Du *et al. J of Biol Res-Thessaloniki* 2016, 23(Suppl 1):S10
DOI 10.1186/s40709-016-0046-7

Journal of Biological Research-Thessaloniki

RESEARCH Open Access

Prediction of protein–protein interaction sites by means of ensemble learning and weighted feature descriptor

Xiuquan Du[1,3*], Shiwei Sun[1], Changlin Hu[1], Xinrui Li[1] and Junfeng Xia[2,3,4*]

Abstract

Background: Reliable prediction of protein–protein interaction sites is an important goal in the field of bioinformatics. Many computational methods have been explored for the large-scale prediction of protein–protein interaction sites based on various data types, including protein sequence, structural and genomic data. Although much progress has been achieved in recent years, the problem has not yet been satisfactorily solved.

Results: In this work, we presented an efficient approach that uses ensemble learning algorithm with weighted feature descriptor (EL-WFD) to predict protein–protein interaction sites. Moreover, weighted feature descriptor was designed to describe the distance influence of neighboring residues on interaction sites. The results on two dataset (Hetero and Homo), show that the proposed method yields a satisfactory accuracy with 83.8 % recall and 96.3 % precision on the Hetero dataset and 84.2 % recall and 96.3 % precision on the Homo dataset, respectively. In both datasets, our method tend to obtain high Mathews correlation coefficient compared with state-of-the-art technique random forest method.

Conclusions: The experimental results show that the EL-WFD method is quite effective in predicting protein–protein interaction sites. The novel weighted feature descriptor was proved to be promising in discovering interaction sites. Overall, the proposed method can be considered as a new powerful tool for predicting protein–protein interaction sites with excellence performance.

Background

Protein–protein interactions (PPIs) are central to all aspects of biological systems including, for example, gene regulation, immunological recognition and protein synthesis [1, 2]. Exploiting the mechanisms of protein interactions plays a pivotal role for understanding the functions of biological systems. Hence, identification of binding sites between two interacting proteins is one of basic problems in the research of protein functions. Knowledge of the three-dimensional (3D) structure of the protein complex provides much valuable information on the protein interaction site. Several experimental technologies such as X-ray crystallography and NMR can be used to obtain such information. However, they cannot meet the requirements of proteomics-generated interaction data since they are time consuming and expensive. Therefore, reliable and efficient computational methods are required to assist the identification of protein–protein interaction sites.

A number of computational methods have been proposed for the prediction of interaction sites in proteins based on the sequence information [3, 4], 3D structure information [5] or a combination of 3D structure and sequence information. Machine learning methods such as support vector machine (SVM) [6–8], neural networks (NN) [9–12], Bayesian networks (BN) [13–16], random forests (RF) [17, 18], conditional random fields (CRF) [19], extreme learning machine (ELM) [20] and L1-logreg classifier [21] have been successful applied for predicting binding sites. Therefore, development of a machine

*Correspondence: dxqllp@163.com; jfxia@ahu.edu.cn
[1] School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China
[2] Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei 230601, Anhui, China
Full list of author information is available at the end of the article

learning based model using protein properties might be a promising strategy to predict unknown PPI sites.

In this study, we present a novel method for PPI sites discovery and prediction that uses weighted feature descriptor derived from protein sequence with ensemble learning algorithm. Firstly, we systematically investigated a wide variety of features from a combination of protein sequence and structure information, and then weighted feature descriptor (WFD) was used to encode the PPI sites. Secondly, meta-algorithm was chosen as the ensemble learning method to identify PPI sites. Finally, a new ensemble classifier, namely EL-WFD, was developed to further improve the prediction accuracy. To demonstrate its effectiveness, the proposed method was applied to both the Hetero and Homo datasets. Empirical studies showed the efficiency and effectiveness of our proposed approach.

Results and discussion

Comparing the prediction performance with/without WFD on the TRS

Four models were generated with/without WFD on the TRS, namely WFD-Hetero, noWFD-Hetero, WFD-Homo and noWFD-Homo, respectively. Then, fivefolds cross validation was used to evaluate the performance of different methods on the TRS. Table 1 shows the detail results of four methods. From Table 1, we can deduce that the average performance of WFD-Hetero is higher about 0.7 % in Accuracy, 0.4 % in Recall, 1.6 % in Precision than the noWFD-Hetero, respectively. The average performance of WFD-Homo is higher about 0.26 % in Accuracy and 0.9 % in Precision than the noWFD-Homo, respectively.

Table 1 The performance of four methods on the Hetero/Homo TRS

Method	Acc	Pre	Rec
noWFD-hetero	92.52	95.1	83.3
WFD-hetero	93.2	96.7	83.7
noWFD-homo	92.92	95.3	82.8
WFD-homo	93.18	96.2	82.8

Acc Accuracy, *Pre* Precision, *Rec* Recall

Comparison with the other methods on the TES

In this study, we compared EL-WFD to J48 algorithm and RF using the same set. The results are shown in Table 2. The overall performance (Accuracy, Recall, Precision, F-measure and MCC) of our method were 93.11, 83.83, 96.3, 89.63 % and 0.8497 on the Hetero dataset. The success rate of J48 and RF was 88.29 and 76.15 % on the Hetero dataset. On the Homo dataset, the success rate of EL-WFD was 93.99 %. Hence, the success rate was improved by at least 6 %, while the overall Recall, Precision, F-measure and MCC were improved by at least about 1, 12, 6 and 10 % respectively.

ROC curves (Fig. 1) are also plotted to compare these three methods objectively on the Homo TES. From Fig. 1, it is found that Bagging is higher than the other methods on Homo test dataset. Figure 2 shows the same result with Fig. 1.

Conclusions

In this paper, we have developed a new approach for PPI sites prediction, which combine ensemble learning method and weighted feature descriptor (EL-WFD). EL offers significant advantages such as fast learning speed, ease of implementation, better generalization performance, and least human intervention. WFD is an effective feature representation method, which can uncover distance influence of neighboring residues on interacting sites. Experimental results show that our method performed significantly well in distinguishing interacting and non-interacting sites. In both datasets, our method tend to obtain high Mathews correlation coefficient (MCC) compared with state-of-the-art technique random forest method. In the future, we will focus on how to predict hot spots in protein interfaces.

Methods

Generation of the datasets

We evaluated the proposed method with the same dataset used in the study of Koike et al. [22]. The PPI sites dataset was collected from the Protein Data Bank (PDB). The protein pairs which contain a protein with few than 100 residues, or have more than 25 % sequence identity

Table 2 The performance comparison using different machine learning methods

Dataset	Classifier	Acc	Rec	Pre	F	MCC
Hetero	J48	88.29	83.29	83.69	83.49	0.7441
	RF	76.15	43.25	80.7	56.32	0.4576
	EL-WFD	93.11	83.83	96.3	89.63	0.8497
Homo	J48	87.97	81.51	80.62	81.06	0.7224
	RF	79.46	82.42	44.46	57.76	0.4956
	EL-WFD	93.99	84.19	96.34	89.86	0.8601

Acc Accuracy, *Rec* Recall, *Pre* Precision, *F* F-measure, *MCC* Mathews correlation coefficient

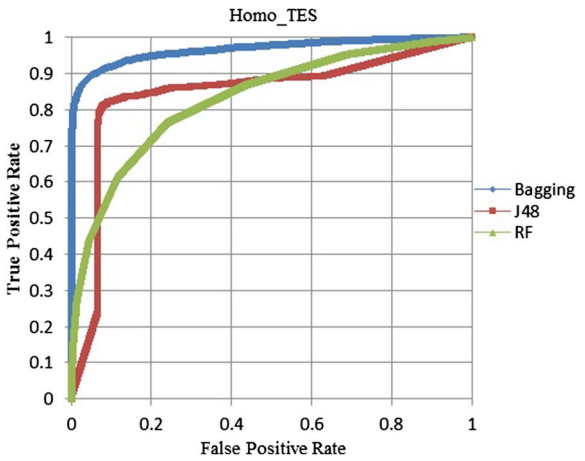

Fig. 1 ROC curves on the homo TES

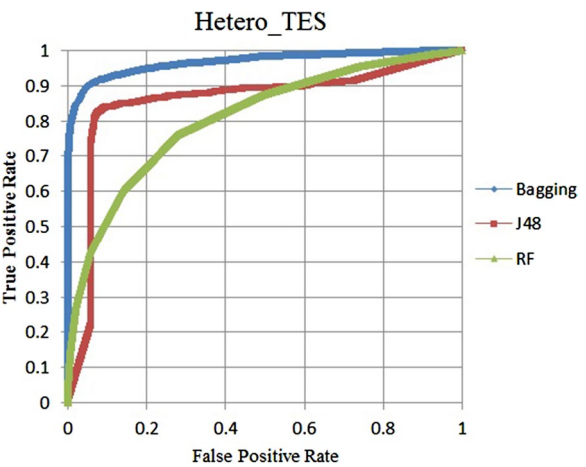

Fig. 2 ROC curves on the hetero TES

were removed. In addition, the number of protein–protein interaction sites less than 20 interfacial residues and 30 interfacial residues for heterocomplex and homocomplexes respectively, or no HSSP entry available [23] are removed. The remaining 559 non-redundant chains, where 270 are from hetero complexes and the other 289 are from homo complexes, comprise the final dataset. We randomly select 202 chains of all chains as Train Set (TRS) and 87 chains as Test Set (TES) from Homo dataset. We also select 189 chains of all chains as TRS and 81 chains as TES from Hetero dataset.

Definition of protein interaction sites

Interfaces are formed mostly by residues that are exposed to the solvent if the partner chain is removed, so we focus on surface residues for later prediction. A residue is considered to be a surface residue if at least 16 % of the solvent accessible surface area (ASA) was exposed to solvent [24]. The ASA of each residue in the unbound molecule (MASA) and in the complex (CASA) is computed using the DSSP program [25]. Meanwhile, a surface residue is defined to be an interface residue if it formed an interfacial contact (|MASA-CASA| >=1).

Feature extraction for residues

To use machine learning methods to predict PPI sites, one of the most important computational challenges is to extract the biological characteristics in which the important information content of amino acid residues is fully encoded. In this study, we extracted feature vectors based residue structure, sequence, and physicochemical information.

Structure based features

1. Accessible surface area: The accessible surface area (ASA) is the atomic surface area exposed to a solvent. The ASA value of each residue calculated by DSSP was used in our work. In addition, protein structure and interaction analyzer (PSAIA) calculates the ASA value for each residue, including backbone ASA, side-chain ASA, polar ASA and non-polar ASA.
2. Relative accessible surface area: Relative accessible surface area (RASA) extracted in this work was calculated by PSAIA [26]. The following residue attributes are calculated by PSAIA: total RASA, backbone RASA, side-chain RASA, polar RASA and non-polar RASA.
3. Depth index: The residue depth is defined as the minimum distance of a residue from any solvent accessible residue and it has been computed by PSAIA. For residue depth, there are six features were calculated by PSAIA. In this paper, the average depth index (DPX) is used.
4. Protrusion index: The protrusion of a non-hydrogen residue is the ration of the volume of a sphere with a radius of 10.0 Å centered at that atom that is not filled with atoms. Same with the DPX, PSAIA calculates six features for the protrusion and the average protrusion index (CX) is adopted.

Sequence based features

1. Properties from HSSP file: The sequence profile in HSSP file for each protein chain are composed of L rows and 20 columns. 'L' stands for the number of amino acids in a chain and 20 kinds of amino acids index columns. $P_{i,j}$ means the probability of j-th amino acid take the place of the i-th residue. We also extracted the other four properties of protein from HSSP [27] database: entropy, relative entropy, con-

servation weight and sequence variability. Entropy measures the conservation of a residue in the location. Relative Entropy is defined as the standardized entropy which normalized to the scale of 0–100. Conservation Weight measures the sequence conservation of a position. Sequence variability contains evolutionary information, on a scale of 0–100 as exported from NAGLIN alignments.

Physicochemical features

1. High-quality-indices: Saha et al. [28] have made a conclusion that physicochemical features of amino acids play a significant role in identifying the PPI sites, thus properties of amino acids are taken into count as important characteristics in discriminating between interacting sites and non-interacting sites. Recently, 544 physicochemical and biochemical properties of amino acids are released in AAIndex1 database. Based on the statistical analyses, Saha et al. [28] categorized these 544 characteristics into eight classes, named high-quality-indices (HQIs). In this work, the HQI8 containing eight clusters named electric properties (BLAM930101), hydrophobicity (BIOV880101), alpha and turn propensities (MAXF760101), physicochemical properties (TSAJ990101), residue propensity (NAKH920108), composition (CEDJ970104), beta propensity (LIFS790101) and intrinsic propensities (MIYS990104), respectively. Each cluster is composed of one value and there are eight indices for each amino acid.
2. Amino acid factors: Based on AAindex1, Atchley et al. [29] made statistical analyses on these 544 properties, as well. Different from HQI, they summarized these properties into five patterns, which reflect polarity, secondary structure, molecular volume, codon diversity and electrostatic charge. These features were also used to evaluate protein interaction sites.

The WFD of the residue

The environment factors for each residue position are very important for PPI sites, so the profiles of sequentially neighboring residues or spatially neighboring residues were adopted as residue features in PPI site prediction in previous report [20]. However, distance effect among these sequentially neighboring residues or spatially neighboring residues was not considered. In other words, as the distance between the query residue and its neighboring residue increases, the neighboring residue will have smaller effect on the query residue, and vice versa. Therefore, we propose a novel WFD which considers the distance effect among the query residue and its neighboring residue. To illustrate the WFD, for example, given the protein sequence segment SLDIQSAA and Q is the interaction site (query residue). In this case, the sliding window is fixed to five and sequentially neighboring is considered. Thus, the feature vector components were arranged in ascending order according to the distance between the neighboring residues, which can be defined as follows,

$(V = V_D, V_I, V_Q, V_S, V_A)$,

where $V_{residue} = (HSSP, PSAIA, HIQ, AAFactors), residue = D, I, Q, S, A$.

Second, we calculate the distance effect according to C_α -atom coordinate of D, I, Q, S and A. Here, the Euclidean distance is used to evaluate the distance effect among the residue, which can be calculated as,

$$\mathrm{ED_{D,Q}} = \sqrt{(x_D - x_Q)^2 + (y_D - y_Q)^2 + (z_D - z_Q)^2}$$

$$\mathrm{ED_{I,Q}} = \sqrt{(x_I - x_Q)^2 + (y_I - y_Q)^2 + (z_I - z_Q)^2}$$

$$ED_{Q,Q} = 1$$

$$\mathrm{ED_{S,Q}} = \sqrt{(x_S - x_Q)^2 + (y_S - y_Q)^2 + (z_S - z_Q)^2}$$

$$\mathrm{ED_{A,Q}} = \sqrt{(x_A - x_Q)^2 + (y_A - y_Q)^2 + (z_A - z_Q)^2}$$

where x_D denotes the x coordinate of C_α-atom, y_D denotes the y coordinate of C_α-atom, and z_D denotes the z coordinate of C_α-atom of residue D, respectively. The rest symbols have similar meanings as those used for residue D.

Finally, the WFD can be written as,

$$\begin{aligned} WFD_Q &= \left(\frac{V_D}{ED_{D,Q}}, \frac{V_I}{ED_{I,Q}}, \frac{V_Q}{ED_{Q,Q}}, \frac{V_S}{ED_{S,Q}}, \frac{V_A}{ED_{A,Q}}\right) \\ &= \left(\frac{V_D}{ED_{D,Q}}, \frac{V_I}{ED_{I,Q}}, V_Q, \frac{V_S}{ED_{S,Q}}, \frac{V_A}{ED_{A,Q}}\right) \end{aligned}$$

The feature space

For each residue, 50 features were extracted including 25 features from HSSP profile, 12 features from structure information (5 features from ASA, 5 features from RASA, 1 feature from DPX, 1 feature from CX), and 13 features from physicochemical information (8 features from HQI8 and 5 features from amino acid factors). In addition, taking into consideration the effect of neighbor residues, 11-size sliding window is used to describe current residue. Therefore, $50 \times 11 = 550$ features were extracted for each residue.

Ensemble meta-algorithm

In this paper, bagging algorithm is used to implement ensemble meta-algorithm which improves the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and avoids over-fitting. Although it is usually applied to decision tree methods, it can be used with any type of method.

Suppose a standard training set D of size n, bagging will produce m new training sets D_i with size n', by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If $n\prime = n$, then for large n the set D_i is expected to have the fraction (1−1/e) (≈63.2 %) of the unique examples of D, the rest being duplicates. This kind of sample is known as a bootstrap sample. The m models are fitted using the above m bootstrap samples and combined by voting for classification.

Performance evaluation

PPI sites prediction is a binary classification problem. In this experiment, precision (Pre), recall (Rec), accuracy (Acc), F-measure (F), and Matthews correlation coefficient (MCC) were employed to measure the performance of classifiers:

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

$$\mathrm{Pr}\,e = \frac{TP}{TP + FP} \tag{2}$$

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

$$F = \frac{2 \times \mathrm{Pr}\,e \times \mathrm{Rec}}{\mathrm{Pr}\,e + \mathrm{Rec}} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

where true positive (TP) denotes the number of true interaction site, true negative (TN) denotes the number of true non-interaction site, FP (False Positive) denotes the number of false interaction site, and false negative (FN) denotes the number of false non-interaction site. The ROC curve is often used to evaluate classifier performance. A classifier conducts predictions on the basis of a threshold, which generally is defined as 0.5. When the threshold value is changed, new predictions can be obtained and a point can be plotted with the true positive rate (TPR) *versus* the false positive rate (FPR) for different threshold values.

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

The area under a curve (AUC) for the receiver operating characteristic (ROC) curve is also used. When the AUC value of a predictor is larger than the area of other ROC curves, such a predictor is considered better than other predictors.

Authors' contributions

XD, SS, CH, XL and JX contributed to algorithm design and implementation. XD and JX contributed to manuscript writing. All authors read and approved the final manuscript.

Author details

[1] School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China. [2] Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei 230601, Anhui, China. [3] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230601, Anhui, China. [4] Institute of Health Sciences, Anhui University, Hefei 230601, Anhui, China.

Acknowledgements

The authors want to thank the editor and anonymous reviewers for helpful comments and suggestions.

Competing interests

The authors declare that they have no competing interests.

Declarations

The publication costs for this article were partly funded by grants from the National Science Foundation of China (61203290 and 31301101), and partly supported by the Anhui Provincial Natural Science Foundation (1408085QF106), the Specialized Research Fund for the Doctoral Program of Higher Education (20133401120011), the Technology Foundation for Selected Overseas Chinese Scholars from Department of Human Resources and Social Security of Anhui Province (No. [2014]-243), the Outstanding Young Backbone Teachers Training (02303301), Provincial Natural Science Research Program of Higher Education Institutions of Anhui province (KJ2016A016) and the Co-Innovation Center for Information Supply & Assurance Technology of Anhui University. This article has been published as part of Journal of Biological Research—Thessaloniki, Volume 23, Supplement 1, 2016: Proceedings of the 2014 International Conference on Intelligent Computing. The full contents of the supplement are available online at http://jbiolres.biomedcentral.com/articles/supplements/volume-23-supplement-1.

Published: 4 July 2016

References

1. Zhou HX. Improving the understanding of human genetic diseases through predictions of protein structures and protein-protein interaction sites. Curr Med Chem. 2004;11:539–49.
2. Zhou HX, Qin S. Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics. 2007;23:2203–9.
3. Chen H, Zhou HX. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins. 2005;61:21–35.
4. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins. 2001;44:336–43.

5. Wang B, Wong HS, Huang DS. Inferring protein–protein interacting sites using residue conservation and evolutionary information. Protein Pept Lett. 2006;13:999–1005.
6. Wong GY, Leung FH, Ling SH. Predicting protein-ligand binding site using support vector machine with protein properties. Ieee Acm T Comput Bi. 2013;10:1517–29.
7. Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics. 2005;21:1487–94.
8. Wang B, Chen P, Huang DS, Li JJ, Lok T-M, Lyu MR. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett. 2006;580:380–4.
9. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem. 2002;269:1356–61.
10. Ofran Y, Rost B. Predicted protein–protein interaction sites from local sequence information. FEBS Lett. 2003;544:236–9.
11. Pettit FK, Bare E, Tsai A, Bowie JU. HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. J Mol Biol. 2007;369:863–79.
12. Li BQ, Feng KY, Ding J, Cai YD. Predicting DNA-binding sites of proteins based on sequential and 3D structural information. Mol Genet Genomics. 2014;289:489–99.
13. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol. 2004;338:181–99.
14. Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR. Insights into protein–protein interfaces using a Bayesian network prediction method. J Mol Biol. 2006;362:365–86.
15. Mayrose I, Graur D, Ben-Tal N, Pupko T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol. 2004;21:1781–91.
16. Bhaskara RM, Padhi A, Srinivasan N. Accurate prediction of interfacial residues in two-domain proteins using evolutionary information: implications for three-dimensional modeling. Proteins. 2013;82:1219–34.
17. Šikić M, Tomić S, Vlahoviček K. Prediction of protein–protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol. 2009;5:e1000278.
18. Li BQ, Feng KY, Chen L, Huang T, Cai YD. Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. PLoS ONE. 2012;7:e43927.
19. Li MH, Lin L, Wang XL, Liu T. Protein–protein interaction site prediction based on conditional random fields. Bioinformatics. 2007;23:597–604.
20. Wang DD, Wang R, Yan H. Fast prediction of protein–protein interaction sites based on Extreme Learning Machines. Neurocomputing. 2014;128:258–66.
21. Dhole K, Singh G, Pai PP, Mondal S. Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier. J Theor Biol. 2014;348:47–54.
22. Koike A, Takagi T. Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel. 2004;17:165–73.
23. Dodge C, Schneider R, Sander C. The HSSP database of protein structure—sequence alignments and family profiles. Nucleic Acids Res. 1998;26:313–5.
24. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins. 1994;20:216–26.
25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
26. Mihel J, Šikić M, Tomić S, Jeren B, Vlahoviček K. PSAIA–protein structure and interaction analyzer. BMC Struct Biol. 2008;8:21.
27. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9:56–68.
28. Saha I, Maulik U, Bandyopadhyay S, Plewczynski D. Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids. 2012;43:583–94.
29. Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci USA. 2005;102:6395–400.