Amino acid encoding methods for protein sequences: a comprehensive review and assessment

Xiaoyang Jing, Qiwen Dong, Daocheng Hong, Ruqian Lu

Abstract—As the first step of machine-learning based protein structure and function prediction, the amino acid encoding play a fundamental role in the final success of those methods. Different with the protein sequence encoding, the amino acid encoding can be used in both residue-level and sequence-level prediction of protein properties by combining with different algorithms. However, it does not attract enough attention in the past decades, and there are no comprehensive reviews and assessments about encoding methods so far. In this article, we make a systematic classification and propose a comprehensive review and assessment for various amino acid encoding methods. Those methods are grouped into five categories according to their information sources and information extraction methodologies, including binary encoding, physicochemical properties encoding, evolution-based encoding, structure-based encoding, and machine-learning encoding. And then sixteen representative methods from five categories are selected and compared on protein secondary structure prediction and protein fold recognition tasks by using large-scale benchmark datasets. The results show that the evolution-based position-dependent encoding method PSSM achieve the best performance, and the structure-based and machine-learning encoding methods also show some potential for further application, the neural network based distributed representation of amino acids in particular may bring new light to this area. We hope that the review and assessment are useful for future studies in amino acid encoding.

Index Terms—amino acid encoding, feature extraction, residue encoding, protein structure and function prediction, protein secondary structure prediction, protein fold recognition.

1 Introduction

The Anfinsen's thermodynamic hypothesis [1] makes it possible to predict protein structure and function from its sequence. Over the past decades, a series of remarkable improvements in protein sequence-based structure and function prediction have been proposed, such as protein secondary structure prediction [2], protein remote homology detection [3], inter-residue contact prediction [4], protein-binding residues prediction [5], etc. Most of these improvements are achieved by using data-driven machine learning methods, especially by using deep learning methods [6]. A successful machine learning based method is usually based on three fundamental factors: adequate training samples, powerful machine learning algorithms, and effective feature representations. In the last few years, the improvements achieved in protein structure and function prediction mainly benefits from the rapid expansion of biological data and the prosperity of machine learning algorithms, such as deep learning algorithms [2]. In comparison, the feature representation methods of amino acid (or amino acid encoding method) has gained limited attention, and hence little progress has been made. In a time when the dividends of data and algorithm have been highly released, paying more attention to amino acid encoding methods may be one of the practical ways to achieve further improvements in protein structure and function studies.

In general, protein sequences are represented by using twenty letters of amino acid alphabet, while such representation cannot be directly processed before it is converted to digital representation. Obtaining the digital representation of amino acid is the first step of machine-learning based protein structure and function prediction methods, and effective digital representation is crucial to the final success of these methods [7]. The digital representation of amino acid is usually called feature extraction, amino acid encoding scheme, or residue encoding scheme, etc. In this article, we use amino acid encoding as the common name. It should be noted that the amino acid encoding is different with the protein sequence encoding. The protein sequence encoding represents the entire protein sequence by using an n-dimensional vector, such as the n-gram [8], pseudo amino acid composition [9], [10], etc. Since the amino acid-specific information is lost, the protein sequence encoding can be only used to predict sequence-level properties (i.e. protein fold recognition). The amino acid encoding represents each amino acid of a protein sequence by using different n-dimensional vectors, thus its vector space for a protein sequence is $n * L$ (L is the length of protein sequence). By combining with different machine learning methods, the

- *Xiaoyang Jing is with the Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University, 220 Handan Road, Shanhai 200433, China. E-mail: xyjing14@fudan.edu.cn*
- *Qiwen Dong is with the School of Data Science and Engineering, East China Normal University, 3663 Zhongshan road, Shanghai 200062, China. E-mail: qwdong@dase.ecnu.edu.cn*
- *Daocheng Hong is with the School of Data Science and Engineering, East China Normal University, 3663 Zhongshan road, Shanghai 200062, China. E-mail: hongdc@dase.ecnu.edu.cn*
- *Ruqian Lu is with the Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University, 220 Handan Road, Shanghai 200433, China. E-mail: rqlu@fudan.edu.cn*

Manuscript received XXXX, 2018; revised XXXX, 2018.

amino acid encoding can be used in protein properties prediction both at residue-level and sequence-level (i.e. protein fold recognition, secondary structure prediction, etc). Here, we are concerned with the amino acid encoding methods. Furthermore, some predicted amino acid properties are used in other properties prediction tasks, for example, the predicted secondary structure and solvent accessibility are used to recognize protein fold types [11]. These properties are also seen as encodings in some literatures, however, they are predicted by the initial amino acid encodings, such as one-hot encodings, physic-chemical properties encodings, etc. Therefore the amino acid encoding discussed in this article refers to the initial amino acid encodings rather than the predicted properties. In the past decades, various amino acid encoding methods have been proposed from different perspectives [12], [13], [14]. The most widely used encodings are the one-hot encoding, the position specific scoring matrix (PSSM) encoding, and some physic-chemical properties encodings. In addition to those encodings, some other encodings have also proposed, such as the encoding estimated from inter-residue contact energies [15], the encoding learned from protein structure alignments [16], and the encoding learned from sequence context [17]. These encoding methods explore the amino acid encoding from new perspectives, and can be the complement of above encodings. Kawashima et al. [18] have proposed a database of numerical indices of amino acids and amino acid pairs, which contains physicochemical and biochemical properties of amino acids. But comprehensive reviews and assessments of amino acid encoding methods are still missing in this area, and there is no systematic classification of amino acid encoding methods until now.

In order to propose a comprehensive review for amino acid encoding methods, we first group these methods into five categories according to their information sources and information extraction methodologies, including binary encoding, physicochemical properties encoding, evolution-based encoding, structure-based encoding, and machine-learning encoding. The structure of this article is as follows. It begins with a review about various amino acid encoding methods of the five categories. Then, a theoretical discussion and analysis of these methods is made. Third, the performances of sixteen representative amino acid encoding methods are compared and discussed based on two tasks: protein secondary structure prediction and protein fold recognition. And finally, the conclusion and future perspective are proposed.

2 AMINO ACID ENCODING METHODS

2.1 Binary encoding

The binary encoding methods use multidimensional binary digits (0 and 1) to represent amino acids in protein sequences. The most commonly used binary encoding is the one-hot encoding, which is also called orthogonal encoding [13]. For the one-hot encoding, each amino acid type of twenty standard amino acids is represented by a twenty dimensional binary vector. Specifically, the twenty standard amino acids are fixed in a specific order, and then the *i*th amino acid type is represented by twenty binary bits with the *i*th bit set to "1" and others to "0". There is only one bit equal to "1" for each vector, hence it is called "one-hot". For example, the twenty standard amino acids are sorted as [A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y], the one-hot code of A is 100000000000000000000, C is 01000000000000000000, and so on. Since protein sequences may contain some unknown amino acids, it should be noted that one more bit is needed to represent the unknown amino acid type in some cases, and the dimension of its binary vector will be twenty-one [19].

Because the one-hot encoding is a high-dimensional and sparse vector representation, there is a simplified binary encoding method based on conservative replacements through evolution [20]. Deriving from the point accepted mutation (PAM) matrices [21], the twenty standard amino acids are divided into six groups: [H, R, K], [D, E, N, Q], [C], [S, T, P, A, G], [M, I, L, V], and [F, Y, W]. And six dimensional binary vectors are used to represent amino acids based on their groups. Another low-dimensional binary encoding scheme is the binary 5-bit encoding introduced by White and Seffens [22]. Theoretically, the binary 5-bit code could represent 32 ($2^5 = 32$) possible amino acid types. In order to represent the twenty standard amino acids, the all zeros encoding, the all ones encoding and those encodings with 1 or 4 ones ($5 + 5 = 10$) are removed, finally 20 encodings ($32 - 1 - 1 - 10 = 20$) are left in total. This binary 5-bit encoding use 5 dimension binary vector to take the place of 20 dimension vector of one-hot encoding, which may lead to less model complexity [13].

2.2 Physicochemical properties encoding

From the perspective of molecular composition, a typical amino acid generally contains a central carbon atom (C) which is attached with an amino group (NH_2), a hydrogen atom (H), a carboxyl group (COOH) and a side chain (R), as shown in Fig. 1. The side chains (R) are usually carbon chains or rings (except the $Proline$) which are attached to various functional groups [13]. The physicochemical properties of those components play critical roles in the formation of protein structures and functions, thus these properties can also be used as features for protein structure and function prediction [7].

Among various physicochemical properties, the hydrophobicity of amino acid is believed to play a fundamental role in organizing the self-assembly of protein [23]. Based on the propensity of the amino acid side chain to be in contact with polar solvent like water, the twenty amino acids can be classified as either hydrophobic or hydrophilic. The free energy of amino acid side chains transferring from cyclohexane to water can be used to quantificationally represent its hydrophobicity [14]. If the free energy is positive value, the amino acid is hydrophobic, and the amino acid with negative value is hydrophilic. The hydrophobic amino acids are usually buried inside the protein core in protein three-dimensional structures, while the hydrophilic amino acids preferentially cover the surface of the protein three-dimensional structures. Furthermore, the hydrophilic amino acids are called as polar amino acids. At typical biological environment, some polar amino acids carry a charge: $Lysine$ (+), $Histidine$ (+), $Arginine$ (+), $Aspartate$ (-) and $Glutamate$ (-), while other polar amino acids, $Asparagine$,

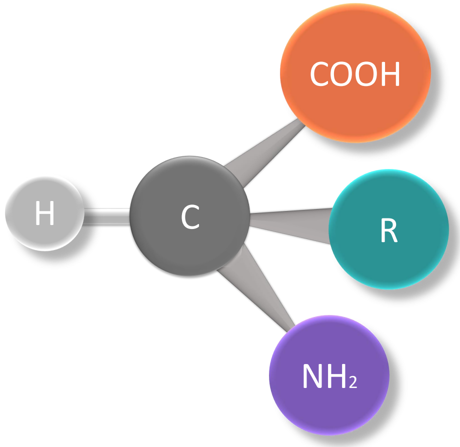


Fig. 1. The schematic of typical amino acid molecular components. A typical amino acid contains a central carbon atom (C) which is attached with an amino group (NH_2), a hydrogen atom (H), a carboxyl group (COOH) and a side chain (R).

Glutamine, *Serine*, *Threonine* and *Tyrosine*, are neutral [24]. A detail classification of the hydrophobicity properties of twenty standard acid sides is shown in Table 1. Other than hydrophobicity properties, the codon diversity and size of amino acids are also used as features. The codon diversity of an amino acid is reflected by the number of codons coding for the amino acid, and the size of an amino acid denotes its molecular volume [7].

TABLE 1
The hydrophobicity properties of twenty standard acid sides

Hydrophobicity	Amino acids	3-letter abbreviation	1-letter abbreviation
Hydrophobic	Alanine	Ala	A
	Isoleucine	Ile	I
	Leucine	Leu	L
	Methionine	Met	M
	Phenylalanine	Phe	F
	Valine	Val	V
	Proline	Pro	P
	Glycine	Gly	G
Charged	Lysine (+)	Lys	K
	Histidine (+)	His	H
	Arginine (+)	Arg	R
	Aspartic (-)	Asp	D
	Glutamic (-)	Glu	E
Polar	Glutamine	Gln	Q
	Asparagine	Asn	N
	Serine	Ser	S
	Threonine	Thr	T
	Tyrosine	Tyr	Y
	Cysteine	Cys	C
	Tryptophan	Trp	W

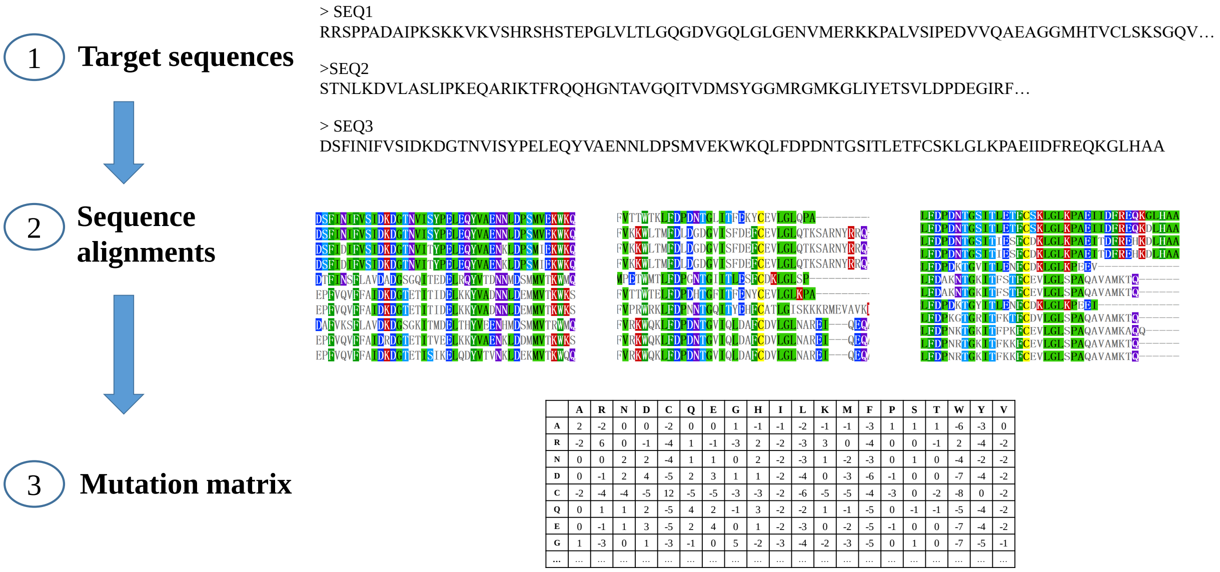
Some physicochemical properties based amino acid encodings have been proposed in previous studies. Fauchre et al. [25] established fifteen physicochemical descriptors of side chains for 20 natural and of 26 non-coded amino acids which reflect hydrophobic, steric, electronic, and other properties of amino acid side chains. Radzicka and Wolfenden [26] obtained the digitized indications of the tendencies of amino acids to leave water and enter a truly nonpolar condensed phase by their experiments. Lohman et al. [27] represented the amino acids by using seven physicochemical properties to predict transmembrane protein sequences: hydrophobicity, hydrophilicity, polarity, volume, surface area, bulkiness, and refractivity. Atchley et al. [7] used the multivariate statistical analyses to produce multidimensional patterns of attribute co-variation for twenty standard amino acids, which reflect the polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge of amino acids.

2.3 Evolution-based encoding

The evolution-based encoding methods extract evolutionary information of residues from sequence alignments or phylogenetic trees to represent amino acids, mainly by using the amino acids substitution probability. These evolution-based encoding methods can be categorized into two groups based on the position relevance: position-independent methods and position-dependent methods.

The position-independent methods encode amino acids by using fixed encodings, regardless of the amino acid position in the sequence and the amino acid composition of the sequence. The most commonly used position-independent encoding are the PAM matrices and the BLOSUM matrices, the common flowchart of them is shown in Fig. 2. The point accepted mutation (PAM) matrices represent the replacement probabilities from a single amino acid to another single amino acid in homologous protein sequences [21], which is focused on the evolutionary process of proteins. The PAM matrices are calculated from protein phylogenetic trees and related protein sequence pairs. The assumption of the PAM matrices is the accepted mutation is similar in physical and chemical properties to the old one and the likelihood of amino acid X replacing Y is the same as that of Y replacing X, thus the PAM matrices are 20*20 symmetry matrices where each row and column represents one of the twenty standard amino acids. Corresponding to different lengths of evolution time, different PAM matrices can be generated. The 250 PAMs, which means the amino acid replacements to be found after 250 evolutionary changes, was found to be an effective scoring matrix for detecting distant relationships by the authors [21], and is widely used in related researches [28], [29]. The blocks amino acid substitution matrices (BLOSUM) [30] are amino acid substitution matrices derived based on conserved regions constructed by the PROTOMAT [31] from non-redundant protein groups. The values in the BLOSUM matrixes represent the probabilities that amino acid pairs will interchange with each other. To reduce the contributions of most closely related protein sequences, the sequences are clustered within blocks. Different BLOSUM matrices can be generated by using different identical percentages for clusters, and the BLOSUM 62 matrix performed better overall [30].

Different with the position-independent matrices, the position-dependent methods encode amino acids at different position by using different encodings, even if the amino acid types are the same. The position-dependent encodings are deduced from the multiple sequence alignments (MSAs) of target sequence, the flowchart is shown in Fig. S1. The position specific scoring matrix (PSSM) is the most widely used encoding method. The PSSM is also called position weight


1 Target sequences
> SEQ1
RRSPPADAIPKSKKVKVSHRSHSTEPGLVLTLGQGDVGQLGLGENVMERKKPALVSIPEDVVQAEAGGMHTVCLSKSGQV…
>SEQ2
STNLKDVLASLIPKEQARIKTFRQQHGNTAVGQITVDMSYGGMRGMKGLIYETSVLDPDEGIRF…
> SEQ3
DSFINIFVSIDKDGTNVISYPELEQYVAENNLDPSMVEKWKQLFDPDNTGSITLETFCSKLGLKPAEIIDFREQKGLHAA
2 Sequence alignments
3 Mutation matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Fig. 2. The flowchart of position-independent amino acid encoding methods. First, the target proteins are selected (step 1). Then, the sequence alignments are constructed based on some criteria (step 2). Finally, the mutation matrix is calculated and is regarded as the amino acid encodings (step 3).

matrix (PWM), which represents the log-likelihoods of the occurrence probabilities of all possible molecule types at each location in a given biological sequence [32]. Generally, the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [33] is used to execute sequence alignment and generate MSA for the target protein sequence. And then the corresponding PSSM is calculated from the MSA. For a protein sequence with length L, its PSSM is an $L * 20$ matrix, in which each row represent the log-likelihoods of the probabilities of twenty amino acids occur at its corresponding position. Besides the PSI-BLAST, the HMM-HMM alignment algorithm HHblits is also widely used to generate the probabilities profile, which is more sensitive than the sequence-profile alignment algorithm PSI-BLAST demonstrated by Remmert et al. [34].

2.4 Structure-based encoding

The structure-based amino acid encoding methods, which can also be called statistical-based methods, encode amino acids by using structure-related statistical potentials, mainly using the inter-residue contact energies [35]. The diagram of protein inter-residue contacts is shown in Fig. 3. The basic assumption is, in a large number of protein native structures, the average potentials of inter-residue contacts can reflect the differences of interaction between residue pairs [36], which play an important role in the formation of protein backbone structures [35]. The inter-residue contact energies of twenty amino acids are usually estimated based on amino acid pairing frequencies from protein native structures [35]. The typical procedure to calculate the contact energies contains three steps. First, a protein structure set is constructed from known protein native structures. Then, the inter-residue contacts of twenty amino acids observed in those structures are counted. Finally, the contact energies are deduced from the amino acid contact frequencies by using the predefined energy function, and different contact energies reflect different contact potentials of amino acids in native structures.

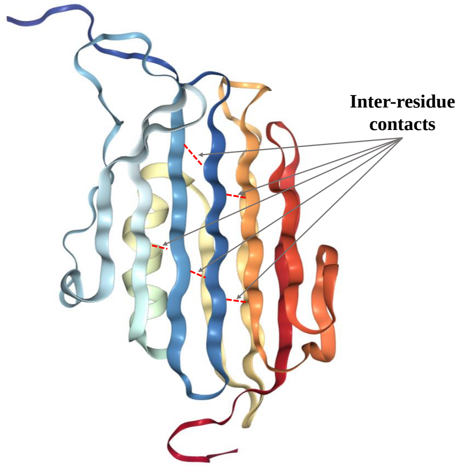


Fig. 3. The schematic diagram of protein inter-residue contacts. The dotted red lines in the figure represent the inter-residue contacts in protein native structure.

Many previous studies have focused on structure-based encodings. In order to account the medium and long range interactions which determine the protein folding conformations, Tanaka and Scheraga [35] evaluated the empirical standard free energies to formulate amino acid contacts from the contacts frequency. By employing the lattice model, Miyazawa and Jernigan [36] estimated contact energies by means of the quasi-chemical approximation with an approximate treatment of the effects of chain connectivity. Later, they re-evaluated the contact energies based on a larger set of protein structures and also estimated an additional

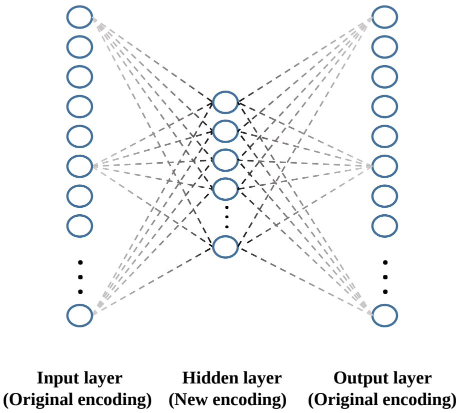


Fig. 4. The typical architecture of machine-learning based amino acid encoding method. The input and output of the neural network are origin amino acid encodings, and the value of hidden layer represents the new encoding of corresponding amino acid.

repulsive packing energy term to provide an estimate of the overall energies of inter-residue interactions [37]. To investigate the validity of the quasi-chemical approximation, Skolnick et al. [38] estimated the expected number of contacts by using two reference states, the first of which treats the protein as a Gaussian random coil polymer and the second includes the effects of chain connectivity, secondary structure and chain compactness. The comparison result show that the quasi-chemical approximation is in general sufficient for extracting the amino acids pair potentials. To recognize native-like protein structures, Simons et al. [39] used distance-dependent statistical contact potentials to develop energy functions. Zhang and Kim [40] estimated sixty residue contact energies which mainly reflect the hydrophobic interactions and show strong dependence on the three secondary structural states. These energies were effective in threading and three-dimensional contact prediction by their tests. Later, Cristian et al. set up an iterative scheme to extract the optimal interaction potentials between the amino acids. [41].

2.5 Machine-learning encoding

Different from above manually defined encoding methods, the machine-learning based encoding methods learn the amino acid encodings from protein sequence or structure data by using machine learning methods, typically using artificial neural networks. In order to reduce the complexity of the model, the neural network for learning amino acids encodings are weight sharing for twenty amino acids. In general, as shown in Fig. 4, the neural network contains three layers: the input layer, the hidden layer, and the output layer. The input layer corresponds with the original encoding of the target amino acid, which can be one-hot encoding, physicochemical encoding, etc. The output layer also corresponds with the original encoding of the related amino acids. The hidden layer, which represents the new encoding of the target amino acid, usually have a reduced dimension compared with the original encoding.

To our knowledge, the earliest concept of learning-based amino acids encodings was proposed by Riis and Krogh [42]. In order to reduce the redundancy of one-hot encoding, they used a $20{*}3$ weight sharing neural network to learn a 3-dimensional real numbers representation of 20 amino acids from the one-hot encoding. Later, Jagla and Schuchhardt [43] also apply the weight sharing artificial neural network to learn a 2-dimensional encoding of amino acids for human signal peptide cleavage sites recognition. Meiler et al. [44] used a symmetric neural network to learn reduced representations of amino acids from amino acid physicochemical and statistical properties. The parameter representations were reduced from five and seven dimensions, respectively, to 1, 2, 3, or 4 dimensions, and then these reduced representations were used to ab initio prediction of protein secondary structure. Lin et al. [16] used artificial neural network to derive encoding schemes of amino acids from protein three-dimensional structure alignments, and each amino acid is described using the values taken from the hidden units of neural network.

In recent years, several new machine-learning based encoding methods [17], [45], [46] have been proposed with reference to distributed word representation in natural language processing. In natural language processing, the distributed representation of words has been proved to be an effective strategy for use in many tasks [47]. The basic assumption is that the words sharing similar contexts will have similar meanings, therefore these methods train the neural network model by using the target word to predict its context words or by predicting the target word from its context words. After trained on unlabeled datasets, the weights of the hidden units for each word is used as its distributed representation. In protein related studies, the similar strategy has been used by assuming that: the protein sequence are sentences, and the amino acids or sub-sequences are words. In previous researches, these distributed representations of amino acids or sub-sequences show potential in protein family classification and disordered protein identification [17], protein function site predictions [45], and protein functional properties prediction [46], etc.

3 Discussions

In this section, we will make a theoretical discussion of amino acid encoding methods. First of all, we investigate the classification criteria of amino acid encoding methods, secondly, discuss the theoretical basis of these methods, and then analyze their advantages and limitations. Finally, we review and discuss the criteria for measuring an amino acid encoding method.

As introduced above, amino acid encoding methods have been divided into five categories according to their information sources and methodologies. However, it should be noted that the methods in one category are not completely different from the others, and there are some fusions between the encoding methods from different categories. For example, the 6-bit one-hot encoding proposed by Wang et al. [20] is a dimension-reduced representation of the common one-hot encoding, but it is based on the six amino acid exchange groups which are derived from PAM matrices [21]. There is another classification criterion based on the position relevance. In the above section, the evolution-based encoding methods have been divided into two categories: position-independent methods and position-dependent methods. We can also group all of the amino acid encoding methods into position-independent category and position-dependent category. Except the position specific scoring matrix (PSSM) and other similar encodings extracting evolution features from multiple sequence alignments are position-dependent methods, all other amino acid encoding methods are position-independent methods. The position-dependent methods can capture the homologous information, while the position-independent methods can reflect the basic properties of amino acid, to some extent these two kinds of methods can be complementary with each other. In practice, the combination of position-independent encoding and position-dependent encoding is often used, such as the combination of one-hot and PSSM [11], the combination of physicochemical properties encoding and PSSM [48].

Theoretically, the functions of a protein are closely related to its tertiary structure, and its tertiary structure is mostly determined by the physicochemical properties of its amino acid sequence [1]. From this perspective, all of the evolution-based encoding, structure-based encoding, and machine-learning encoding methods extract information based on the physicochemical properties of amino acid by using difference strategies. Specifically, different amino acids may have different mutation tendencies in the evolutionary process due to their hydrophobicity, polarity, volume, and other properties. These mutation tendencies will be reflected in the sequence alignments and are detected by the evolution-based encoding methods. Similarly, the physicochemical properties of amino acid could affect the inter-residue contact potentials in protein tertiary structure which is the basis of the structure-based encoding methods. And the machine-learning encoding methods also learn amino acid encoding from its physicochemical representation or evolution information (such as homologous protein structure alignments), which can be seen as another variant of physicochemical properties. In spite of these encoding methods share similar theoretical basis, their performance is different due to their restrictions in implementation. To the one-hot encoding method, there is no artificial correlation between amino acids, but it is highly sparse and redundant, which will lead to a complex machine learning model. The physicochemical properties of amino acids play fundamental roles in protein folding process, theoretically the physicochemical properties encoding methods should be effective. However, as the protein folding-related physicochemical properties and their digital metrics are unknown, developing effective physicochemical properties encoding method is still an open problem. The evolution-based encoding methods extract evolution information just from protein sequences, which could enjoy the dividends of large-scale protein sequence data. Especially the PSSM has achieved significant performance in many researches [3]. But for those proteins without homologous sequences, the performances of evolution-based methods are limited. The structure-based encoding methods encode amino acids based on the potential of inter-residue contact which is a low-dimensional representation of protein structure. Because of the limited number of known protein structures, their performances are limited. The early machine-learning encoding methods also face the problem of insufficient data samples, while several methods developed recently overcome this problem by taking advantage of unlabeled sequence data [17], [45], [46].

As discussed above, different amino acid encoding methods have specific advantages and limitations, what is the effective encoding method? Wang et al. [20] defined that the best encoding should highly reduce the uncertainty of the output of prediction model, or the encoding could capture both the global similarity and the local similarity of protein sequences, where the global similarity refers to the overall similarity among multiple sequences, and the local similarity refers to motifs in the sequences. Riis and Krogh [42] proposed that the redundancy encodings will lead the prediction model to be over-fitting, and need to be simplified. Meiler et al. [44] also tried to use reduced representations of amino acids physicochemical and statistical properties for protein secondary structure prediction. Zamani and Kremer [12] stated that an effective encoding must preserve information relative to the problem at hand, while diminishing superfluous data. By summing up previous studies, an effective amino acid encoding should be information-rich and non-redundant. The information-rich property means the encoding contains enough information that is highly relevant to the protein structure and function, such as the physicochemical properties, evolution information, contact potential, and so on. The non-redundant property means the encoding is compact and does not contain noise or other unrelated information. For example, to neural network based protein structure and function prediction, a redundancy encoding will lead to complicated networks with a very large number of weights, which leads to over-fitting and restricts the generalization ability of the model. Therefore, under the premise of containing sufficient information, the more compact encoding will be more generative.

Over the past two decades, several studies have been proposed to investigate the effective amino acid encoding methods [13]. David [49] examined the effectiveness of various hydrophobicity scales by using parallel cascade identification algorithm for the structure or function classification of protein sequences. Zhong et al. [50] compared the orthogonal encoding, hydrophobicity encoding, BLOSUM62 encoding and PSSM encoding utilizing the denoeux belief neural network for protein secondary structure prediction. Hu et al. [14] combined the orthogonal encoding, hydrophobicity encoding and BLOSUM62 encoding to find the most optimal encoding scheme by using the SVM with sliding window training scheme for protein secondary structure prediction. According to their test results, the combination of orthogonal and BLOSUM62 matrix showed the highest accuracy compared with all other encoding schemes. Zamani and Kremer [12] investigated the efficiency of fifteen amino acid encoding schemes, which contain orthogonal encoding, physicochemical encoding, secondary structures and BLOSUM62 related encoding, by training artificial neural networks to approximate the substitution matrices. Their experimental results indicate that the number (dimension) and the types (properties) of amino acid encodings methods are the two key factors for the efficiency of the encoding performance. Dongardive and Abraham [51] compared the orthogonal, hydrophobicity, BLOSUM62, PAM250 and hybrid encoding schemes of amino acid for protein secondary structure prediction and found the best performance was achieved by the BLOSUM62 matrix. These previous studies explore amino acid encoding methods from different perspectives, but they just evaluated a part of encoding methods on small data sets. To present a comprehensive and systematic comparisons, in this article, we performed a large-scale comparative assessment of various amino acid encoding methods based on two tasks: protein secondary structure prediction and protein fold recognition, which is proposed in next sections. It should be noted we are concerned with assessing how much effective information contained in different encoding methods, rather than exploring the optimal combination of encoding methods.

4 The Assessment of Encoding Methods for Protein Secondary Structure Prediction

In computational biology, protein sequence labeling tasks, such as protein secondary structure prediction, solvent accessibility prediction, disorder region prediction, and torsion angle prediction, have gained a great deal of attention from researchers. Among those sequence labeling tasks, protein secondary structure prediction is the most representative task [2], and several previous amino acid encoding studies also paid attention to this topic [14], [42], [50], [51]. Therefore, we first perform the assessment of various amino acid encoding methods based on the protein secondary structure prediction task.

4.1 Encoding methods selection and generation

To perform a comprehensive assessment of different amino acid encoding methods, we select sixteen representative encoding methods from each category for evaluation. The brief introduction of the sixteen selected encoding methods is shown in Table 2, and the demonstration of twelve position-independent encodings in 2D space using t-SNE is shown in supplementary materials Fig. S2. Except the PSSM and HMM encodings, most of these encodings are position-independent encodings and can be used directly to encode amino acids, the python script to encode protein sequences is at: https://github.com/xyjing-works/SequenceEncoding. It should be noted that some protein sequences may contain unknown amino acid types, these amino acids will be expressed by the average value of the corresponding column if the original encodings do not deal with this situation. For the ProtVec [17], which is a 3-gram encoding, we encode each amino acid by adding its left and right adjacent amino acid to form the corresponding 3-gram word. Since the start and end amino acids do not have enough adjacent amino acids to form 3-grams, they are represented by the "<unk>" encoding in ProtVec. Recently the further work of ProtVec (ProtVecX [52]) has demonstrated that the concatenation of ProtVec and k-mers could achieve better performance, here we also evaluate the performance of ProtVec concatenated with 3-mers (named as ProtVec-3mer). For position-dependent encoding methods PSSM and HMM, we follow the common practice to generate them. Specifically, for the PSSM encoding of each protein sequence, we ran the PSI-BLAST [33] tool with e-value threshold of 0.001 and three iterations against the UniRef90 [53] sequence database which is filtered at 90% sequence identity. The HMM encoding is extracted from the HMM profile by running the HHblits [34] against the UniProt20 [53] protein database with parameters -n 3 -diff inf -cov 60. According to the HHsuite user guide, we use the first 20 columns of HMM profile and convert the integers in HMM profile to amino acid emission frequencies by using the formula: $h^{fre} = 2^{-0.001*h}$. Where h is the initial integer in HMM profile and h^{fre} is the corresponding amino acid emission frequency. The h is set to 0 if it is an asterisk.

4.2 Benchmark datasets for protein secondary structure prediction

Following several representative protein secondary structure prediction works [19], [48], [54], [55], [56] published recent years, we use the CullPDB dataset [57] as training data and use four widely used test datasets: the CB513 dataset [58], the CASP10 dataset [59], the CASP11 dataset [60] and the CASP12 dataset [61] as test data to evaluate the performance of different features. The CullPDB dataset is a large non-homologous sequence set produced by using the PISCES server [57] which culls subsets of protein sequences from the Protein Data Bank based on sequence identity and structural quality criteria. Here we retrieved a subset of sequences that have structures with better than 1.8 angstroms resolution and share less than 25% sequence identity with each other. We also remove those sequences sharing more than 25% identity with sequences from test dataset to ensure there is no homology between training and test dataset, and finally the CullPDB dataset contains 5748 protein sequences with lengths ranging from 18 to 1455. The CB513 dataset contains 513 proteins with less than 25% sequence similarity. The Critical Assessment of techniques for protein Structure Prediction (CASP) is a highly recognized community experiment to determine the state-of-the-art methods in protein structure prediction from amino acid [61], the recent released CASP 10, CASP 11 and CASP 12 datasets are adopted as test datasets. It should be noted that the protein targets from CASP used here are based on protein domain. Specifically, the CASP10 dataset contains 123 protein domains whose sequence lengths range from 24 to 498, the CASP11 dataset contains 105 protein domains whose sequence lengths range from 34 to 520 and the CASP12 dataset contains 55 protein domains whose sequence lengths range from 55 to 463.

Protein secondary structure labels are inferred by using the DSSP program [62] from corresponding experimentally determined structures. The DSSP specifies 8 secondary structure states to each residue, here we adopt 3-state secondary structure prediction as benchmark task by converting 8 assigned states to 3 states: G, H, and I to H; B and E to E; and S, T, and C to C.

4.3 The performance comparison by using the Random Forests method

In order to use the information of neighboring residues, many previous protein secondary structure prediction methods apply the sliding window scheme and have achieve

TABLE 2
The brief introduction of sixteen selected amino acid encoding methods

Category	Encoding method	Dimension	Description
Binary	One-hot	20	The general one-hot method with one bit to encode one amino acid type.
	One-hot (6-bit)	6	The dimension-reduced one-hot method by using six bit [20].
	Binary 5-bit	5	The binary encoding method by using five binary bit [22].
Physicochemical properties	Hydrophobicity matrix	20	The hydrophobicity matrix based on the hydrophobicity index [26].
	Meiler parameters	7	Seven parameters of physicochemical related properties provided by Meiler et al. [44].
	Acthely factors	5	Five numerical values reflectting various physicochemical properties provided by Atchley et al. [7].
Evolution-based	PAM250	20	The 250 PAM matrix proposed by Dayhoff [21].
	BLOSUM62	20	The BLOSUM 62 matrix proposed by Henikoff and Henikoff [30].
	PSSM	20	The position weight matrix generated by using the PSI-BLAST [33].
	HMM	20	The position weight matrix generated by using the HHblits [34].
Structure-based	Miyazawa energies	20	The inter-residue contact energies estimated by Miyazawa and Jernigan [15].
	Micheletti potentials	20	The residue interaction potentials extracted from coarse-grained description of proteins [41].
Machine-learning	AESNN3	3	An amino acid encoding learned from protein structure alignments [16].
	ANN4D	4	A reduced representation of amino acids learned from physicochemical and statistical properties [44].
	ProtVec	100	A distributed representation of 3-grams amino acids learned from Swiss-Prot sequences [17].
	ProtVec-3mer	163	ProtVec [17] concatenated with 3-mer.

considerable performances [2]. Refer to those methods, we also use the sliding window scheme to evaluate different amino acid encoding methods, the diagram is shown in supplementary materials Fig. S3. The evaluation is based on the random forests method from the Scikit-learn toolboxes [63], the window size is 13 and the number of trees in the forest is 100. The comparison results are shown in Table 3.

First of all, we analyze and discuss the performance of different methods in same category. For the binary encoding methods, the one-hot encoding is the most widely used encoding method. The one-hot (6-bit) encoding and the binary 5-bit encoding are two dimension-reduced representations of the one-hot encoding. In Table 3, the best performance is achieved by one-hot encoding method, which demonstrates that some effective information should be lost after the artificial dimension reduction for one-hot (6-bit) encoding and binary 5-bit encoding. For the physicochemical properties encodings, the hydrophobicity matrix just contains the hydrophobicity related information and performs poorly. While the Meiler parameters and the Acthely factors are constructed from multiple physic-chemical information sources and perform better. This indicates the integration of multiple physic-chemical information is valuable. For the evolution-based encodings, it is obvious that the position-dependent encodings (PSSM and HMM) are much more powerful that those position-independent encodings (PAM250 and BLOSUM62), which shows that the homologous information is strongly associate with the protein structures. For the two structure-based encodings, they have comparative performances. For the three machine-learning encodings, the ANN4D performs better than the AESNN3 and the ProtVec, while the ProtVec-3mer encoding achieves similar performance compared with the ProtVec encoding. Second, on the whole, the position-dependent evolution-based encoding methods (PSSM and HMM) achieved the best performance. This result suggests that the evaluation information extracted from the MSAs is more conserved than that global information extracted from other sources. Third, the performances of different encoding methods show a certain degree of correlation with encoding dimensions, and the low-dimensional encodings, i.e. the one-hot (6-bit), binary 5-bit, and two machine-learning encodings, achieve poorer performances than those high-dimensional encodings. This correlation should be due to the sliding window scheme and random forests algorithm, larger feature dimension is more conducive to recognize the secondary structure states, but too large dimensions will lead to poor performance (ProtVec and ProtVec-3mer).

4.4 The performance comparison by using the BRNN method

Recent years, the deep learning based methods for protein secondary structure prediction have achieved significant improvements [2]. One of the most important advantages of deep learning methods is that they can capture both neighboring and long-range interactions, which could avoid the shortcomings of sliding window methods with handcrafted window size. For example, Heffernan et al. [48] have achieved the state-of-the-art performances by using the long short-term memory (LSTM) bidirectional recurrent neural networks. Therefore, to exclude the potential influence of the handcrafted window size, we also perform the assessment by using the bidirectional recurrent neural networks (BRNN) with long short-term memory cells. The model used here is similar to the model used in Heffernans work [48],

TABLE 3
The protein secondary structure prediction accuracy of sixteen amino acid encoding methods by using the Random Forests method

Category	Encoding method	Dimension	CB513	CASP10	CASP11	CASP12	Mean[a]
Binary	One-hot	20	60.21%	56.72%	59.31%	59.70%	58.98%
	One-hot (6-bit)	6	50.00%	47.97%	48.86%	48.78%	48.90%
	Binary 5-bit	5	48.09%	44.76%	45.78%	47.59%	46.55%
Physicochemical properties	Hydrophobicity matrix	20	56.46%	54.08%	55.48%	54.33%	55.09%
	Meiler parameters	7	62.29%	59.67%	61.19%	60.63%	60.94%
	Acthely factors	5	61.42%	59.04%	60.44%	60.49%	60.35%
Evolution-based	PAM250	20	61.65%	59.62%	60.85%	60.54%	60.67%
	BLOSUM62	20	62.30%	59.84%	61.53%	60.89%	61.14%
	PSSM	20	72.80%	71.58%	71.79%	71.75%	71.98%
	HMM	20	72.20%	72.27%	69.04%	68.51%	70.50%
Structure-based	Miyazawa energies	20	61.83%	59.68%	61.20%	60.86%	60.89%
	Micheletti potentials	20	59.61%	56.60%	58.19%	58.85%	58.31%
Machine-learning	AESNN3	3	54.80%	51.71%	52.60%	53.77%	53.22%
	ANN4D	4	58.01%	55.26%	56.41%	57.97%	56.91%
	ProtVec	100	50.00%	49.00%	50.29%	51.63%	50.23%
	ProtVec-3mer	163	51.87%	47.94%	49.41%	51.51%	50.18%

[a] Mean: the mean value of accuracies on three test datasets.

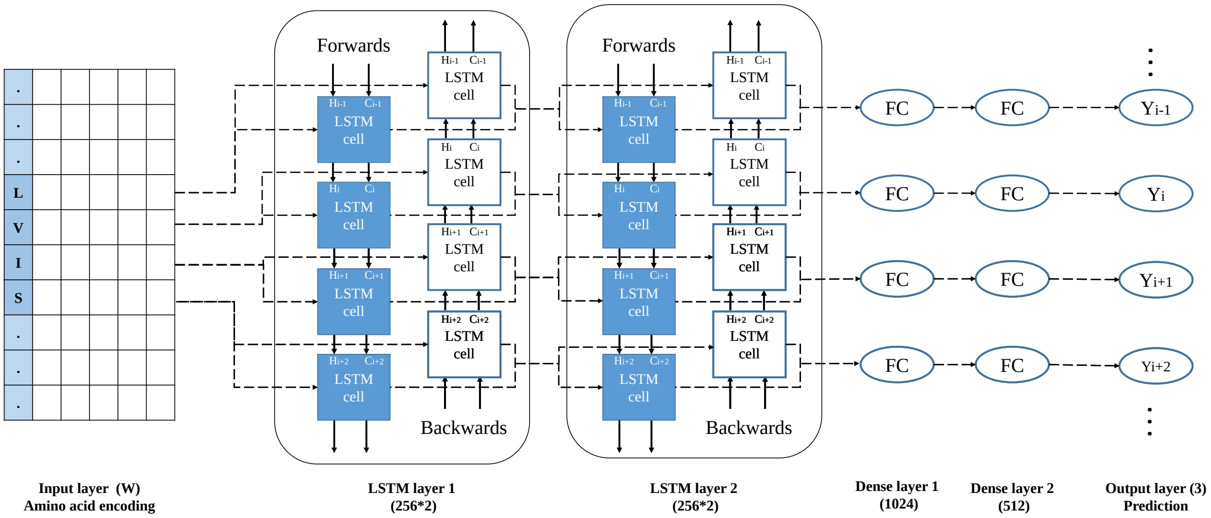


Fig. 5. The architecture of the long short-term memory (LSTM) bidirectional recurrent neural networks for protein secondary structure prediction.

as shown in Fig. 5, which contains two BRNN layers with 256 LSTM cells and two fully connected (dense) layers with 1024 and 512 nodes, and is implemented based on the open-sourced deep learning library TensorFlow [64].

The corresponding comparison results of sixteen selected encoding methods are shown in Table 4. From the overall view, the BRNN based method has better performances compared with the random forests based method, but there are also some specific similarities and differences between them. For the binary encoding methods, the one-hot encoding still achieves the best performance, which once again confirms the information loss of the one-hot (6-bit) and the binary 5-bit encoding methods. For the physicochemical properties encodings, the Meiler parameters performs not as well as the Acthely factors, suggesting that the Acthely factors is more efficient for deep learning methods. For the evolution-based encodings, the PSSM encoding achieves best accuracy, while the HMM encoding just achieves as much accuracy as those position-independent encodings (PAM250 and BLOSUM62). The difference could be due to the different level of homologous sequence identity. The HMM encoding is extracted from the UniProt20 database with 20% sequence identity, while the PSSM encoding is extracted from the UniRef90 database with 90% sequence identity. Therefore, for a certain protein sequence, its MSA from the UniProt20 database mainly contains remote homologous sequences, while its MSA from the UniRef90 database usually contains more homologous sequences. From the results in Table 4, the evaluation information of homologous sequences is more powerful to distinguish different protein secondary structures than that of remote homologous sequences. For the structure-based encodings, the Micheletti potentials has much better performance based on BRNN method than that based on random forests method. For machine-learning encodings, the ProtVec and ProtVec-3mer achieves significantly better performance compared with Table 4, which demonstrates the potential of machine-learning encoding. It is worth noting that ProtVec-3mer has better performances than ProtVec on BRNN algorithm, corresponding to the authors' recent work [52]. Overall, for the deep learning algorithm BRNN, the position-dependent PSSM encoding still performs best among all encoding methods. For the position-independent encoding methods, the Micheletti potentials achieve the best performance, which demonstrates the structure related information has the application potential in protein structure and function studies.

To measure the degree of performance differences between sixteen amino acid encoding methods, we perform the Students t-test on protein secondary structure prediction accuracy, the p-values of students t-test are shown in supplementary materials Table S1. The p-values greater than 0.05 indicate the similar performance between the two methods. From the Table S1, most p-values are small enough, which means that differences between these methods are statistically significant.

5 Assessments of encoding methods for protein fold recognition

In addition to the protein sequence labeling tasks, protein sequence classification tasks have also received a lot of attentions, such as protein remote homology detection [3] and protein fold recognition [11], [65]. Here, we perform another assessment of above sixteen selected amino acid encoding methods based on the protein fold recognition task. Many machine learning methods have been developed to classify protein sequences into different fold categories for protein fold recognition [3]. Especially the deep learning methods can automatically extract the discriminative patterns from variable-length protein sequences and achieve significant success [11]. Referring to the Hous work [11], we use the one-dimensional deep convolution neural network (DCNN) to assess the usefulness of sixteen selected encoding methods for protein fold recognition. As shown in Fig. 6, the deep convolution neural network used here has 10 hidden layers of convolution, 10 filters of each convolution layer with two window sizes (6 and 10), 20 maximum values at the max pooling layer, and a flatten layer which is fully connected with the output layer to output the corresponding probability of each fold type.

5.1 Benchmark datasets for protein fold recognition

The most commonly used dataset to evaluate protein fold recognition methods is the SCOP database [66] and its extended version SCOPe database [67]. The SCOP is a manual structural classification of protein whose three-dimensional structures have been determined. All of the proteins in SCOP are classified in four hierarchy levels: class, fold, superfamily and family. Folds represent the main characteristics of protein structures, and protein fold could reveal the evolutionary process between protein sequence and its corresponding tertiary structure [68]. Here we use the F184 dataset which was constructed by Xia et al. [65] based on the SCOPe database. The F184 dataset contains 6451 sequences with less than 25% sequence identity from 184 folds. Each fold contains at least 10 sequences, which could ensure that there are enough sequences for training and test purpose. Then we randomly selected 20% of the sequences as test data from each fold, leaving 80% of the sequence as training data. Finally, we got 5230 sequences for training and 1221 sequence for test.

5.2 Performances of different encodings on protein fold recognition task

The comparison results of sixteen selected encoding methods for protein fold recognition are listed in Table 5. It should be noted that the training process for each encoding method is repeated 10 times to eliminate stochastic effects. Different with the performances for protein secondary structure prediction, the performances of most position-independent encoding methods are similar. All of the binary, the physicochemical, and the machine-learning based encoding methods (except the ProtVec) achieve about 30% mean accuracies, which demonstrates that the position-independent encodings could just offer limited information for protein fold classification. The two structure-based encodings have better accuracies near 33%, demonstrating that the structure potential is more related with the protein fold type. The two evolution-based methods PAM250 and BLOSUM62 perform best among twelve position-independent encoding methods, which means the evaluation information is more coupled with the protein structure. The position-dependent encoding PSSM and HMM achieve better performances, especially the PSSM encoding has a significant performance. It again indicates that the protein evaluation information is tightly coupled with the protein structure, and the homologous information is more useful than remote homologous information. The machine-learning based AESNN3 and ANN4D encodings achieve comparable performances with other position-independent encoding methods but have much low dimensions (3 for the AESNN3 and 4 for the ANN4D), showing its potential for further application. The performance of the ProtVec encoding is poor, it should be caused by the overlapping strategy [69], [70] that has also been mentioned by the author [17]. While the ProtVec-3mer encoding has better performance, demonstrating the effectiveness of the combination of ProtVec and 3-mer.

It should be noted that the benchmark presented here is based on the DCNN method, and these encodings may achieve different performances by using other machine learning methods. The DCNN method could handle with variable-length sequences and achieve significant success on fold recognition task, which are the main reasons that it is chosen here.

5.3 The performance difference of different encodings

We also perform the Students t-test between sixteen amino acid encoding methods on Top 10 accuracy for protein fold recognition, the p-values are shown in supplementary materials Table S2. In Table S2, there are more p-values are not small compared with the p-values in Table S1. The results indicate that these methods may have similar performance for protein fold recognition.

6 Conclusion and future perspective

Amino acid encoding is the first step of protein structure and function prediction, and it is one of the foundations to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2911677, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 4
The protein secondary structure prediction accuracy of sixteen amino acid encoding methods by using the BRNN method

Category	Encoding method	Dimension	CB513	CASP10	CASP11	CASP12	Mean[a]
Binary	One-hot	20	66.32%	66.54%	68.75%	67.87%	67.37%
	One-hot (6-bit)	6	58.42%	59.84%	59.81%	59.14%	59.30%
	Binary 5-bit	5	55.04%	52.39%	52.43%	52.31%	53.04%
Physicochemical properties	Hydrophobicity matrix	20	55.69%	56.27%	57.24%	58.57%	56.94%
	Meiler parameters	7	52.88%	47.98%	49.69%	49.97%	50.13%
	Acthely factors	5	63.79%	61.54%	63.70%	64.07%	63.28%
Evolution-based	PAM250	20	66.67%	68.02%	70.22%	68.74%	68.41%
	BLOSUM62	20	66.18%	64.85%	67.58%	66.36%	66.24%
	PSSM	20	80.46%	82.25%	81.45%	79.44%	80.90%
	HMM	20	70.51%	68.55%	67.32%	66.30%	68.17%
Structure-based	Miyazawa energies	20	65.44%	61.19%	61.42%	61.46%	61.46%
	Micheletti potentials	20	70.85%	70.65%	72.76%	70.01%	71.07%
Machine-learning	AESNN3	3	49.80%	51.19%	51.74%	51.72%	51.72%
	ANN4D	4	57.18%	53.09%	54.53%	54.67%	54.87%
	ProtVec	100	67.48%	66.39%	68.62%	67.13%	67.41%
	ProtVec-3mer	163	70.69%	69.43%	72.43%	72.43%	71.24%

[a] Mean: the mean value of accuracies on three test datasets.

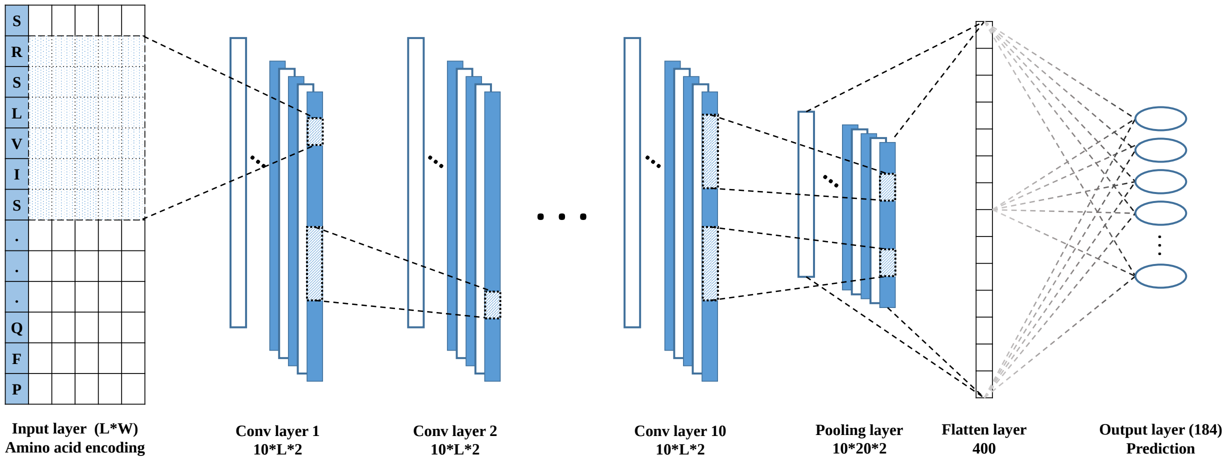


Fig. 6. The architecture of the one-dimensional deep convolution neural network for protein fold recognition.

TABLE 5
The protein fold recognition accuracy of sixteen amino acid encoding methods by using the DCNN method

Category	Encoding method	Dimension	Top 1[a]	Top 5[b]	Top 10[c]	Mean[d]
Binary encoding	One-hot	20	13.50%	35.50%	48.09%	32.37%
	One-hot (6-bit)	6	13.39%	35.01%	47.87%	32.09%
	Binary 5-bit	5	12.57%	33.05%	46.26%	30.63%
Physicochemical properties	Hydrophobicity matrix	20	12.68%	33.56%	46.53%	30.92%
	Meiler parameters	7	12.62%	33.05%	45.82%	30.50%
	Acthely factors	5	14.97%	36.93%	50.70%	34.20%
Evolution-based	PAM250	20	18.70%	43.69%	56.66%	39.68%
	BLOSUM62	20	16.86%	41.80%	55.25%	37.97%
	PSSM	20	47.04%	74.38%	84.57%	68.66%
	HMM	20	20.02%	43.51%	56.40%	39.98%
Structure-based	Miyazawa energies	20	13.95%	36.32%	49.62%	33.30%
	Micheletti potentials	20	14.86%	37.61%	50.30%	34.26%
Machine-learning	AESNN3	3	11.85%	32.58%	45.63%	30.02%
	ANN4D	4	11.77%	31.91%	45.0%	29.58%
	ProtVec	100	6.18%	22.91%	33.42%	20.84%
	ProtVec-3mer	163	12.01%	33.41%	45.12%	30.18%

[a] Top 1: the accuracy calculated in the case that the first predicted folding type is the actual folding type.
[b] Top 5: the accuracy calculated in the case that the top 5 predicted fold types contain the actual fold type.
[c] Top 5: the accuracy calculated in the case that the top 10 predicted fold types contain the actual fold type.
[d] Mean: the mean value of accuracies on Top 1, Top 5, and Top 10.

achieve the final success of those studies. In this article, we propose the first systematic classification of various amino acid encoding methods and review the methods of each category. According to the information sources and information extraction methodologies, these methods are grouped into five categories: binary encoding, physicochemical properties encoding, evolution-based encoding, structure-based encoding, and machine-learning encoding. To benchmark and compare different amino acid encoding methods, we first select sixteen representative methods from those five categories. And then, based on the two representative protein related studies: protein secondary structure prediction and protein fold recognition, we construct three machine learning models referring to the state-of-the-art studies. Finally, we encode the protein sequence and implement the same training and test phase on the benchmark datasets for each encoding method. The performance of each encoding method is regarded as the indicator of its potential in protein structure and function studies.

The assessment results show that the evolution-based position-dependent encoding method PSSM consistently achieve the best performance both on protein secondary structure prediction and protein fold recognition tasks, suggesting its important role in protein structure and function prediction. However, another evolution-based position-dependent encoding method HMM does not perform well, the main reason could be that the remote homologous sequences only provide limited evaluation information for the target residue. For the one-hot encoding method, it is highly sparse and leads to complex machine learning models, while its two compressed representations, the one-hot (6-bit) encoding and the binary 5-bit encoding, lose more or less valuable information and cannot be widely used in related researches. More reasonable strategies to reduce the dimension of one-hot encoding need to be developed. For the physicochemical properties encodings, the variety of properties and extraction methodologies are two important factors to construct a valuable encoding. The structure-based encodings and the machine-learning encodings achieve comparable or even better performances when compared with other widely used encodings, suggesting more attention need to be paid on these two categories.

In the time when the dividends of data and algorithm have been highly released, exploring more effective encoding schemas for amino acids should be a key factor to further improve the performance of protein structure and function prediction. In the following, we provide some perspectives for future related researches. First, updated position-independent encodings should be constructed based on new protein data sets. Except for the one-hot encoding, all other position-independent encoding methods construct their encodings based on the information extracted from the native protein sequences or structures. There is no doubt that random errors are unavoidable for those encodings and larger data sets will help to reduce those errors. As the development of sequencing and structure detection techniques, the number of protein sequences and structures is growing rapidly in the past years. Considering that most of the position-independent encoding methods are proposed one decade ago, it should be valuable to reconstruct them by using new datasets. Second, the structure-based or function-based encoding methods require more attention. The structure-based encoding methods have been demonstrated by its ability in protein secondary structure prediction and protein fold recognition. These encodings reflect the structural potential of amino acids, which should be highly correlated with the protein structure and function. With the growing of the number of proteins with known structure, the prospect of structure-based encodings is considerable. Furthermore, the encodings reflecting function potentials may be more useful than others for protein function prediction, thus exploring function-based encoding methods is a worthwhile topic. Third, the machine-learning encoding methods are promising in future studies. As the amino acid encoding is an open problem, most encoding methods are based on artificially defined basis, i.e. the physicochemical properties encodings are constructed from protein fold related properties observed by researchers, which will inevitably bring some unknown deviations. However, the machine-learning methods can avoid those artificial deviations by learning the amino acid encoding from biological data automatically. The protein sequences and natural languages share some similarities to certain extent, for instance, the protein sequences can be seen as the sentences, and the amino acid or polypeptide chain can be seen as the words in languages. Considering that the word distributed representation has achieved comprehensive improved performances in natural language processing tasks, the protein sequences should also gain improvements by using the distributed representations of amino acid or n-gram amino acids. Some recent studies have demonstrated the potential of amino acid distributed representations in protein family classification, disordered protein identification and protein functional properties prediction, but most of these methods are concerned with the n-gram amino acid distributed representations that cannot be directly used to the residue-level properties prediction. The residue-level distributed representations of amino acid need more attention.

ACKNOWLEDGMENTS

Daocheng Hong is the corresponding author of this paper. This work has been supported by the National Natural Science Foundation of China under Grant No. 61672234, U1401256, U1711262, 61402177, the National Key Research and Development Program of China under grant 2016YFB1000905.

REFERENCES

[1] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973, 7317.

[2] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y. Zhou, "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?" *Briefings in Bioinformatics*, p. bbw129, 2016.

[3] J. Chen, M. Guo, X. Wang, and B. Liu, "A comprehensive review and comparison of different computational methods for protein remote homology detection," *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 231–244, Mar. 2018.

[4] Q. Wuyun, W. Zheng, Z. Peng, and J. Yang, "A large-scale comparative assessment of methods for residue-residue contact prediction," *Briefings in Bioinformatics*, p. bbw106, 2016.

[5] J. Zhang and L. Kurgan, "Review and comparative assessment of sequence-based predictors of protein-binding residues," *Briefings in Bioinformatics*, 2017.

[6] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, p. bbw068, 2016.

[7] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, pp. 6395–6400, 2005, 256.

[8] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top- n -grams and latent semantic analysis.(Research article)," *BMC Bioinformatics*, vol. 9, no. 510, p. 510, 2008.

[9] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. W1, pp. W65–W71, 2015.

[10] B. Liu, "BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings in bioinformatics*, 2017.

[11] J. Hou, B. Adhikari, and J. Cheng, "DeepSF: deep convolutional neural network for mapping protein sequences to folds," *Bioinformatics*, vol. 34, no. 8, pp. 1295–1303, Apr. 2018, 00000.

[12] M. Zamani and S. C. Kremer, "Amino acid encoding schemes for machine learning methods," in *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Nov. 2011, pp. 327–333.

[13] P. D. Yoo, B. B. Zhou, and A. Y. Zomaya, "Machine learning techniques for protein secondary structure prediction: an overview and evaluation," *Current Bioinformatics*, vol. 3, no. 2, pp. 74–86, 2008.

[14] H.-J. Hu, Y. Pan, R. Harrison, and P. C. Tai, "Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier," *IEEE Transactions on NanoBioscience*, vol. 3, no. 4, pp. 265–271, 2004, 91.

[15] S. Miyazawa and R. L. Jernigan, "Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 1, pp. 49–68, Jan. 1999, 183.

[16] K. Lin, A. C. W. May, and W. R. Taylor, "Amino Acid Encoding Schemes from Protein Structure Alignments: Multi-dimensional Vectors to Describe Residue Types," *Journal of Theoretical Biology*, vol. 216, no. 3, pp. 361–365, Jun. 2002, 24.

[17] E. Asgari and M. R. K. Mofrad, "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics," *PLOS ONE*, vol. 10, no. 11, p. e0141287, Nov. 2015, 00000.

[18] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D202–D205, Jan. 2008.

[19] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields," *Scientific Reports*, vol. 6, 2016.

[20] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu, "New techniques for extracting features from protein sequences," *IBM Systems Journal*, vol. 40, no. 2, pp. 426–441, 2001, 127.

[21] M. O. Dayhoff, "A model of evolutionary change in proteins," *Atlas of protein sequence and structure*, vol. 5, pp. 89–99, 1978, 649.

[22] G. White and W. Seffens, "Using a neural network to backtranslate amino acid sequences," *Electronic Journal of Biotechnology*, vol. 1, no. 3, pp. 17–18, Dec. 1998.

[23] G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, no. 4716, pp. 834–838, Aug. 1985.

[24] M. J. Betts and R. B. Russell, "Amino acid properties and consequences of substitutions," in *Bioinformatics for geneticists*, 2003, vol. 317, p. 289, 500.

[25] J.-L. Fauchre, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *Chemical Biology &amp; Drug Design*, vol. 32, no. 4, pp. 269–278, Oct. 1988, 326.

[26] A. Radzicka and R. Wolfenden, "Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution," *Biochemistry*, vol. 27, no. 5, pp. 1664–1670, 1988, 624.

[27] Lohmann Reinhard, Schneider Gisbert, Behrens Dirk, and Wrede Paul, "A neural network model for the prediction of membranespanning amino acid sequences," *Protein Science*, vol. 3, no. 9, pp. 1597–1601, 1994, 46.

[28] A. Elofsson, "A study on protein sequence alignment quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 46, no. 3, pp. 330–339, 2002.

[29] E. E. Oren, C. Tamerler, D. Sahin, M. Hnilova, U. O. S. Seker, M. Sarikaya, and R. Samudrala, "A novel knowledge-based approach to design inorganic-binding peptides," *Bioinformatics*, vol. 23, no. 21, pp. 2816–2822, 2007.

[30] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, Nov. 1992, 5799.

[31] ——, "Automated assembly of protein blocks for database searching," *Nucleic acids research*, vol. 19, no. 23, pp. 6565–6572, 1991.

[32] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the Perceptronalgorithm to distinguish translational initiation sites in E. coli," *Nucleic acids research*, vol. 10, no. 9, pp. 2997–3011, 1982, 572.

[33] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLASTa tool for discovery in protein databases," *Trends in biochemical sciences*, vol. 23, no. 11, pp. 444–447, 1998, 643.

[34] M. Remmert, A. Biegert, A. Hauser, and J. Sding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature methods*, vol. 9, no. 2, p. 173, 2012, 694.

[35] S. Tanaka and H. A. Scheraga, "Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins," *Macromolecules*, vol. 9, no. 6, pp. 945–950, Nov. 1976, 457.

[36] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," *Macromolecules*, vol. 18, no. 3, pp. 534–552, May 1985, 1622.

[37] ——, "Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading," *Journal of Molecular Biology*, vol. 256, no. 3, pp. 623–644, Mar. 1996.

[38] J. Skolnick, A. Godzik, L. Jaroszewski, and A. Kolinski, "Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?" *Protein science*, vol. 6, no. 3, pp. 676–688, 1997, 217.

[39] Simons Kim T., Ruczinski Ingo, Kooperberg Charles, Fox Brian A., Bystroff Chris, and Baker David, "Improved recognition of nativelike protein structures using a combination of sequencedependent and sequenceindependent features of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 1, pp. 82–95, Oct. 1999, 473.

[40] C. Zhang and S.-H. Kim, "Environment-dependent residue contact energies for proteins," *Proceedings of the National Academy of Sciences*, vol. 97, no. 6, pp. 2550–2555, 2000, 120.

[41] M. Cristian, S. Flavio, B. J. R., and M. Amos, "Learning effective amino acid interactions through iterative stochastic techniques," *Proteins: Structure, Function, and Bioinformatics*, vol. 42, no. 3, pp. 422–431, Jan. 2001, 50.

[42] S. K. Riis and A. Krogh, "Improving Prediction of Protein Secondary Structure Using Structured Neural Networks and Multiple Sequence Alignments," *Journal of Computational Biology*, vol. 3, no. 1, pp. 163–183, Jan. 1996, 211.

[43] B. Jagla and J. Schuchhardt, "Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites," *Bioinformatics*, vol. 16, no. 3, pp. 245–250, Mar. 2000, 47.

[44] J. Meiler, M. Mller, A. Zeidler, and F. Schmschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *Molecular modeling annual*, vol. 7, no. 9, pp. 360–369, Sep. 2001, 145.

[45] Y. Xu, J. Song, C. Wilson, and J. C. Whisstock, "PhosContext2vec: a distributed representation of residue-level sequence contexts and its application to general and kinase-specific phosphorylation site prediction," *Scientific Reports*, vol. 8, May 2018, 00000.

[46] K. K. Yang, Z. Wu, C. N. Bedbrook, and F. H. Arnold, "Learned protein embeddings for machine learning," *Bioinformatics*, 2018, 00000.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[48] R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, "Capturing Non-Local Interactions by Long Short Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers, and Solvent Accessibility," *Bioinformatics*, p. btx218, 2017.

[49] R. David, "Applications of nonlinear system identification to protein structural prediction," PhD Thesis, Massachusetts Institute of Technology, 2000, 00000.

[50] W. Zhong, G. Altun, X. Tian, R. Harrison, P. C. Tai, and Y. Pan, "Parallel protein secondary structure prediction based on neural networks," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 2. IEEE, 2004, pp. 2968–2971, 00000.

[51] J. Dongardive and S. Abraham, "Reaching optimized parameter set: protein secondary structure prediction using neural network," *Neural Computing and Applications*, vol. 28, no. 8, pp. 1947–1974, Aug. 2017, 3.

[52] E. Asgari, A. C. McHardy, and M. R. Mofrad, "Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (dimotif) and sequence embedding (protvecx)," *Scientific reports*, vol. 9, no. 1, p. 3577, 2019.

[53] T. U. Consortium, "The Universal Protein Resource (UniProt)," vol. 36, no. Database issue, pp. D190–D195, 2008, 00000.

[54] Z. Li and Y. Yu, "Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks," *arXiv preprint arXiv:1604.07176*, 2016.

[55] A. R. Johansen, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Deep recurrent conditional random field network for protein secondary prediction," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 73–78.

[56] J. Zhou, H. Wang, Z. Zhao, R. Xu, and Q. Lu, "Cnnh_pss: protein 8-class secondary structure prediction by convolutional neural network with highway," *BMC bioinformatics*, vol. 19, no. 4, p. 60, 2018.

[57] G. Wang and R. L. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, Aug. 2003.

[58] Cuff James A. and Barton Geoffrey J., "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 4, pp. 508–519, Sep. 1999.

[59] Moult John, Fidelis Krzysztof, Kryshtafovych Andriy, Schwede Torsten, and Tramontano Anna, "Critical assessment of methods of protein structure prediction (CASP) round x," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. S2, pp. 1–6, Oct. 2013.

[60] L. N. Kinch, W. Li, R. D. Schaeffer, R. L. Dunbrack, B. Monastyrskyy, A. Kryshtafovych, and N. V. Grishin, "CASP 11 Target Classification," *Proteins Structure Function & Bioinformatics*, vol. 84, no. S1, pp. 20–33, 2016.

[61] Moult John, Fidelis Krzysztof, Kryshtafovych Andriy, Schwede Torsten, and Tramontano Anna, "Critical assessment of methods of protein structure prediction (CASP)Round XII," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. S1, pp. 7–15, Dec. 2017.

[62] Kabsch Wolfgang and Sander Christian, "Dictionary of protein secondary structure: Pattern recognition of hydrogenbonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Feb. 2004.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[64] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2016.

[65] J. Xia, Z. Peng, D. Qi, H. Mu, and J. Yang, "An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier," *Bioinformatics*, vol. 33, no. 6, pp. 863–870, Mar. 2017.

[66] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, Apr. 1995.

[67] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: Structural Classification of Proteinsextended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic Acids Research*, vol. 42, no. D1, pp. D304–D309, Jan. 2014.

[68] T. Damoulas and M. A. Girolami, "Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection," *Bioinformatics*, vol. 24, no. 10, pp. 1264–1270, May 2008.

[69] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman, "Comparative n-gram analysis of whole-genome protein sequences," in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 76–81.

[70] S. M. Srinivasan, S. Vural, B. R. King, and C. Guda, "Mining for class-specific motifs in protein sequence classification," *BMC bioinformatics*, vol. 14, no. 1, p. 96, 2013.

Xiaoyang Jing received the B.E. degree from East China University of Science and Technology in 2014. He is currently a doctoral student in School of Computer Science and Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China.

His main research interests include bioinformatics, big data and machine learning.

Qiwen Dong received the B.E., M.E., and Ph.D. degrees from Harbin Institute of Technology. Currently, he is a professor in School of Data Science and Engineering at East China Normal University, Shanghai, China.

His main research interests include bioinformatics, network informatics, big data processing and computational education.

Daocheng Hong received his Bachelor Degree from Wuhan University, Master Degree and Ph.D. Degree from Fudan University. He was the Visiting Fellow at Victoria University sponsored by the China Scholarship Council (CSC) and a faculty member at Fudan University. Currently, he is an associate professor of Data Science and Engineering School at East China Normal University, Shanghai, China.

His core research experience is in the field of machine learning, knowledge-based systems and data management.

Ruqian Lu received the Diploma degree in mathematics from Jena University, Jena, Germany, in 1959. Currently, he is a Professor at the Institute of Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, and the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China.

His research interests include quantum information processing, formal semantics, artificial intelligence, knowledge engineering, knowledge-based software engineering, bioinformatics, etc.

Mr. Lu is a Fellow of the Chinese Academy of Sciences.