Computational and Structural Biotechnology Journal 20 (2022) 3503–3510

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

Research progress of reduced amino acid alphabets in protein analysis and prediction

Yuchao Liang [a], Siqi Yang [a], Lei Zheng [a], Hao Wang [a], Jian Zhou [a], Shenghui Huang [a], Lei Yang [b,*], Yongchun Zuo [a,*]

[a] State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot, China
[b] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

ARTICLE INFO

Article history:
Received 14 March 2022
Received in revised form 30 June 2022
Accepted 1 July 2022
Available online 4 July 2022

Keywords:
Reduced amino acid alphabets
Machine learning
Sequence alignment
Protein classification
Structure analysis

ABSTRACT

Proteins are the executors of cellular physiological activities, and accurate structural and function elucidation are crucial for the refined mapping of proteins. As a feature engineering method, the reduction of amino acid composition is not only an important method for protein structure and function analysis, but also opens a broad horizon for the complex field of machine learning. Representing sequences with fewer amino acid types greatly reduces the complexity and noise of traditional feature engineering in dimension, and provides more interpretable predictive models for machine learning to capture key features. In this paper, we systematically reviewed the strategy and method studies of the reduced amino acid (RAA) alphabets, and summarized its main research in protein sequence alignment, functional classification, and prediction of structural properties, respectively. In the end, we gave a comprehensive analysis of 672 RAA alphabets from 74 reduction methods.

Contents

1. Introduction . . . 3503
2. The reduction methods of natural amino acid alphabets . . . 3504
3. The application of reduced amino acid alphabets for sequence alignment . . . 3505
4. The classification of protein function based on reduced amino acid alphabets . . . 3505
5. The prediction of protein structure property based on reduced amino acid alphabets . . . 3506
6. A comprehensive analysis of the 672 reduced amino acid alphabets . . . 3507
7. Conclusion . . . 3508
Funding . . . 3508
CRediT authorship contribution statement . . . 3508
Declaration of Competing Interest . . . 3508
Acknowledgements . . . 3508
Appendix A. Supplementary data . . . 3508
References . . . 3508

1. Introduction

As the direct execution molecules of cellular life activities, the study of proteins has received much attention in the past few decades. With the maturity of technologies such as high-throughput sequencing, mass spectrometry, and co-immunoprecipitation, more and more protein sequence, structure, and function data have been annotated and published, which opened the way for human proteomics research [1,2]. However, it has been gradually discovered that there are many drawbacks in the method of annotating protein information experimentally, such as time-wasting, expensive consumables, inefficiency, etc.

* Corresponding authors.
E-mail addresses: leiyang@hrbmu.edu.cn (L. Yang), yczuo@imu.edu.cn (Y. Zuo).

https://doi.org/10.1016/j.csbj.2022.07.001

In recent years, the analysis and prediction methods based on machine learning and artificial intelligence have been continuously developed and applied to the research of biology and bioinformatics, which greatly shorten the experimental time and improve the experimental efficiency [3,4]. However, researchers' work is hindered by the cumbersome feature engineering, the increased complex network model architectures, and ever-upgrading hardware requirements [5,6]. To this end, people are also seeking balance, resulting in various feature analysis and optimization methods, such as principal component analysis (PCA), relief algorithm, F-score, linear dimension reduction algorithm (LDA) and more streamlined model architectures such as deep residual networks (ResNet) [7–11].

The simplified amino acid composition greatly reduces the dimensions of traditional feature engineering, effectively suppresses the negative effects of noise, and provides the model with richer biological prior knowledge to extract key features [12,13]. In addition, it is highly inclusive and has good compatibility with many existing methods, which helps to promote the further integration of traditional machine learning and biology [14,15].

RAA alphabets are not a recent product, which had been mentioned as early as the 1960s. Morita et al. proposed in 1967 that three-clusters random polypeptide segments (Glu, Lys, Ala) can form α helix [16]. In 1992, Heinz et al. confirmed the existence of a lot of redundant information in the amino acid sequence through the phage T4 lysozyme mutation experiment [17]. In the same year, the evolution of amino acid types from simple to complex was demonstrated by Osawa et al. [18]. A five-clusters reduction scheme was proposed by Riddle et al. in 1997 through the phage SH3 domain [19], which was tested by Wolynes from the perspective of energy [20]. Schafmeister et al. also proposed the use of a seven-clusters reduction scheme to synthesize 4 helical protein bundles [21]. In 1999, Wang et al. proposed a minimal mismatch-based RAA alphabets named HP, which laid a theoretical foundation for the research on RAA alphabets [22]. Their model still plays an important role in many theories until now.

As part of feature engineering, the most important feature of the reduced amino acid (RAA) composition is the fundamental redefinition of sequence. For any protein sequence, 20 amino acid residues can be grouped by specific methods and assigned new identifiers to each class (Fig. 1A). We construct sequences using the c residues and map them one-to-one with the natural sequence (Fig. 1C). According to specific clustering rules, we construct RAA alphabets of different clusters (Size 2–19), which is more conducive to the wide adaptation of the same reduced alphabet to different protein data (Fig. 1B).

In the following, we systematically review the methodological studies of the reduced amino acid alphabets and their major progress in protein sequence alignment, functional classification, and prediction of structural properties. The 672 RAA alphabets of the 74 reduction methods will be comprehensively discussed in the end.

2. The reduction methods of natural amino acid alphabets

Since the 21st century, the rapid development of computer technology and the raise of various amino acid mutation matrices (such as Miyazawa and Jernigan's MJ-matrix [23], BLOSUM matrix [24,25], PAM matrix [26,27], JTT matrix [28], WAG matrix [29]) have expanded the application direction of RAA. Murphy et al. used the BLOSUM50 mutation matrix to illustrate the effect of RAA on protein folding and predict that only 10–12 clusters of RAA alphabets would be required to represent different families of proteins [30]. Kosiol et al. constructed the new RAA alphabets using a Markov model based on PAM matrix and WAG matrix which were famous in the field of sequence alignments and phylogenetic trees [31]. Cannata et al. used multiple substitution matrices such as PAM and BLOSUM to perform an exhaustive analysis of all possible RAA alphabets and built it into the WebServer platform AlphaSimp [32].

Then, some biologists boldly put the RAA alphabets into practical application, trying to apply RAA alphabets to existing research. Akanuma et al. replaced 88% of the amino acid sequence with AAA reduced sequences (A, D, G, L, P, R, T, V, Y) by site-directed mutagenesis of Escherichia coli whey phosphoglycosyltransferase, which did not affect the structure and function of the protein [33]. Davies et al. developed a G protein-coupled receptor (GPCR)

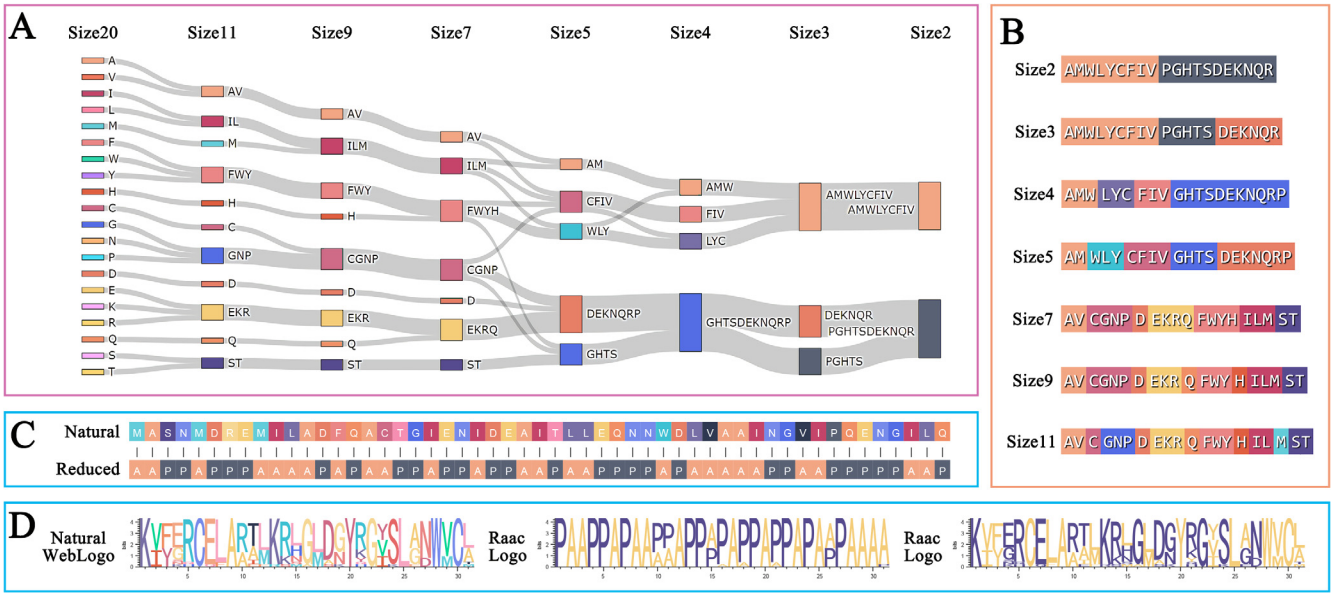

Fig. 1. A reduced amino acid alphabet of two-clusters (AMWLYCFIV-PGHTSDEKNQR) can be represented in protein sequence by AP. A: Sankey diagram of RAA alphabet, the different colors on the left represent 20 different amino acids; the right side represents 20 amino acids are gradually clustered into two clusters. B: RAA alphabets of different clusters under the same reduction method (Size 2–11). C: Alignment of original sequence and RAA sequence. D: The application of RAA alphabet in WebLogo. Left, middle and right respectively represent the original Weblogo, RaacLogo by the first letter of each cluster and RaacLogo by color of each cluster.

classifier through artificial immune algorithm (AIS) combined with RAA alphabets, and achieved great results [34].

In the past research, the RAA alphabets based protein prediction methods mostly relied on traditional machine learning techniques like support vector machine (SVM) [35]. They achieved superior performance in many scenarios. For example, in 2004, Weathers et al. used RAA alphabets based on SVM to classify and predict intrinsically disordered proteins, and achieved an accuracy of about 87% [36]. In 2009, Bohnstingl et al. used the RAA-based BioHEL to predict the number of contacts and relative solvent accessibility of protein structures [37]. Yang et al. proposed an RAA-SVM model for predicting protein subcellular localization in 2015, and compared the prediction performance of different machine learning models in detail [38].

A new generation of deep learning based machine learning algorithms greatly enhanced the customization and application of RAA alphabets. In 2001, Meiler et al. published an RAA alphabets generation method based artificial neural network, and proposed that each amino acid can be replaced by several sets of physical features [39]. In 2020, Oberti et al. used a convolutional neural network based RAA alphabets to predict the intrinsically disordered regions of proteins [40].

3. The application of reduced amino acid alphabets for sequence alignment

Sequence alignment and sequence search algorithms are not only one of the most commonly used methods in bioinformatics but also the cornerstone of many mainstream protein analysis methods. However, with the continuous increase of protein data and sequence complexity, the efficiency of multiple sequence alignment in huge databases is gradually unsatisfactory. There have been a lot of studies to improve the speed of sequence alignment from different methods, among which the RAA composition has been used as a common dimension reduction method in many excellent studies.

Algorithms for sequence alignment of proteins usually have high time complexity due to the diversification of sequences. Murphy et al. analyzed in detail the protein alignment effect of the RAA alphabets with different sizes, and pointed out that alphabets with less than 10 clusters would greatly lose sequence information. Ye et al. developed the fast protein similarity search tools RAPSearch and RAPSearch2 based on the 10-clusters RAA alphabet, which are 20–90 times faster than BLAST, and more significantly for shorter reads [41,42]. Buchfink et al. constructed DIAMOND, which is a fast protein sequence alignment algorithm using an algorithm based on a double index of the RAA alphabet [43]. It is 40–20,000 times faster than BLAST and has close sensitivity, which greatly improves alignment efficiency in large databases. Steinegger et al. proposed that the Kmer based on RAA alphabets and BLOSUM62 matrix can greatly improve the efficiency of sequence alignment, and developed a series of sequence search/clustering algorithms and tools for MMSeq based on this method [44–46]. Melo et al. used the RAA composition to align distant homologous sequences, and pointed out that fewer amino acid species would improve the alignment performance of conserved structures of distant homologous sequences [47].

4. The classification of protein function based on reduced amino acid alphabets

With the exponential expansion of proteomic data, using machine learning methods to mine the sequence intrinsic regularities behind the functions of known proteins from massive data and make accurate predictions about the functions, families and cellular localization of unknown proteins has become a focus of the current research work. Simplified amino acid alphabets greatly expand the method of protein sequence feature representation, and restore the seemingly complex and disordered sequence due to evolutionary mutation to a more conservative and concise state. It not only explains the sequence properties and evolutionary direction of proteins in biology, but also improves the prediction performance of the model.

In 2007, Chen et al. constructed a six-clusters reduced alphabet based on amino acid hydrophilicity and hydrophobicity, which successfully predicted the subcellular localization of apoptotic proteins, emphasizing the importance of hydrophilicity in the study of protein subcellular localization sex [48]. In 2012, Lin et al. constructed a multi-classification model of the ketoacyl synthase family based on RAA-SVM, which enabled SVM to obtain important compositional features of proteins [49]. In 2013, Feng et al. developed iHSP-PseRAAAC for predicting heat shock proteins and achieved good performance in complex classification tasks [50]. In 2014, Liu et al. published a prediction model for DNA-binding proteins based on RAA alphabets, which greatly reduced the feature dimension of traditional pseudo-amino acids and improved the prediction performance [51]. Similarly, our previous works successfully applied RAA alphabets in important research fields such as protein subtype classification, protein subfamily classification, and protein subcellular localization [52–54]. Veltri et al. published a reduced alphabet model based on deep learning in 2018, and successfully improved the recognition accuracy of antimicrobial peptides [55].

It is worth noting that the reduction alphabets of amino acids directly affect the performance of classification prediction, and it is important to choose the most suitable reduction scheme among a large number of imputation models. In 2008, Davies et al. used the artificial immune system (AIS) to screen the RAA alphabets most suitable for G protein-coupled receptors, and analyzed the contribution and significance of the reduced alphabet in the GPCR classification model through classifier prediction results [34]. By comparing different reduction alphabets, they found that cysteines always tend to be grouped independently, which is closely related to the formation of disulfide bonds and the maintenance of spatial structure of GPCRs, and is a key feature of GPCR classification. In 2019, we used a RAA-based Kmer method to predict defensins, small antimicrobial proteins that play an important role in cellular nonspecific immunity [12]. By modeling the predictions for the K = 2 and K = 3 features of more than 600 reduced alphabets, the best prediction performance was finally achieved in the "PGEKRQDSNTHClVW-YF-ALM" scheme with K = 2, and the highest prediction scores were achieved in different species and different excellent results were obtained in the defensin prediction of the family.

In addition, a large number of researchers are also working on the construction and popularization of RAA alphabets platforms, which can also be obtained in Table 1. In 2007, Shimizu proposed POODLE-S, a protein disorder prediction platform based on amino acid physicochemical properties and position-specific scoring matrix, which has received extensive attention and citations [56]. In 2017, our group built an RAA platform PseKRAAC based on pseudo-amino acids and Kmers, and integrated 16 amino acid sequence reduction schemes, which facilitated non-bioinformatics researchers [57]. In 2019, Xi et al. proposed a mapping tool platform based on RAA method, RaaMLab. They organize a large database of amino acid physicochemical properties and support user-defined reduced alphabets [58]. In recent years, we successively constructed iDEF-PseRAAC, RaacLogo, RaacBook, OGFE-RAAC and other protein analysis and prediction platforms based on RAA alphabets, which enriched the application scope of RAA alphabets and emphasized the important role of simplified amino acid composition in sequence-structure–function (Fig. 1D and Table 1) [12–15,59].

Table 1
RAA Webserver platform summary.

Webserver Name	Link	Cite
PseKRAAC	http://bigdata.imu.edu.cn/	[57]
RAACBook	http://bioinfor.imu.edu.cn/raacbook	[14]
RaacLogo	http://bioinfor.imu.edu.cn/raaclogo	[59]
iSP-RAAC	http://bioinfor.imu.edu.cn/ispraac/public	[60]
iDEF-PseRAAC	http://bioinfor.imu.edu.cn/idpf	[12]
iHEC-RAAC	http://bioinfor.imu.edu.cn/ihecraac	[13]
POODLE-S	http://mbs.cbrc.jp/poodle/poodle-s.html (Inaccessible)	[56]
RaaMLab	https://github.com/bioinfo0706/RaaMLab	[58]
iHSP-PseRAAAC	http://lin-group.cn/server/iHSP-PseRAAAC	[50]
OGFE-RAAC	http://bioinfor.imu.edu.cn/ogferaac	[15]
iDNA-Prot	http://bioinformatics.hitsz.edu.cn/iDNA-Prot_dis/	[51]
PROFEAT	http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi (Inaccessible)	[61]
cnnAlpha	https://github.com/mauricioob/shiny-pred	[40]
iDPF-PseRAAAC	http://wlxy.imu.edu.cn/college/biostation/fuwu/iDPF-PseRAAAC/index.asp (Inaccessible)	[54]

5. The prediction of protein structure property based on reduced amino acid alphabets

The structure of protein is a decisive factor in its functioning. A large number of proteins with unique functions are obviously conserved in their natural structures. For example, GPCRs have seven transmembrane domains, and their structures show clear rules of solvent accessibility. However, the detection methods of protein structure and properties are complicated, and the manual analysis is inefficient, which has been plaguing the whole biological world. The traditional identification of protein structure properties requires professional technicians to gradually explore through methods such as X-ray crystallography and nuclear magnetic resonance, which takes a long time. After the rise of bioinformatics,

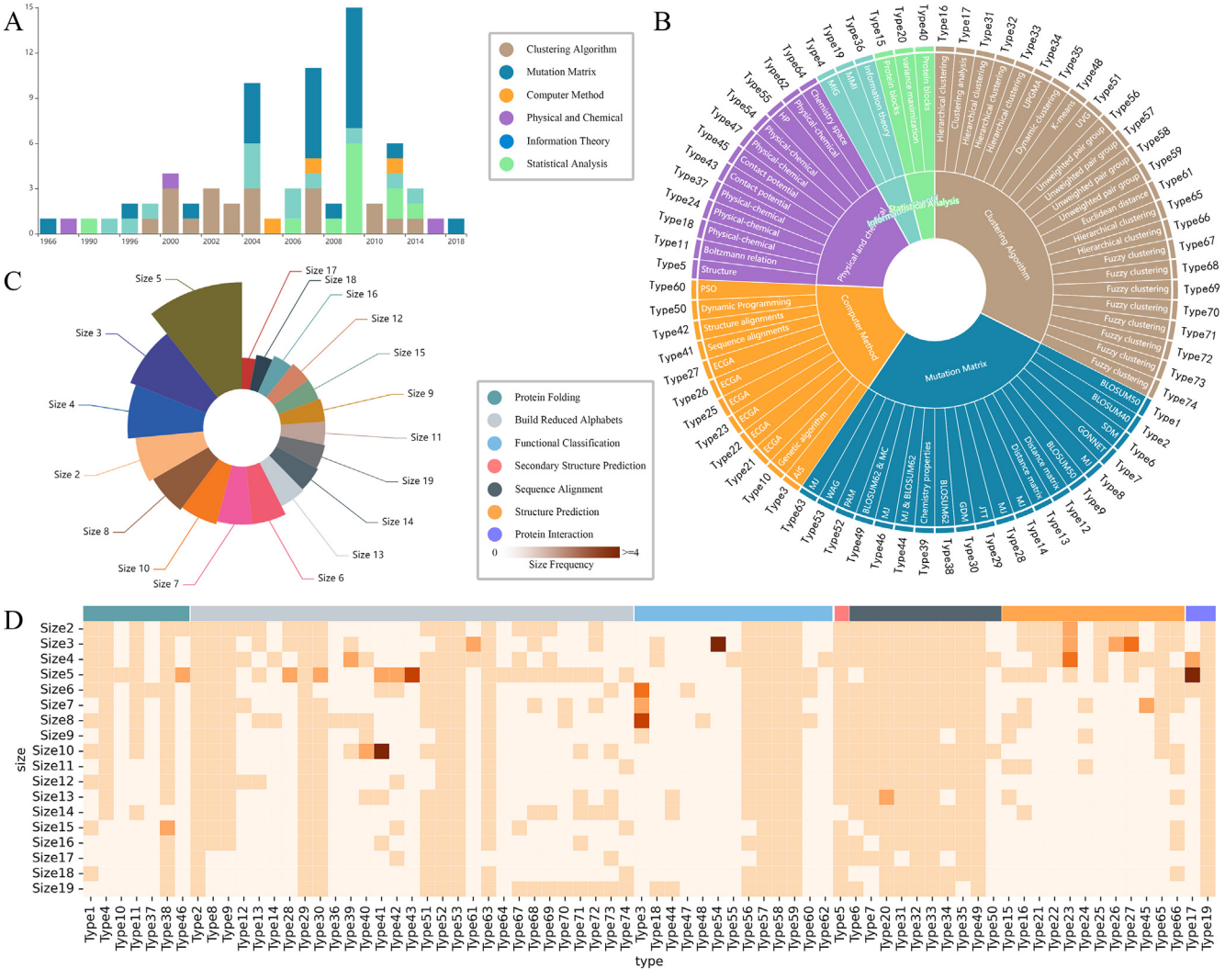

Fig. 2. Statistics of 672 RAA alphabets in 74 reduction methods. A: The 74 reduction methods are divided into 6 categories according to different principles, and are arranged on the timeline. B: The 672 RAA alphabets are divided into 6 categories according to different principles, and correspond to Type. C: RAA alphabets of different clusters (Size 2–19) in the 74 alphabets. D: Summarize all RAA alphabets contained in the 74 reduction methods and cluster according to application scenarios, and the shade of color indicates the number of reduced clusters. See the attachment for the full content.

Table 2
The 6 reduction categories of 74 reduction methods.

Categories	Reduction Alphabets	Reduced clusters	Cite
Clustering Algorithm	24	259	[55,63–71]
Mutation Matrix	20	239	[22,30–32,51,72–80]
Computer Method	12	60	[34,37,47,78,81,82]
Physical and Chemical Method	12	52	[36,37,48,61,78,83–89]
Information Theory	3	32	[90–92]
Statistical Analysis	3	30	[63,78,93]

people used early experimental data to analyze structural laws through machine learning methods, and tried to predict protein structure properties, such as intrinsic disorder, solvent accessibility and contact number.

Weathers et al. used hydrophilicity and hydrophobicity as the reduction rule for functional classification prediction of intrinsically disordered proteins, and pointed out that hydrophobic amino acids play a central role in stabilizing folded proteins in 2004 [36]. In 2006, Melo proposed the use of RAA alphabets to improve sequence alignment and protein folding accuracy [47]. They developed a new genetic algorithm to obtain a five-clusters reduction scheme based entirely on structural information, and supposed that the five-clusters-based reduction model also has good predictive performance in evaluating protein folding. In 2009, Bacardit et al. proposed a method for predicting protein structure contact number and solute accessibility on the basis of the mutual information reduced alphabet, and emphasized that the reduction well preserved the physicochemical properties of amino acid residues and improved the accuracy [37]. In 2020, Oberti et al. used a convolutional neural network based on simplified amino acid composition to predict the intrinsically disordered regions of proteins, and proposed that RAA alphabets help convolution to recognize complex patterns in sequences [40].

In recent years, the AlphaFold series created by Google DeepMind has raised the accuracy and efficiency of protein structure prediction to a new level based on a powerful artificial neural network architecture. With the support of AlphaFold structure database, a large number of protein structural properties analysis predictions continue to emerge. Recently, a protein structure analysis platform RaacFold based on RAA alphabets has been constructed. It combines RAA alphabets with the structural database predicted by AlphaFold2 and previous protein structure database, which provides users with a convenient protein structure and property analysis service by using different RAA alphabets [62]. The 3D rendering service of reduction structure properties provided by RaacFold enriched the application of RAA alphabets in the analysis of protein sequence and structural properties.

6. A comprehensive analysis of the 672 reduced amino acid alphabets

In recent years, we have collected a large number of RAA alphabets and achieved many excellent results in predicting protein functional classification by using these RAA alphabets. Based on our research work, 672 RAA alphabets from 74 reduction methods have been arranged, and annotated with the source and reduction method of each reduced alphabet in detail (Please refer to the supplementary file for full data). According to different principles, we

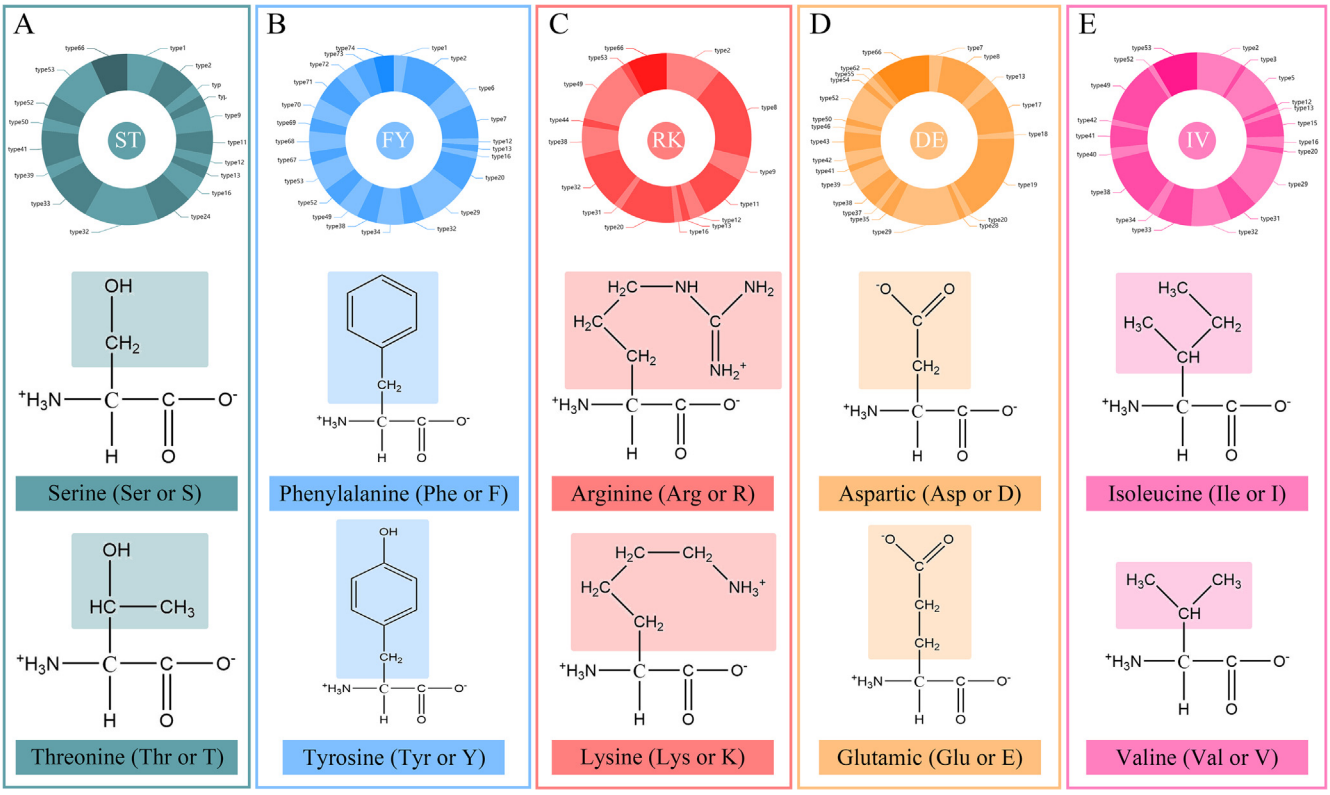

Fig. 3. Five high-frequency words and their structures. A: The word ST, which is composed of serine and threonine, is contained in 18 reduced methods and occurs 43 times, and their R groups are both polar OH^-. B: The word FY, which is composed of phenylalanine and tyrosine, is contained in 23 reduced methods and occurs 77 times, and their R groups are both phenyl rings. C: The word RK, which is composed of arginine and lysine, is contained in 15 reduced methods and occurs 66 times, and both of them contain amino groups in their R groups. D: The word IV, which is composed of isoleucine and valine, is contained in 20 reduced methods and occurs 94 times, and both of them have nonpolar R groups. E: The word DE, which is composed of aspartic and glutamic is contained in 23 reduced methods and occurs 77 times, and their R groups are both carboxyl groups.

summarize the 74 reduction methods into 6 types, namely Clustering Algorithm, Mutation Matrix, Computer Method, Physical and Chemical Method, Information Theory and Statistical Analysis (Fig. 2B and Table 2). Clustering Algorithm and Mutation Matrix are widely used in RAA research, accounting for more than half of the papers published in the past 20 years. Many RAA alphabets are still in use today (Fig. 2A).

We counted 672 RAA alphabets and the reduced sizes they contained (Fig. 2C), and classified them into seven categories according to the application scenarios of each reduced alphabet, namely protein folding, build reduced alphabets, functional classification, secondary structure prediction, sequence alignment, structure prediction, and protein interaction (Fig. 2D). Among all alphabets, Size2-Size5 has the largest proportion, which is related to the early results of a large number of RAA studies by Wang et al (Fig. 2C) [22].

However, with the development of research, a large number of research results pointed out that too small simplified alphabets can easily lead to a large loss of sequence information. Reduced alphabets of Size 10 and above perform better for most jobs while retaining the protein information [30,75]. Of the 672 RAA alphabets, nearly half of the alphabets have only been created and not put into specific research work. Most of the rest are devoted to protein alignment, folding, and functional structure prediction, laying a solid foundation for protein diversification analysis.

The combined frequencies of all words showed that the five words "ST", "FY", "RK", "DE" and "IV" were distributed more frequently (over 40 times) in most alphabets (Fig. 3). This means that these five words may be recognized by many researchers due to their similar properties in a lot of cases. For example, Wang's article points out that "DE" (Asp and Glu) can be reduced to one class by MJ matrix and contact potential, which is verified in Yu's article by a multi-species classification model, and the same reduction results are obtained in Mirny's article by structurally derived substitution matrices [22,84,86].

7. Conclusion

The research on the structure and function of proteins has been accelerating, and the methods and tools that have been kept in dust for many years have gradually shown their powerful advantages. Protein analysis and prediction methods based on machine learning improve analytical efficiency, achieve higher precision, and solve deeper biological problems.

As an important part of protein feature engineering, the reduction of amino acid alphabets has realized the redefinition of sequence and structure. It not only has strong inclusive power, allowing it to be used as an upstream processing step for almost all existing methods, but also provides the model with richer biological prior knowledge, which greatly optimizes the biological background of traditional computer models and is expected to decipher proteins under the complex structures.

In addition, it provides better solutions to problems such as the cumbersomeness and dimension explosion of current machine learning and artificial intelligence methods, and is more suitable for deployment on small and medium-sized computers and servers to reduce the computing pressure of equipment.

The current research results and evaluation criteria for RAA alphabets have not formed a set of recognized systems, and RAA alphabets have not been fully and maturely used in current research. Under the joint promotion of all researchers, simplified amino acid composition still has space for optimization and important significance in the new era, and the technology and platform based on RAA alphabets may still create higher and far-reaching value in the future.

Funding

This work was supported by the National Nature Scientific Foundation of China (No: 62171241, 62061034, 61861036), the Key Technology Research Program of Inner Mongolia Autonomous Region (2021GG0398), and the Science and Technology Major Project of Inner Mongolia Autonomous Region of China to the State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock (2019ZD031).

CRediT authorship contribution statement

Yuchao Liang: Writing - original draft, Investigation, Formal analysis. **Siqi Yang:** Investigation, Writing - review & editing. **Lei Zheng:** Software. **Hao Wang:** Writing - review & editing. **Jian Zhou:** Writing - review & editing. **Shenghui Huang:** Software, Investigation. **Lei Yang:** Writing - review & editing. **Yongchun Zuo:** Writing - Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Mingzhu Liu, Pengfei Liang and others for their contributions to the proofreading of the thesis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.001.

References

[1] Zhang Z, Wu S, Stenoien DL, Paša-Tolić L. High-throughput proteomics. *Annu Rev Anal Chem (Palo Alto Calif)* 2014;7:427–54.
[2] Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: technologies and their applications. J Chromatogr Sci 2017;55:182–96.
[3] Sonsare PM, Gunavathi C. Investigation of machine learning techniques on proteomics: A comprehensive survey. Prog Biophys Mol Biol 2019;149:54–69.
[4] Wen B, Zeng WF, Liao Y, Shi Z, Savage SR, Jiang W, et al. Deep learning in proteomics. *Proteomics* 2020;20:e1900335.
[5] Li C, Luo X, Qi Y, Gao Z, Lin X. A new feature selection algorithm based on relevance, redundancy and complementarity. Comput Biol Med 2020;119:103667.
[6] Zhao X, Zhang Y, Du X. DFpin: Deep learning-based protein-binding site prediction with feature-based non-redundancy from RNA level. Comput Biol Med 2022;142:105216.
[7] Li Z, Lin Y, Elofsson A, Yao Y. Protein contact map prediction based on ResNet and DenseNet. Biomed Res Int 2020;2020:7584968.
[8] David CC, Jacobs DJ. Principal component analysis: a method for determining the essential dynamics of proteins. Methods Mol Biol 2014;1084:193–226.
[9] Le TT, Urbanowicz RJ, Moore JH, McKinney BA. STatistical Inference Relief (STIR) feature selection. Bioinformatics 2019;35:1358–65.
[10] Liang P, Yang W, Chen X, Long C, Zheng L, Li H, et al. Machine learning of single-cell transcriptome highly identifies mRNA signature by comparing F-score selection with DGE analysis. Mol Ther Nucleic Acids 2020;20:155–63.
[11] Wirsing L, Klawonn F, Sassen WA, Lünsdorf H, Probst C, Hust M, et al. Linear discriminant analysis identifies mitochondrially localized proteins in Neurospora crassa. J Proteome Res 2015;14:3900–11.
[12] Zuo Y, Chang Y, Huang S, Zheng L, Yang L, Cao G. iDEF-PseRAAC: identifying the defensin peptide by using reduced amino acid composition descriptor. *Evol Bioinform Online* 2019;15:1176934319867088.
[13] Wang H, Xi Q, Liang P, Zheng L, Hong Y, Zuo Y. IHEC_RAAC: a online platform for identifying human enzyme classes via reduced amino acid cluster strategy. Amino Acids 2021;53:239–51.
[14] Zheng L, Huang S, Mu N, Zhang H, Zhang J, Chang Y, et al. RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. Database (Oxford) 2019;2019.

[15] Zhou J, Bo S, Wang H, Zheng L, Liang P, Zuo Y. Identification of disease-related 2-oxoglutarate/Fe (II)-dependent oxygenase based on reduced amino acid cluster strategy. Front Cell Dev Biol 2021;9:707938.
[16] Morita K, Simons ER, Blout ER. Polypeptides. 53. Water-soluble copolypeptides of L-glutamic acid, L-lysine, and L-alanine. Biopolymers 1967;5:259–71.
[17] Heinz DW, Baase WA, Matthews BW. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. Proc Natl Acad Sci U S A 1992;89:3751–5.
[18] Osawa S, Jukes TH, Watanabe K, Muto A. Recent evidence for evolution of the genetic code. Microbiol Rev 1992;56:229–64.
[19] Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, et al. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–9.
[20] Wolynes PG. As simple as can be? Nat Struct Biol 1997;4:871–4.
[21] Schafmeister CE, LaPorte SL, Miercke LJ, Stroud RM. A designed four helix bundle protein with native-like structure. Nat Struct Biol 1997;4:1039–46.
[22] Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–8.
[23] Miyazawa S, Jernigan RL. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng 1993;6:267–78.
[24] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89:10915–9.
[25] Mount DW. Using BLOSUM in sequence alignments. CSH Protoc 2008;2008. pdb.top39.
[26] Mount DW. Using PAM Matrices in Sequence Alignments. CSH Protoc 2008;2008. pdb.top38.
[27] Mount DW. Comparison of the PAM and BLOSUM amino acid substitution matrices. CSH Protoc 2008;2008. pdb.ip59.
[28] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–82.
[29] Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18:691–9.
[30] Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng Des Sel 2000;13:149–52.
[31] Kosiol C, Goldman N, Buttimore NH. A new criterion and method for amino acid classification. J Theor Biol 2004;228:97–106.
[32] Cannata N, Toppo S, Romualdi C, Valle G. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. Bioinformatics 2002;18:1102–8.
[33] Akanuma S, Kigawa T, Yokoyama S. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci U S A 2002;99:13549–53.
[34] Davies MN, Secker A, Freitas AA, Clark E, Timmis J, Flower DR. Optimizing amino acid groupings for GPCR classification. Bioinformatics 2008;24:1980–6.
[35] Cherkassky V. The nature of statistical learning theory∼. IEEE Trans Neural Netw 1997;8:1564.
[36] Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Lett 2004;576:348–52.
[37] Bacardit J, Stout M, Hirst JD, Valencia A, Smith RE, Krasnogor N. Automated alphabet reduction for protein datasets. BMC Bioinf 2009;10:6.
[38] Yang H, Huimin XU, Yan S, Chen J, Geng L, Yao Y, Univerdsity QB. Protein subcellular localization prediction based on reduced representation of amino acid and statistical characteristic. Chin J Bioinf 2015.
[39] Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. Mol Model Annu 2001;7:360–9.
[40] Oberti M, Vaisman II. cnnAlpha: Protein disordered regions prediction by reduced amino acid alphabets and convolutional neural networks. Proteins Struct Funct Bioinf 2020;88.
[41] Ye Y, Choi JH, Tang H. RAPSearch: a fast protein similarity search tool for short reads. BMC Bioinf 2011;12:159.
[42] Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 2011;28:125–6.
[43] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12:59–60.
[44] Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun 2018;9:2542.
[45] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35:1026–8.
[46] Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics 2021;37:3029–31.
[47] Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins 2006;63:986–95.
[48] Chen YL, Li QZ. Prediction of the subcellular location of apoptosis proteins. J Theor Biol 2007;245:775–83.
[49] Chen W, Feng P, Lin H. Prediction of ketoacyl synthase family using reduced amino acid alphabets. J Ind Microbiol Biotechnol 2012;39:579–84.
[50] Feng PM, Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem 2013;442:118–25.
[51] Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, et al. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS ONE 2014;9:e106691.
[52] Zuo Y-C, Li Q-Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides 2009;30:1788–93.
[53] Feng P, Lin H, Chen W, Zuo Y. Predicting the types of J-proteins using clustered amino acids. Biomed Res Int 2014;2014:935719.
[54] Zuo Y, Lv Y, Wei Z, Yang L, Li G, Fan G. iDPF-PseRAAAC: a web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. PLoS ONE 2015;10:e0145541.
[55] Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics 2018;34:2740–7.
[56] Shimizu K, Hirose S, Noguchi T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics 2007;23:2337.
[57] Zuo Y, Li Y, Chen Y, Li G, Yan Z, Yang L. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics 2017;33:122–4.
[58] Xi B, Tao J, Liu X, Xu X, He P, Dai Q. RaaMLab: A MATLAB toolbox that generates amino acid groups and reduced amino acid modes. Biosystems 2019;180:38–45.
[59] Zheng L, Liu D, Yang W, Yang L, Zuo Y. RaacLogo: a new sequence logo generator by using reduced amino acid clusters. Brief Bioinform 2021;22.
[60] Zhang H, Xi Q, Huang S, Zheng L, Yang W, Zuo Y. iSP-RAAC: identify secretory proteins of malaria parasite using reduced amino acid composition. Comb Chem High Throughput Screen 2020;23:536–45.
[61] Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 2006;34:W32–7.
[62] Zheng L, Liu D, Li YA, Yang S, Liang Y, Xing Y, et al. RaacFold: a webserver for 3D visualization and analysis of protein structure by using reduced amino acid alphabets. Nucleic Acids Res 2022.
[63] Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 2007;36:1059–69.
[64] Jardin C, Stefani AG, Eberhardt M, Huber JB, Sticht H. An information-theoretic classification of amino acids for the assessment of interfaces in protein-protein docking. J Mol Model 2013;19:3901–10.
[65] Li J, Wang W. Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. Sci China C Life Sci 2007;50:392–402.
[66] Sneath PH. Relations between chemical structure and biological activity in peptides. J Theor Biol 1966;12:157–95.
[67] Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 2005;102:6395–400.
[68] Stanfel LE. A new approach to clustering the amino acids. J Theor Biol 1996;183:195–205.
[69] Adamian L, Liang J. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. J Mol Biol 2001;311:891–907.
[70] Li X, Hu C, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins 2003;53:792–805.
[71] Georgiou DN, Karakasidis TE, Nieto JJ, Torres A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 2009;257:17–26.
[72] Prlić A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–50.
[73] Liu X, Liu D, Qi J, Zheng WM. Simplified amino acid alphabets based on deviation of conditional probability from random background. Phys Rev E Stat Nonlin Soft Matter Phys 2002;66:021906.
[74] Pape S, Hoffgaard F, Hamacher K. Distance-dependent classification of amino acids by information theory. Proteins 2010;78:2322–8.
[75] Shepherd SJ, Beggs CB, Jones S. Amino acid partitioning using a Fiedler vector model. Eur Biophys J 2007;37:105–9.
[76] Susko E, Roger AJ. On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol 2007;24:2139–50.
[77] Tanping, Li, Ke, Fan, Jun, Wang, Wei. Reduction of protein sequence complexity by residue grouping. Protein Eng Wang 2003.
[78] Stephenson JD, Freeland SJ. Unearthing the root of amino acid similarity. J Mol Evol 2013;77:159–69.
[79] Cieplak M, Holter NS, Maritan A, Banavar JR. Amino acid classes and the protein folding problem. J Chem Phys 2001.
[80] Esteve JG, Falceto F. A general clustering approach with application to the Miyazawa-Jernigan potentials for amino acids. Proteins 2004;55:999–1004.
[81] Smith RF, Smith TF. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc Natl Acad Sci U S A 1990;87:118–22.
[82] Zhang H, Kurgan L. Improved prediction of residue flexibility by embedding optimized amino acid grouping into RSA-based linear models. Amino Acids 2014;46:2665–80.
[83] Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A 1996;93:11628–33.
[84] Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–96.

[85] Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–88.
[86] Yu ZG, Anh V, Lau KS. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J Theor Biol 2004;226:341–8.
[87] Han P, Zhang X, Norton RS, Feng ZP. Predicting disordered regions in proteins based on decision trees of reduced amino acid composition. J Comput Biol 2006;13:1723–34.
[88] Ilardo MA, Freeland SJ. Testing for adaptive signatures of amino acid alphabet evolution using chemistry space. *J Syst Chem,5,1(2014-01-21)* 2014;5:1.
[89] Andersen CA, Brunak S. Representation of protein-sequence information by amino acid subalphabets. *AI Mag* 2004;25:97-97.
[90] Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins 2000;38:149–64.
[91] Solis AD. Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. Proteins 2015;83:2198–216.
[92] Robson B, Suzuki E. Conformational properties of amino acid residues in globular proteins. J Mol Biol 1976;107:327–56.
[93] Wrabl JO, Grishin NV. Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. Proteins 2005;61:523–34.