liu, Y.; Deng, L. Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Bioinformatics* **2018**, *34*, 1750–1757. [CrossRef] [PubMed]
64. Burdukiewicz, M.; Sobczyk, P.; Rödiger, S.; Duda-Madej, A.; Mackiewicz, P.; Kotulska, M. Amyloidogenic motifs revealed by n-gram analysis. *Sci. Rep.* **2017**, *7*, 12961. [CrossRef] [PubMed]
65. Lin, C.; Chen, W.; Qiu, C.; Wu, Y.; Krishnan, S.; Zou, Q. Libd3c: Ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing* **2014**, *123*, 424–435. [CrossRef]
66. Liu, B. BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* **2017**. [CrossRef] [PubMed]
67. Varadi, M.; De Baets, G.; Vranken, W.F.; Tompa, P.; Pancsa, R. Amypro: A database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.* **2018**, *46*, D387–D392. [CrossRef] [PubMed]
68. Wei, L.; Tang, J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2017**, *384*, 135–144. [CrossRef]
69. Zhang, N.; Huang, T.; Cai, Y.D. Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Mol. Genet. Genom.* **2015**, *290*, 343–352. [CrossRef] [PubMed]
70. Zou, Q.; Li, X.; Jiang, Y.; Zhao, Y.; Wang, G. Binmempredict: A web server and software for predicting membrane protein types. *Curr. Proteom.* **2013**, *10*, 2–9. [CrossRef]
71. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef] [PubMed]

72. Basu, S.; Söderquist, F.; Wallner, B. Proteus: A random forest classifier to predict disorder-to-order transitioning binding regions in intrinsically disordered proteins. *J. Comput. Aided Mol. Des.* **2017**, *31*, 453–466. [CrossRef] [PubMed]
73. Liu, Z.-P.; Wu, L.-Y.; Wang, Y.; Zhang, X.-S.; Chen, L. Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics* **2010**, *26*, 1616–1622. [CrossRef] [PubMed]
74. Zhang, N.; Li, B.Q.; Gao, S.; Ruan, J.S.; Cai, Y.D. Computational prediction and analysis of protein γ-carboxylation sites based on a random forest method. *Mol. Biosyst.* **2012**, *8*, 2946–2955. [CrossRef] [PubMed]
75. Shu, Y.; Zhang, N.; Kong, X.; Huang, T.; Cai, Y.D. Predicting A-to-I RNA editing by feature selection and random forest. *PLoS ONE* **2014**, *9*, e110607. [CrossRef] [PubMed]
76. Manavalan, B.; Lee, J.; Lee, J. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE* **2014**, *9*, e106542. [CrossRef] [PubMed]
77. Dao, F.-Y.; Yang, H.; Su, Z.-D.; Yang, W.; Wu, Y.; Hui, D.; Chen, W.; Tang, H.; Lin, H. Recent advances in conotoxin classification by using machine learning methods. *Molecules* **2017**, *22*, 1057. [CrossRef] [PubMed]
78. Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* **2018**. [CrossRef] [PubMed]
79. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. Aippred: Sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.* **2018**, *9*, 276. [CrossRef] [PubMed]
80. Manavalan, B.; Lee, J. Svmqa: Support–vector-machine-based protein single-model quality assessment. *Bioinformatics* **2017**, *33*, 2496–2503. [CrossRef] [PubMed]
81. Lin, H.; Ding, C.; Song, Q.; Yang, P.; Ding, H.; Deng, K.-J.; Chen, W. The prediction of protein structural class using averaged chemical shifts. *J. Biomol. Struct. Dyn.* **2012**, *29*, 643–649. [CrossRef] [PubMed]
82. Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. *Front Microbiol.* **2018**, *9*, 476. [CrossRef] [PubMed]
83. Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N^4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **2017**, *33*, 3518–3523. [CrossRef] [PubMed]
84. Lai, H.-Y.; Chen, X.-X.; Chen, W.; Tang, H.; Lin, H. Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* **2017**, *8*, 28169–28175. [CrossRef] [PubMed]
85. Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* **2018**, *9*, 1944–1956. [CrossRef] [PubMed]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).