Biointerface Research in Applied Chemistry
Platinum Open Access Journal (ISSN: 2069-5837)

Review
Volume 11, Issue 4, 2021, 11233 - 11243
https://doi.org/10.33263/BRIAC114.1123311243

Recent Trends in Machine Learning-based Protein Fold Recognition Methods

Apurva Mehta [1,*], Himanshu Mazumdar [2]

[1] Assistant Professor, Department of Computer Engineering, Faculty of Technology, Dharmsinh Desai University; apurvamehta.ce@ddu.ac.in (A.M.);
[2] Head, R&D Center, Faculty of Technology, Dharmsinh Desai University; hsmazumdar@ddu.ac.in (H.M.);
* Correspondence: apurvamehta.ce@ddu.ac.in;

Scopus Author ID 57212214531

Received: 11.11.2020; Revised: 5.12.2020; Accepted: 7.12.2020; Published: 10.12.2020

Abstract: Proteins are macromolecules that enable life. Protein function is due to its three-dimensional structure and shape. It is challenging to understand how a linear sequence of amino acid residues folds into a three-dimensional structure. Machine learning-based methods may help significantly in reducing the gap present between known protein sequence and structure. Identifying protein folds from a sequence can help predict protein tertiary structure, determine protein function, and give insights into protein-protein interactions. This work focuses on the following aspects. The kind of features such as sequential, structural, functional, and evolutionary extracted for representing protein sequence and different methods of extracting these features. This work also includes details of machine learning algorithms used with respective settings and protein fold recognition structures. Detailed performance comparison of well-known works is also given.

Keywords: evolutionary; protein fold; protein secondary structure; random forest algorithm; spatial; structural classification of protein

1. Introduction

Protein Sequences consist of 20 standard amino acids, which fold into their respective three-dimensional structure. The protein function is due to its three-dimensional structure and shape [1,2]. It is challenging to understand how a chain of amino acid residues is folded into its three-dimensional structure. There is a wide gap in sequence and structure availability. Many experimental methods are currently used to determine protein structure, including X-ray crystallography [3,4] and NMR spectroscopy [5]. These methods cannot help reduce the vast amount of gap present between sequence and structure, as they are slow and much costlier [6]. The machine learning-based methods may help significantly in reducing this gap. It is a challenging task to predict a protein tertiary structure from a protein sequence directly. Identifying a protein fold from a protein sequence can help predict a protein tertiary structure and function. An in-silico method for protein fold recognition has many applications in biology, chemistry, and medicine [7–12].

Identifying a fold category of a protein sequence is called fold recognition [13–16]. The process of protein fold recognition is summarized in the following Figure 1. The different types of feature vectors are created from input protein sequences. Features vectors are combined for use by machine learning algorithms. Combining features plays an important role in representing protein sequence information for building a machine learning model.

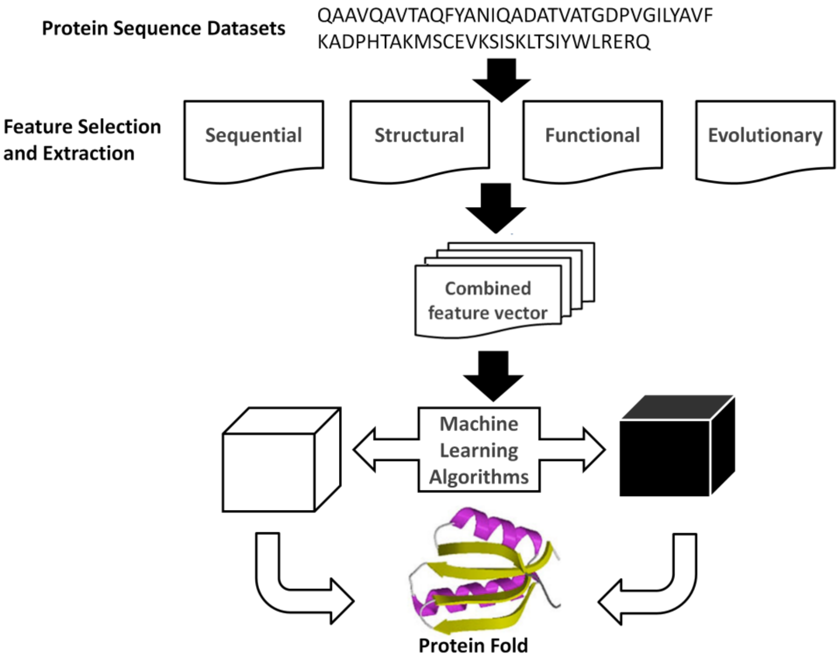
Protein Sequence Datasets
QAAVQAVTAQFYANIQADATVATGDPVGILYAVF
KADPHTAKMSCEVKSISKLTSIYWLRERQ
Feature Selection and Extraction
Sequential
Structural
Functional
Evolutionary
Combined feature vector
Machine Learning Algorithms
Protein Fold

Figure. 1. Protein fold recognition process using machine learning algorithms.

2. Datasets

The popular datasets used in protein fold recognition are DD [17], EDD [18], and TG [19]. SCOPe [20] dataset also contains fold information (i.e. 1003 folds) for protein sequence[21,22]. Statistically, the SCOPe dataset is better in terms of the number of records per fold compared to other datasets like DD, EDD, and TG [21]. Following Figure 2 highlights the growth of the Structural Classification of Protein Dataset. Each entry in SCOP or SCOPe provides lots of information, viz. protein structural class, protein fold, protein super-family, protein family belonging to a string containing amino acid resides. It also provides additional information of PDB id and PDB chain id for each entry that concisely links with other resources. [20,23]

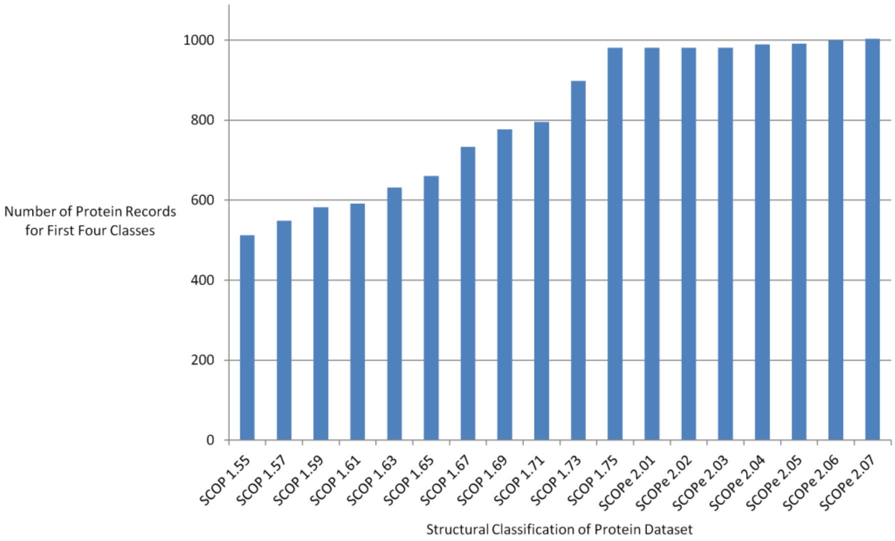

Figure. 2 Growth of structural classification of protein dataset.

3. Methods of feature extractions

The Protein sequence is in a textual format, and each sequence is likely to be of a different length. Thus, it is not amenable to use sequence it-self as the only feature for model building. It is also known that the linear sequence of amino acids contains all the information that is necessary to determine the final tertiary structure [24–26]. So, there is a requirement to look for different feature sets that can represent a given amino acid sequence. Majorly four kinds of features are used: sequential, structural, functional, and evolutionary [27–29].

The sequential feature extracted from the protein sequence is the amino acid composition. The number of times a particular amino acid is present in a sequence without considering its spatial location [30]. It helps in identifying frequently and rarely appearing amino acids in a given sequence. Although amino acid occurrences vary as the length vary. So, the frequency of occurrence must need to be normalized concerning the length of the sequence. The normalized scores can be directly used to compare protein sequences based on their amino acid composition. The amino acid composition is used as a feature in [17,19,31]. Any protein sequence P with N amino acid residues can be represented as in Equation 1. There are 20 standard amino acids available. Composition for a single amino acid residue is expressed by Equation 2. Then protein P of Equation 1 can be suitably represented by the composition of its 20 amino acid residues as P’ using Equation 3. As per Equation 3, each protein sequence can be represented as a 20-dimensional vector consisting of scalar values.

$$P = R_1 R_2 R_3 ... R_i ... R_N \quad (1)$$

where Ri is one of the amino acids among the 20 standard amino acids, i indicates the physical position of Ri in the protein sequence of length N.

$$c_i = \frac{100}{N} \times \sum (n_i) \quad (2)$$

Here, 1≤i≤20, $\sum n_i$ indicates the count of i amino acids in a given protein sequence, and N is the total number of residues in a given protein sequence.

$$P' = (c_1, c_2, ..., c_{20}) \quad (3)$$

In a study [32], to avoid completely ignoring the sequence-order effects, the pseudo-amino acid composition was used to replace the conventional amino acid composition. According to the typical PseAAC discrete model, protein *P’* of Equation 3 can be represented as P’’ in following Equation 4 [33].

$$P'' = (c_1, c_2, ..., c_{20}, c_{20+1}, c_{20+2}, ... c_{20+\lambda}), \lambda < L \quad (4)$$

where, L is the length of protein sequence, λ reflects the rank of correlation and

$$\begin{aligned} c_u &= \frac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, (1 \le u \le 20) \\ c_u &= \frac{\omega \tau_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, (20+1 \le u \le 20+\lambda) \end{aligned} \quad (5)$$

where f_u is the frequency of amino acid u, f_i is the frequency of amino acid i, ω is the weight factor, τ_{u-20} the u-20th and τ_k the kth tier correlation factor that reflects the hopping sequence order correlation between all the u-20th and kth most contiguous residues respectively as formulated by Equation 6. τ_{u-20} and τ_k are computed for different function/physicochemical properties (i.e., hydrophobic, polar, etc.) as given in Equation 7.

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k}, (k < L) \quad (6)$$

$$J_{i,i+k} = \frac{1}{\Gamma}\sum_{q=1}^{r}[\Phi_q(R_{i+k}) - \Phi_q(R_i)]^2 \quad (7)$$

where $\Phi_q(R_i)$ is the q^th function/physicochemical property of the amino acid R_i and Γ the total number of functions/ physicochemical properties considered.

Table 1. Physicochemical and Structural properties used for Fold Recognition

Physico-chemical and Structural Property	References
Hydrophobicity	[17,32,34-38]
Predicted Solvent Accessibility	[32,37-40]
Normalized van-der Waals volume	[17,31-32,34-37]
Polarity	[17,31-32,34-37]
Polarizability	[17,31-32,34-37]

The physicochemical and structure-based properties of amino acids used for protein fold recognition are shown in Table 1. The Hydrophobic effect is believed to play a pivotal role in the process of protein folding [41]. The Hydrophobicity property is used to measure how soluble an amino acid residue is in water. The Predicted solvent accessibility determines the solvent-exposed area of a protein. The Normalized van-der Waals volume is used to determine the level of packing density of molecules' interior. The polarity of the amino acids affects the overall structure of a protein. Polarizability is included in determining the dynamic response of a fold or structure to external fields. These all properties represent environmental factors affecting the formation of protein fold. Predicted Secondary Structure helps specify types of structural elements viz. helix, sheet, and turn/ loop present in the structure. Types, frequency, and location of secondary structural elements play a key role in predicting protein fold using secondary structure motifs. Three descriptors, “composition” (C), “transition” (T), and “distribution” (D), are calculated for a given property to describe the global percent composition of each of the three groups in a protein, the percent frequencies with which the attribute changes its index along the entire length of the protein, and the distribution pattern of the attribute along the sequence, respectively. The authors of [42] proposed a method of k-separated bi-gram probabilities extracted from Position Specific Scoring Matrix (PSSM)[43,44] representing sequential evolution probabilities, where k is an integer in the range 1 to 11. The following equation is useful for calculating k-separated bi-gram.

$$T_{m,n}(k) = \sum_{i=1}^{L-k} N_{i,m}N_{i+k,n} \quad (8)$$

where, 1≤m≤20, 1≤n≤20, 1≤k≤11, L is the length of protein sequence, and N can be PSSM matrix or composition matrix.

The authors use PSSM instead of the original primary sequence to avoid zero in the resulting bi-gram feature vector. The Bi-gram probabilities calculated from PSSM are used as features in [45]. They are not considering k-separated bi-grams. The Bi-gram probabilities are also used in [34] for the TG dataset. Similarly, in [46], a tri-gram extraction technique is given. They have used PSSM linear probabilities of a given protein sequence to compute individual trigrams' probabilities to form the 3-dimensional probability matrix. The PSSM feature itself is used in [36] as one among other features like physicochemical properties and functional domain-based features from Conserved Domain Database (CDD). Previous studies of the PSI-BLAST profile show that evolutionary information is more informative than the query sequence itself, so the PSSM is transformed into a fixed-length vector by AutoCross Covariance (ACC) [18]. Evolutionary features are extracted using the profile-profile sequence alignment method HHblits using PSSM [47]. In one of the most recent works [13], physicochemical, evolutionary, and structural features are combined to create a multi-view model for protein fold recognition. Pse-AAC is used for representing the physicochemical profile. Evolutionary features are extracted using ACC transformation and the HHblits method of profile-profile sequence alignment. Secondary features are extracted from secondary structure profiles predicted using the PSI-PRED server [48,49]. Secondary structure feature vectors primarily include the probability of secondary structure elements (helix, strand, and coil), the entropy of secondary structure elements, ACC, Bi-gram, and Tri-gram of predicted secondary structure. Probability and entropy-based features are given in Equation 9. Later they combine the multi-view model with template-based methods HHblits and HMMER to create an ensemble model.

$$P_C = \frac{N_C}{L}, P_E = \frac{N_E}{L}, P_H = \frac{N_H}{L} \qquad (9)$$
$$entropy = -(P_C \ln P_C + P_E \ln P_E + P_H \ln P_H)$$

where, N_C, N_E, N_H represent the frequency of coil, strand, and helix respectively, L is the protein sequence's length.

4. Machine learning algorithms

The exponential growths of biological data need algorithms to identify important parameters and features while performing tasks intelligently [50–54]. Machine learning models are the one which can fulfill this requirement. Currently, for protein fold recognition, single classifier and ensemble classifier based methods are popularly in use. The single classifier-based methods classify new records based on the classifier's prediction, while ensemble classifier methods classify new records based on the vote of their classifiers’ predictions [55–57].

Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used as a single classifier for protein fold recognition. ANN is an ML approach that models a network mathematically based on neurons' model in living organisms to carry out learning and other computational tasks [53,58]. Neurons of networks are arranged in layers, and normally the network has three layers: Input, Hidden, and Output. The feature vector is provided to the input layer. Learning takes place in hidden layers with weight propagation and weight update mechanism. The output layer lets you interpret output based on learning performed [59]. A Support Vector Machine predicts by finding the decision boundary that maximizes the target classes' margin [60].

Three-layer feed-forward neural networks (NN) are used with the NN weights adjusted by conjugate gradient minimization. Authors use physical, chemical, and structural properties for constructing feature vectors in the form of CTD feature vectors from protein sequences. In this work [37], a separate training set is constructed for each fold in the database, and NN is trained. Many discriminative methods use one-vs-others methods for prediction; authors in their study [17] investigate one-vs-others and all-vs-all methods with NN and SVM as a base classifier with the same global composition representation CTD for protein sequences. The

Grow and Learn NN with one-vs-other protein fold recognition methods used 125-dimensional data constructed from physicochemical properties of amino acids [31]. The properties considered are amino acid composition, predicted secondary structure, hydrophobicity, normalized van der Waals volume, polarity, and polarizability [61]. SCOPe 1.75 dataset sequences are utilized in deep neural network-based methods [62]. This work focuses on features based on sequence pair alignments and secondary structure prediction.

The pseudo amino acid composition with other chemical and evolutionary properties is used as a feature vector in [63] for predicting protein fold using a k-NN classifier. The pseudo amino acid composition is also used in an unsupervised machine learning method, where [64] each protein is associated with its corresponding PseAAC. This work uses the spectral graph clustering method for the prediction of protein fold.

In a study [18] based on autocross-covariance transformation LIBSVM, implementation of SVM is used with RBF as the kernel function. Kernel trick transforms linearly inseparable data into linearly separable data by mapping original data in higher dimensional space [65]. Normally RBF kernel transforms input data by taking squared Euclidean distance of input feature vectors. The RBF kernel is generally preferred with SVM compared to linear and polynomial kernel functions [66]. SVM algorithm is also used [67] for protein fold recognition using features of PSSM and SSPM. LIBSVM is used for tuning parameters. SVM implementation based on LIBSVM with RBF kernel is used in a work [47] based on sequence-sequence profiles for protein fold prediction.

An ensemble framework consists of 9 classifiers is used to reduce the variance caused by peculiarities of one training set combining all features. The classifier used is OET-KNN (optimized evidence-theoretic k-nearest neighbors). An ensemble output is selected by a voting scheme [32]. A two-level classification method that first predicts class and then folds uses MLP networks, RBF networks, SVM, and an ensemble of classifiers. Simple majority voting scheme and five folds cross-validation are used to fuse prediction outcomes in both levels [68]. A set of 11 SVM classifiers were used in [42] to build a model from feature vectors of k-separated bi-grams generated using PSSM probabilities. The work of [36] shows that the DSSP feature has a significant impact on improving classifier performance. After validating the random forest's predictive quality, SVM, nearest neighbor, Naïve Bayes, and multiple logistic regression ensemble classifiers, random forest is employed by PFP-RFSM [40]. An ensemble classifier consisting of five classifiers Random Forest, Naïve Bayes, Bayes Net, LibSVM, and SVM with SMO is constructed in WEKA [69]. A novel ensemble classifier comprising template free and template-based methods is proposed [13]. It utilizes sequential, evolutionary, and structural features for constructing template free linear regression model and profiles generated from HMM using homology templates of query sequence found using HHblits and HMMER in a template-based method. The template-based approach is also used with features derived from sequence and structural evolutionary information [39].

5. Performance comparison

There are many works using machine learning algorithms for protein fold recognition.

Table 2. Comparison of Protein Fold Recognition for Benchmark Datasets.

Dataset	Feature types	ML algorithm	Accuracy (%)	References
DD	PSSM, Functional Domain composition, Amino Acid	Ensemble Classifier	76.2	[36]

Dataset	Feature types	ML algorithm	Accuracy (%)	References
	Composition, Physico-chemical properties,			
DD	PSSM Bi-gram	GA, SVM	71.5	[42]
DD	Physicochemical property and Predicted Secondary Structure	Grow and Learn Neural Network	81.2	[31]
DD	Physicochemical property, Predicted Secondary Structure, PSSM	Ensemble Method	83.5	[13]
EDD	PSSM, Functional Domain composition, Amino Acid Composition, Physico-chemical properties,	Ensemble Classifier	93.2	[36]
EDD	PSSM Bi-gram	GA, SVM	87.7	[42]
EDD	Physicochemical property, Predicted Secondary Structure, PSSM	Ensemble Method	94.8	[13]
TG	PSSM, Functional Domain composition, Amino Acid Composition, Physico-chemical properties,	Ensemble Classifier	94.3	[36]
TG	PSSM Bi-gram	GA, SVM	75.8	[42]
TG	Physicochemical property, Predicted Secondary Structure, PSSM	Ensemble Method	85.1	[13]

These machine learning algorithms can be compared effectively using benchmark datasets. Table 2 provides a performance evaluation of significant methods reported since 2016.

6. Conclusions and Challenges

In this article, key insights are provided for protein fold recognition from protein sequences using a machine learning algorithm. They are in the form of dataset availability, feature representation methods for protein sequence, and popularly used machine learning algorithms with their performance comparison.

Many physicochemical properties are used to find features from a linear sequence of amino acids. Still, as performance is not as expected, other properties related to R-group like Aliphatic, Aromatic, Acidic, etc. may be included for feature representation. Apart from physicochemical properties, attention may be given to structure-based features for protein sequence representation as protein folds due to arrangements of its secondary structure in space relative to one another. The search space for performing protein fold recognition is too large. Efforts may be performed to reduce search space, which eventually impacts the protein fold recognition algorithm's performance. One key way to reduce search space is to predict protein structure class and then predict protein fold for each protein structure class.

Funding

This research received no external funding.

Acknowledgments

The authors are thankful to the Research and Development Center of Dharmsinh Desai University for allowing the use of their tools and resources necessary for current research work.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Sela, M.; Anfinsen, C.B.; Harrington, W.F. The correlation of ribonuclease activity with specific aspects of tertiary structure. *Biochim. Biophys. Acta* **1957**, *26*, 502-512, https://doi.org/10.1016/0006-3002(57)90096-3.
2. Kinoshita, K.; Nakamura, H. Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **2003**, *13*, 396-400, https://doi.org/10.1016/s0959-440x(03)00074-5.
3. Bragg, L. The development of X-ray analysis. *Contemporary Physics* **1976**, *17*, 103-104, https://doi.org/10.1080/00107517608210844.
4. Formanek, H.; Formanek, S. Protein crystallography by T. L. Blundell and L. N. Johnson. *Acta Crystallographica Section B* **1977**, *33*, 2702, https://doi.org/10.1107/S0567740877009339.
5. Baldwin, E.T.; Weber, I.T.; St Charles, R.; Xuan, J.C.; Appella, E.; Yamada, M.; Matsushima, K.; Edwards, B.F.; Clore, G.M.; Gronenborn, A.M. Crystal structure of interleukin 8: symbiosis of NMR and crystallography. *Proceedings of the National Academy of Sciences* **1991**, *88*, 502, https://doi.org/10.1073/pnas.88.2.502.
6. Gromiha, M.M. *Protein bioinformatics: from sequence to function*. academic press: **2010.**
7. Apurva, M.; Mazumdar, H. Predicting structural class for protein sequences of 40% identity based on features of primary and secondary structure using Random Forest algorithm. *Comput. Biol. Chem.* **2020**, *84*, 107164, https://doi.org/10.1016/j.compbiolchem.2019.107164.
8. Li, C.-C.; Liu, B. MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinform.* **2019**, https://doi.org/10.1093/bib/bbz133.
9. Liu, B.; Zhu, Y.; Yan, K. Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Brief. Bioinform.* **2019**, https://doi.org/10.1093/bib/bbz139.
10. Shao, J.; Liu, B. ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm. *Brief. Bioinform.* **2020**, https://doi.org/10.1093/bib/bbaa192.
11. Yu, C. H., Qin, Z., Martin-Martinez, F. J., & Buehler, M. J. A self-consistent sonification method to translate amino acid sequences into musical compositions and application in protein design using artificial intelligence. *ACS nano* ***2019***, *13*(7), 7471-7482. https://doi.org/10.1021/acsnano.9b02180
12. Refahi, M.S.; Mir, A.; Nasiri, J.A. A novel fusion based on the evolutionary features for protein fold recognition using support vector machines. *Sci. Rep.* **2020**, *10*, 14368, http://dx.doi.org/10.1038/s41598-020-71172-x.
13. Yan, K.; Fang, X.; Xu, Y.; Liu, B. Protein fold recognition based on multi-view modeling. *Bioinformatics* **2019**, *35*, 2982-2990, https://doi.org/10.1093/bioinformatics/btz040.
14. Sudha, P.; Ramyachitra, D.; Manikandan, P. Enhanced Artificial Neural Network for Protein Fold Recognition and Structural Class Prediction. *Gene Reports* **2018**, *12*, 261-275, https://doi.org/10.1016/j.genrep.2018.07.012.
15. Zheng, W.; Zhang, C.; Wuyun, Q.; Pearce, R.; Li, Y.; Zhang, Y. LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res.* **2019**, *47*, W429-W436, https://doi.org/10.1093/nar/gkz384.
16. Guo, Y.; Yan, K.; Wu, H.; Liu, B. ReFold-MAP: Protein remote homology detection and fold recognition based on features extracted from profiles. *Anal. Biochem.* **2020**, *611*, 114013, https://doi.org/10.1016/j.ab.2020.114013.
17. Ding, C.H.Q.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349-358, https://doi.org/10.1093/bioinformatics/17.4.349.
18. Dong, Q.; Zhou, S.; Guan, J. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* **2009**, *25*, 2655-2662, https://doi.org/10.1093/bioinformatics/btp500.
19. Taguchi, Y.h.; Gromiha, M.M. Application of amino acid occurrence for discriminating different folding

types of globular proteins. *BMC Bioinformatics* **2007**, *8*, 404, https://doi.org/10.1186/1471-2105-8-404.

20. Fox, N.K.; Brenner, S.E.; Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2014**, *42*, D304-D309, https://doi.org/10.1093/nar/gkt1240.
21. Chandonia, J.-M.; Fox, N.K.; Brenner, S.E. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res.* **2019**, *47*, D475-D481, https://doi.org/10.1093/nar/gky1134.
22. Bankapur, S.; Patil, N. An Enhanced Protein Fold Recognition for Low Similarity Datasets using Convolutional and Skip-Gram Features with Deep Neural Network. *IEEE Trans. NanoBiosci.* **2020**, https://doi.org/10.1109/TNB.2020.3022456.
23. Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536-540, https://doi.org/10.1016/S0022-2836(05)80134-2.
24. Anfinsen, C.B.; Haber, E.; Sela, M.; White, F.H. The Kinetics Of Formation Of Native Ribonuclease During Oxidation Of The Reduced Polypeptide Chain. *Proceedings of the National Academy of Sciences* **1961**, *47*, 1309, http://dx.doi.org/10.1073/pnas.47.9.1309.
25. Petegrosso, R.; Li, Z.; Srour, M.A.; Saad, Y.; Zhang, W.; Kuang, R. Scalable remote homology detection and fold recognition in massive protein networks. *Proteins: Structure, Function, and Bioinformatics* **2019**, *87*, 478-491, https://doi.org/10.1002/prot.25669.
26. Mishra, S.; Looger, L.L.; Porter, L.L. Inaccurate secondary structure predictions often indicate protein fold switching. *Protein Sci.* **2019**, *28*, 1487-1493, https://doi.org/10.1002/pro.3664.
27. Komal, P.; Usha, C. Relevance of Machine Learning Techniques and Various Protein Features in Protein Fold Classification: A Review. *Curr. Bioinform.* **2019**, *14*, 688-697, https://doi.org/10.2174/1574893614666190204154038.
28. Xiaoyu, T.; Daozheng, C.; Jun, G. An Overview on Protein Fold Classification via Machine Learning Approach. *Curr. Proteomics* **2018**, *15*, 85-98, https://doi.org/10.2174/1570164614666171030160312.
29. Stapor, K.; Roterman-Konieczna, I.; Fabian, P. Machine Learning Methods for the Protein Fold Recognition Problem. In *Machine Learning Paradigms*, Springer: 2019; 101-127, https://doi.org/10.1007/978-3-319-94030-4_5.
30. Ju, Z.; Wang, S.-Y. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. *Gene* **2018**, *664*, 78-83, https://doi.org/10.1016/j.gene.2018.04.055.
31. Polat, Ö.; Dokur, Z. Protein fold classification with Grow-and-Learn network. *Turkish Journal of Electrical Engineering & Computer Sciences* **2017**, *25*, 1184-1196, https://doi.org/10.3906/elk-1506-126.
32. Shen, H.-B.; Chou, K.-C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics* **2006**, *22*, 1717-1722, https://doi.org/10.1093/bioinformatics/btl170.
33. Arif, M.; Hayat, M.; Jan, Z. iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition. *J. Theor. Biol.* **2018**, *442*, 11-21, https://doi.org/10.1016/j.jtbi.2018.01.008.
34. Ibrahim, W.; Saniee, M. Protein fold recognition using Deep Kernelized Extreme Learning Machine and linear discriminant analysis. *Neural Comput. Appl.* **2018**, *0123456789*, https://doi.org/10.1007/s00521-018-3346-z.
35. Yan, K.; Xu, Y.; Fang, X.; Zheng, C.; Liu, B. Protein fold recognition based on sparse representation based classification. *Artif. Intell. Med.* **2017**, *79*, 1-8, https://doi.org/10.1016/j.artmed.2017.03.006.
36. Chen, D.; Tian, X.; Zhou, B.; Gao, J. Profold: Protein fold classification with additional structural features and a novel ensemble classifier. *BioMed research international* **2016**, *2016*, https://doi.org/10.1155/2016/6802832.
37. Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S.-H. Recognition of a protein fold in the context of the SCOP classification. *Proteins: Structure, Function, and Bioinformatics* **1999**, *35*, 401-407, https://doi.org/10.1002/(SICI)1097-0134(19990601)35:4<401::AID-PROT3>3.0.CO;2-K.
38. Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.H. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences* **1995**, *92*, 8700,, https://doi.org/10.1073/pnas.92.19.8700.
39. Ghouzam, Y.; Postic, G.; Guerin, P.-E.; de Brevern, A.G.; Gelly, J.-C. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Sci. Rep.* **2016**, *6*, 28268, https://doi.org/10.1038/srep28268.
40. Li, J.; Wu, J.; Chen, K. PFP-RFSM: protein fold prediction by using random forests and sequence motifs. *J. Biomed. Sci. Eng.* **2013**, *6*, 1161, http://doi.org/10.4236/jbise.2013.612145.
41. Rose, G.D.; Geselowitz, A.R.; Lesser, G.J.; Lee, R.H.; Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. *Science* **1985**, *229*, 834, http://doi.org/10.1126/science.4023714.
42. Saini, H.; Raicar, G.; Lal, S.P.; Dehzangi, A.; Imoto, S.; Sharma, A. Protein fold recognition using genetic algorithm optimized voting scheme and profile bigram. *Journal of Software* **2016**, *11*, 756-767, https://doi.org/10.17706/jsw.11.8.756-767.
43. Wang, J.; Yang, B.; Revote, J.; Leier, A.; Marquez-Lago, T.T.; Webb, G.; Song, J.; Chou, K.-C.; Lithgow,

T. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* **2017**, *33*, 2756-2758, https://doi.org/10.1093/bioinformatics/btx302.

44. Liu, B.; Li, C.-C.; Yan, K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* **2020**, *21*, 1733-1741, https://doi.org/10.1093/bib/bbz098.
45. Sharma, A.; Lyons, J.; Dehzangi, A.; Paliwal, K.K. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *J. Theor. Biol.* **2013**, *320*, 41-46, https://doi.org/10.1016/j.jtbi.2012.12.008.
46. Paliwal, K.K.; Sharma, A.; Lyons, J.; Dehzangi*, A. A Tri-Gram Based Feature Extraction Technique Using Linear Probabilities of Position Specific Scoring Matrix for Protein Fold Recognition. *IEEE Trans. NanoBiosci.* **2014**, *13*, 44-50, https://doi.org/10.1109/tnb.2013.2296050.
47. Lyons, J.; Dehzangi*, A.; Heffernan, R.; Yang*, Y.; Zhou, Y.; Sharma, A.; Paliwal*, K. Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models. *IEEE Trans. NanoBiosci.* **2015**, *14*, 761-772, https://doi.org/10.1109/tnb.2015.2457906.
48. McGuffin, L.J.; Bryson, K.; Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404-405, https://doi.org/10.1093/bioinformatics/16.4.404.
49. Buchan, D.W.A.; Minneci, F.; Nugent, T.C.O.; Bryson, K.; Jones, D.T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **2013**, *41*, W349-W357, https://doi.org/10.1093/nar/gkt381.
50. van Ijzendoorn, D.G.P.; Szuhai, K.; Briaire-de Bruijn, I.H.; Kostine, M.; Kuijjer, M.L.; Bovée, J.V.M.G. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comp. Biol.* **2019**, *15*, e1006826, https://doi.org/10.1371/journal.pcbi.1006826.
51. Hossain, M.A.; Saiful Islam, S.M.; Quinn, J.M.W.; Huq, F.; Moni, M.A. Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *J. Biomed. Inf.* **2019**, *100*, 103313, https://doi.org/10.1016/j.jbi.2019.103313.
52. Tang, B.; Pan, Z.; Yin, K.; Khateeb, A. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in genetics* **2019**, *10*, 214, https://doi.org/10.3389/fgene.2019.00214.
53. L Larrañaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J.A.; Armañanzas, R.; Santafé, G.; Pérez, A.; Robles, V. Machine learning in bioinformatics. *Brief. Bioinform.* **2006**, *7*, 86-112, https://doi.org/10.1093/bib/bbk007.
54. Mehta, A.A.; Mazumdar, H.S. Discovery of significant mirna-biomarkers for breast cancer using decision tree classifier. *Int. J. Emerg. Technol.* **2020**, *11*, 453–460.
55. Dietterich, T.G. *Ensemble Methods in Machine Learning. In: Multiple Classifier Systems*; 2000; 1-15, http://dx.doi.org/10.1007/3-540-45014-9_1.
56. Salo, F.; Nassif, A.B.; Essex, A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer Networks* **2019**, *148*, 164-175, https://doi.org/10.1016/j.comnet.2018.11.010.
57. Qiu, W.-R.; Sun, B.-Q.; Xiao, X.; Xu, Z.-C.; Jia, J.-H.; Chou, K.-C. iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. *Genomics* **2018**, *110*, 239-246, https://doi.org/10.1016/j.ygeno.2017.10.008.
58. Peurifoy, J.; Shen, Y.; Jing, L.; Yang, Y.; Cano-Renteria, F.; DeLacy, B.G.; Joannopoulos, J.D.; Tegmark, M.; Soljačić, M. Nanophotonic particle simulation and inverse design using artificial neural networks. *Science Advances* **2018**, *4*, eaar4206, http://doi.org/10.1126/sciadv.aar4206.
59. Zador, A.M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications* **2019**, *10*, 3770, http://dx.doi.org/10.1038/s41467-019-11786-6.
60. Ben-Hur, A.; Ong, C.S.; Sonnenburg, S.; Schölkopf, B.; Rätsch, G. Support Vector Machines and Kernels for Computational Biology. *PLoS Comp. Biol.* **2008**, *4*, e1000173, https://doi.org/10.1371/journal.pcbi.1000173.
61. Alpaydin, E. GAL: Networks That Grow When They Learn And Shrink When They Forget. *International Journal of Pattern Recognition and Artificial Intelligence* **1994**, *08*, 391-414, https://doi.org/10.1142/S021800149400019X.
62. Jo, T.; Hou, J.; Eickholt, J.; Cheng, J. Improving Protein Fold Recognition by Deep Learning Networks. *Sci. Rep.* **2015**, *5*, 17573, https://doi.org/10.1038/srep17573.
63. Kavousi, K.; Moshiri, B.; Sadeghi, M.; Araabi, B.N.; Moosavi-Movahedi, A.A. A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM. *Comput. Biol. Chem.* **2011**, *35*, 1-9, https://doi.org/10.1016/j.compbiolchem.2010.12.001.
64. Tripathi, P.; Pandey, P.N. A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. *J. Theor. Biol.* **2017**, *424*, 49-54, https://doi.org/10.1016/j.jtbi.2017.04.027.
65. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, 2000; https://doi.org/10.1017/CBO9780511801389.
66. Kavzoglu, T.; Colkesen, I. A kernel functions analysis for support vector machines for land cover

classification. *International Journal of Applied Earth Observation and Geoinformation* **2009**, *11*, 352-359, https://doi.org/10.1016/j.jag.2009.06.002.

67. Paliwal, K.K.; Sharma, A.; Lyons, J.; Dehzangi, A. Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information. *BMC Bioinformatics* **2014**, *15*, S12, http://dx.doi.org/10.1186/1471-2105-15-S16-S12.
68. Ghanty, P.; Pal, N.R. Prediction of Protein Folds: Extraction of New Features, Dimensionality Reduction, and Fusion of Heterogeneous Classifiers. *IEEE Trans. NanoBiosci.* **2009**, *8*, 100-110, https://doi.org/10.1109/TNB.2009.2016488.
69. Wei, L.; Liao*, M.; Gao, X.; Zou, Q. Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique. *IEEE Trans. NanoBiosci.* **2015**, *14*, 649-659, https://doi.org/10.1109/TNB.2015.2450233.