

Protein Structure Prediction: Biological Basis, Processing Methods and Deep Learning

Jingkai Wen¹^a

College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

Keywords: Protein Structure Prediction, Processing Methods, Deep Learning.

Abstract: As the speed of finding new proteins exceeds that of structural analysis, traditional experimental ways are time-consuming and cannot meet the need to decipher the structure of proteins in a relatively short time, which leads to the appearance of protein structure prediction. Protein structure prediction uses deposited protein structures to predict the newfound and has developed fast with the increase of computational resources and the refinement of algorithms. This review introduces protein structure prediction based on machine learning, including sequence encoding and feature extraction. After that, we focus on deep learning and interpret several common methods used in deep learning algorithms, including sequence alignment, residues contact profile. Finally, we introduce several representative algorithms and their methods.

1 INTRODUCTION

Protein Structure prediction is crucial to understand the protein function and inter-protein contact. Traditionally, scientists use experimental methods to analyse the three-dimensional (3D) structure, like X-ray Crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and Cryogenic Electron Microscopy (Cryo-EM) (<https://www.genome.jp/dbget/aaindex.html>); (Anand, 2008); (Anfinsen) But now, the gap between protein sequences and know structures is getting bigger (Bateman, 2021). Meanwhile, the speed of discovering new proteins surpass that of analysing, so it is needed to predict the structure of new-finding proteins based on what we have known. Proteins have infinite patterns of structure, but the basic forming elements are conservative among all species. Also, according to self-assembly theory, only the amino acid residue is adequate to model the final 3D structure (Bengio, 1994). That's the theoretical base for computational prediction. Many methods were created to predict 3D structures based on sequences. According to Critical Assessment of Protein Structure Prediction (CASP), protein structure prediction can be divided into two categories, template-based modeling (TBM) and free modeling


(FM) (Bernardes, 2013). TBM compares target sequences with those in Protein Database (PDB) (<https://www.rcsb.org>) and finds homologous fragments, then takes known motifs together and thread several parts to present the whole 3D structure. FM, also called ab initio prediction, predict the target sequence based on inter-residues interaction and evolutionary relationship.

The Dictionary of Protein Secondary Structures defines eight states of a single amino acid residue to make the sequence easier to be processed by the procedure, so the algorithm can easily categorize residues. Different methods are applied to process the sequence and get information (Bonetta, 2020).

2 PROTEIN STRUCTURES

2.1 Primary Structure

The primary structure of proteins refers to the sequence of amino acids in the polypeptide chains, like ACDE, which is determined by the sequence of DNA. After transcription and translation, the genetic information is transformed from DNA to mRNA and finally to protein (Cai, 2000). Each amino acid is joined by peptide bonds, formed by dehydration

^a <https://orcid.org/0000-0003-1252-5843>

between amino groups and carboxyl groups. In structure prediction, 20 amino acids are encoded with 20 letters, so the input to the algorithm is actually character strings. The first protein deciphered was insulin. Frederick Sanger discovered its amino acid sequence in 1951 and brought up that proteins have defining amino acid sequences (Chou, 1995).

2.2 Secondary Structure

Secondary structure refers to regular local sub-structures defined by the patterns of hydrogen bonds. The one-dimensional sequence can form three dimensional local segments through the hydrogen bond between amino and carboxyl oxygen. Because secondary structures are elements for protein folding, prediction from sequence to local segments is critical and therefore challenging. Pauling assigned secondary structures to eight types based on hydrogen bonding patterns. For convenient expression and encoding, The Dictionary of Protein Secondary Structures (DSSP) is commonly used to describe secondary structures with corresponding eight letters (Bonetta, 2020).

2.3 Tertiary Structure

Tertiary structure refers to the three-dimensional structure of a single protein folding with one or several domains driven by non-specific hydrophobic interaction. Tertiary structure is basically the spatial arrangement of multiple secondary structures, sometimes accompanied by metal ions.

Therefore, models for structure prediction actually try to extract local features from the amino acid sequence (input) and transform them into octet-state secondary structures, and then assemble elements into three-dimensional protein structures.

Table 1. The Dictionary of Protein Secondary Structures.

Code	Secondary Structure
G	3_{10} helix (3-turn helix)
H	\square helix
I	\square helix
T	hydrogen-bonded turn
E	extended strand in parallel and/or anti-parallel β -sheet
B	Isolated β -bridge
S	Bend
C	Coil

3 SELF-ASSEMBLY THEORY

In 1961, Anfinsen treated bovine pancreatic ribonuclease with 8 M urea and got a randomly coiled polypeptide chain with cysteine residues. Then he found that though the ribonuclease was denatured, it could regain the activity under optimal conditions of polypeptide concentration and pH (Bengio, 1994). Therefore, he proposed that proteins have their natural structures, determined by one-dimensional sequence, and peptide chains will automatically fold into that conformation. That's the self-assembly theory. It is the basis for protein structure prediction because the theory assumes that no other variables are influencing the final structure of proteins except for the sequence.

4 FEATURE EXTRACTION

After the input of sequence, a specific encoding scheme is needed to generate a set of features to represent the properties of each protein and use those features as input to machine learning (ML) algorithms (Chou, 2020). In the past 30 years, scientists have developed a number of different descriptors of proteins for different aspects of prediction, including fold classification, subcellular location prediction and membrane protein type prediction (Chou, 2001); (Chou, 2019). Descriptors are designed to show some information of the proteins, like isoelectric point (pI), amino acid residues composition. Here we introduce some strategies to extract protein features.

4.1 Amino Acid Composition

Proteins are composed of amino acid residues whose arrangement determines how proteins will fold. So basically, amino acid composition (AAC) help to find a specific spatial structure. Originally, AAC was utilized as a feature descriptor. Based on sequence, a vector with 20 elements is yielded, and each element represents the frequency of a specific amino acid residue (Comet, 2002); (Dayhoff, 1983); (Deng, 2018). However, it was found that only the AAC descriptor cannot simulate the spatial structure, and the bias is inevitable. Scientists think some important information may be neglected, so Zhou proposed the concept of pseudo amino acid composition (PseAAC) (Ding, 2013); (Dubchak, 1995). It is just some additional digital information adding to the

original 20 elements, yielding a $20 + \lambda$ -dimension vector:

$$X = [x_1, x_2, \dots, x_{20}, x_{20+1}, \dots, x_{20+\lambda}]$$

The factors x_1, x_2, \dots, x_{20} are frequencies of 20 natural amino acids, and the factors $x_{20+1}, \dots, x_{20+\lambda}$ are the information along the sequence as complementary input. PseAAC includes hydrophobicity, hydrophilicity, etc. With the development of analysis, more PseAACs showed up, representing more important information (Dubchak, 1999); (Eddy, 2002). Free resources containing 63 different kinds of PseAACs can be accessed through a web server named "PseAAC", which was established by Shen and Chou (<https://www.csbio.sjtu.edu.cn/bioinf/PseAAC/>) (Edgar, 2004); (Elmlund, 2015).

4.2 Sequence Order

Lin and Li take two adjacent amino acid residues as an unit and use it to predict secondary structure (Gondro, 2007); (Gonzalez-Lopez, F.). For example, dipeptide composition is utilized, which means it yields 400 possible theoretical arrangements. The output is then a vector containing the occurrence frequency of a combination of specific two residues. Similarly, Yu and coworkers use dipeptide, tripeptide, and tetrapeptide elements to promote modeling accuracy (Hall, 1964); (Hanson). What's more, they convert polypeptide composition into a structural class tendency sequence, which was then used as a new feature descriptor. Tetrapeptide arrangement predicts regular structures, for i -th residue usually interact with $i+4$ -th residue (Hanson).

4.3 Physicochemical Properties

Each amino acid has its specific side chain, which determines its physicochemical properties, like isoelectric point, polarizability and hydrophobicity (Dayhoff, 1983); (Henikoff, 1992); (Hornak, 2006). Those properties are accessible online at the amino acid index database (<https://www.genome.jp/dbget/aaindex.html>) (Hornak, 2006); (Ilonen, 2003). Chou extracted information from physicochemical properties with a set of correlation factors, yielding PseAAC descriptors. Using different functions, sequence order correlation factors can be calculated.

According to global protein sequence descriptors (GPSD) theory, amino acids are classified according to their unique properties (Henikoff, 1992). GPSD includes three dimensions: composition, transition

and distribution. Same as AAC, the composition means the occurrence frequency of each amino acid residue type. The transition means frequencies that a specific type of amino acid changes to another one. The distribution is position-specific information, showing the distribution of each amino acid residue along the sequence.

4.4 Secondary-structure-based Features

According to DSSP above, each amino acid has its tendency to appear in one or more secondary structures. So the protein sequence can be converted into secondary structure descriptors by extracting information with GPSD. Helix (H), strand (E) and coil (C) are commonly utilized (Jararweh, 2019). Also, this method can be integrated with several methods above to improve accuracy. For example, Secondary structure features help to calculate the correlation factors or yield PseAAC (Jumper, J.); (Kabsch, 1983).

5 SEQUENCE ALIGNMENT

One letter represents a specific amino acid during the prediction, and the input is a character string. The algorithm cannot get the information about the character of each amino acid, so the first thing to do is to find if deposited proteins are containing the same sequence fragments. To achieve this, the target sequence must be compared with every sequence in the database once, which needs the application of Pairwise Alignment (PA). If there are some counterparts in the database, the algorithm will just take the known structure of counterparts to construct the model, based on the self-assembly rule proposed by Christian B. Anfinsen. If there is no sequence counterpart in the database, then we have to apply the algorithm to make an ab initio prediction, which is normally Multiple Sequence Alignment (MSA). Through MSA, we can know the evolutionary relationship between the target sequence and existing sequences and then reason probable 3D structure.

5.1 Pairwise Sequence Alignment (PSA)

Pairwise sequence alignment (PSA) is a method assessing the similarity between two sequences. The classic method is Dynamic Programming, also named the Needleman-wunsch (NW) algorithm

(Kawashima, 2000). By using the trace-back process, DP can provide optimum alignment and predict the objective function (Kinch, 2016). When the input contains two sequences, which is the simplest case, DP builds an $i \times j$ matrix based on two sequences (i, j is the sequence length). Each position in the matrix has a score, representing the similarity of the row and the column. Finding the path with the highest scores can get an alignment pattern of two sequences (Kurgan, 2007). However, DP is time-consuming and requires high computation resources. The complexity grows exponentially along with the increase in sequence length, not to mention that two or more optimal paths are available. Moreover, DP suffers from high-dimensional problems in multiple sequence alignment. Even if the computation resources are adequate, the optimum alignment from DP is rarely biologically optimum. For high accuracy of prediction, protein sequence alignment substitution matrices were established (Landan, 2009). The substitution scoring matrix includes PAM and BLOSUM, (Landan, 2009); (Lecun, 1998). Jararweh and co-workers improved the needleman-wunsch algorithm by applying three sets of parallel implementations. They utilized three hardware solutions: POSIC Threads-based, SIMD Extensions-based and GPU-based implementations (Lewicki, 2003).

5.2 Multiple Sequence Alignment (MSA)

MSA aligns multiple sequences, which are normally related, to get more biological information, like the estimation of evolutionary divergence and ancestral sequence profiling (Lin, 2013). It is normally implemented when we want to know the evolutionary trace of target proteins when it comes to protein prediction. By simulating the process of evolution with MSA, we hope to reason the possible structure of the target sequence right now. To achieve it, substitution matrix and scoring were assessed (Lin, 2007). DP can work on MSA if the sequence number is small, but it cannot be applied on hundreds of sequences for the spur of complexity (Liu, 2020). Based on that situation, scientists create heuristic algorithms, sacrificing some accuracy for higher computational efficiency. For example, MUSCLE, CLUSTAL and T-COFFEE use progressive alignment (Moult, 1995); (Moult, 2007), and MUMMALS and PROMALS use hidden Markov model based alignments (Nanni); (Needleman, 1970). Comet and Henry present a method that integrates other information to the classical dynamic

programming algorithms, like the pattern of PROSITE (Notredame, 1998).

6 CRITICAL ASSESSMENT OF PROTEIN STRUCTURE PREDICTION

Critical Assessment of Protein Structure Prediction (CASP) is a biennial event that assesses the model of protein structure prediction, held firstly in 1994. For every CASP competition, participants are asked to construct models within a stipulated time to predict the structure of the given protein sequences. Participants do not know the actual structure of the sequences, but the actual structure was determined by experiments before (Pei, 2007). After modeling, the predicted structure will be compared to the counterpart from the experiments, and the similarity will be evaluated. There are two categories for modeling since CASP7, free modeling (FM) and template based modeling (TBM) (Pei, 2008). TBM predicts those targets whose homologies, which shares similar sequences, have been deposited in PDB. The FM category is a big challenge for low prediction accuracy, and fragment-based approaches, like Rosetta, I-TASSER, and QUARK, dominated CASP for many years until deep learning was introduced

(<https://www.csbio.sjtu.edu.cn/bioinf/PseAAC/>).

6.1 Deep Learning

Deep learning, as a sub-field of machine learning, is based on artificial neural networks. It was introduced in predicting protein structure in 2016 because contact prediction, a key intermediate step for prediction, emerged in CASP12. A deep learning-based method, Raptor-X, showed up and reached about 50% precision when evaluating top L/5 long-range predictions, which was twice as much precision in CASP11 (Qian, 1988). After CASP12, an improved version of the Raptor-X and open-source deep learning method DNCON2 were released. In CASP13, AlphaFold and other high-performing methods were upgraded to use 'distogram' rather than just contacts (Qin, 2015); (Rumelhart, 1986).

Deep learning (DL) simulates biological neural networks. Deep learning uses multiple connected layers to transform input into corresponding output. To some extent, Deep learning is the advancement of the Feed Forward Neural Network (FFNN) (Sadique, 2020); (Sanger, 1951). FFNN is an artificial neural

network system that contains no cycles, dividing nodes into groups (layers) and process the input through them. Each node has its parameter and weight and all the nodes in the same layer calculate a vector. The layer $i+1$ gets its values exclusively from layer i , until the output layer is valued. The FFNN model can be trained with examples by propagation algorithm and have universal approximation properties, which has been proven (Shen, 2008); (Simpkin, A.J). Typically, the "windowed" version is applied to protein prediction. A fixed number of amino acids as a segment is regarded as the input, and the target is the PSA of the segment. Based on FFNN, the Deep Learning method varies the connectivity between the layers, allowing the algorithm to be applied to different data types. Like FFNN, Deep Learning can be trained with many examples by backpropagation automatically (Smyth, 2000). However, containing numbers of internal parameters leads to data-greedy and large samples required. Two mainstream DL algorithms are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

6.2 Convolutional Neural Networks

Convolutional Neural Networks is designed to process spatial dependent data (like a base pair in the DNA sequence). Taking this advantage, CNN layers apply local convolutional filters on positions in the data. This strategy avoids the overfitting problem and is translation invariant. A module of CNN contains multiple consecutive layers to encode more complex features (Wang, 2017). "Windowed" FFNN is a specific, shallow version of CNN, so we keep referring to CNN as FFNN.

6.3 Recurrent Neural Networks

Recurrent Neural Networks aim to learn global features from sequential data. RNN modules use parameterized sub-module to process the sequence into an internal state vector and use the vector to summarize the original sequence. Previous state vector and current input elements determine the current internal state vector (Wang, 2018). However, RNN easily suffers from the gradient vanishing or gradient explosion problem because RNN keeps using the same function repeatedly (Xu, 2019). Then Long Short Term Memory, Gated Recurrent Unit and other Gated RNN are created to contain the problems.

6.4 Alphafold

AlphaFold (AlphaFold v2.0) outperform other methods in the CASP14 TM group. Its central essence is the CNN method (Qin, 2015). AlphaFold implements a distogram for protein structure prediction by constructing very deep residual neural networks with 220 residual blocks processing a representation of dimensionality $64 \times 64 \times 128$. By customizing the length of the given sequence and plot an $L \times L$ map, AlphaFold can predict inter-residue distances for sub-regions. Then the map is transformed into 3D models using minimized distance potential, which implements sampling and gradient-descent-based methods. Trained by the proteins in PDB, it reaches high accuracy even for targets with fewer homologous sequences.

AlphaFold contains two main stages: the trunk of the network and the structure module. After the input of amino acid sequence, the trunk of the network process it through multiple repeated layers of neural network block (named Evoformer) and produce two arrays. $N_{seq} \times N_{res}$ array represents a processed multiple sequence alignment (MSA), achieved by aligning the sequence with those of other species in Protein Database (PDB) and outputting a matrix representing the similarity between the target and deposited sequence. On the other hand, Evoformer also produces a $N_{res} \times N_{res}$ array. By assessing the interaction of every two amino acid residues, the $N_{res} \times N_{res}$ array can show residue pairs that likely attract or repulse each other, then predict the three-dimensional position of residues. The innovative point of Evoformer is a new mechanism for information exchange within MSA and pair representation. It helps to reason spatial and evolutionary relationship.

Besides, AlphaFold utilizes end-to-end structure prediction with pair representation. By assuming each residue is a free-floating rigid body, the module constructs residue gas. Residue gas represents 3D backbone structure as the form of N_{res} independent rotations and translations, which prioritize the orientation of the C backbone, so the side chains are highly constrained. Meanwhile, the C backbone is completely unconstrained, and the network frequently violates the chain constraint, hoping to find the conformation with minimum global energy. Finally, exact enforcement of peptide geometry is completed through post-prediction relaxation with gradient descent in the Amber force field (Ye, 2011); (Yu, 2007).

6.5 RaptorX-Contact

RaptorX-Contact was created for contact map prediction by Xu's group based on Deep Learning. It uses a model called deep ResNet, containing two major residual neural network modules, respectively called 1D deep ResNet and 2D deep dilated ResNet (Qian, 1988); (Qin, 2015); (Zhang, T.-L., Y.-S. Ding, and K.-C. Chou). The 1D and 2D ResNets play different roles, respectively capturing long-range sequential and pairwise context. 1D ResNet extracts sequential features and conducts 1D convolutional transformations from a $L \times 26$ matrix, as the input, into a $L \times n$ matrix (L is sequence length). After converting the $L \times n$ matrix into a $L \times L \times 3n$ pairwise feature matrix, 2D feature is obtained, derived from 1D feature. It merges with a $L \times L \times 3$ pairwise feature matrix, forming the output. The output from the 1D ResNet is then fed into 2D ResNet. 2D ResNet is a Residual Neural Network module, conducting 2D convolutions. It transforming $L \times L \times (3 + 3n)$ matrix into a $L \times L$ predicted distance matrix by softmax. Eventually, the output from the 2D module is fed into logistic regression. The 1D and 2D ResNets contains of ~ 7 and ~ 60 convolutional layers and kernel size of 15 and 5×5 , respectively. The 1D and 2D ResNets for 1D and 2D feature learning is a calculative way to save computational resources. That's also why RaptorX-Contact produced better results in CASP11, comparing with other existing approaches.

For classification, interatomic distances are discretized into 25 bins: <4.5 , $4.5-5$, $5-5.5$, ..., $15-15.5$, $15.5-16$, and >16 Å, and treat each bin as a label. Contact prediction is achieved by summing up the probability of all the $C_\beta-C_\beta$ distance bins that fall into interval $[0, 8$ Å].

In CASP12, average long-range contact prediction accuracy of RaptorX Postdict in L, L/2, L/5, L/10 are respectively 40.18%, 50.20%, 58.87%, 63.93%, ranking the top. In CASP13 RaptorX-Contact also did the best, when top L, L/2, and L/5 long-range predicted contacts are evaluated, on the FM targets it has precision 44.731%, 57.787%, and 70.054%, respectively, and F1 values 0.411, 0.362, and 0.233, respectively.

7 CONCLUSION

With the advancement of computation and experiments, we will rely more on algorithm prediction to predict more protein structures. Frankly, high prediction accuracy of AlphaFold on FM illustrates the power of deep learning, and contact

profile and “distogram” are also huge successes and promote the accuracy of modeling. But in some way, it also embodies the deficiency we have in the theoretical field of structural biology.

Proteins with poor alignment are still hard to determine the 3D structure. For TBM, deficient sequence alignment means the difficulties to do homologous modeling, which for FM it means poor evolutionary relationship accessible. That is to say, we cannot fully get rid of empiricism at present. Thus, more sufficient methods for crystallographic structural analysis are needed to abate the gap between known proteins and those experimentally determined. Also, some sequences with low similarities share with a similar structure, so we need to find out what really matters to determine secondary structure, or we can define new feature descriptors. Thus, more factors that influence the spatial bias are remained to find. On top of that, new deep learning methods are indeed successful in prediction, but integrating with other traditional methods is still a challenge. It is necessary to build a comprehensive model to maximize the accuracy rather than approach the real model through different aspects. We cannot know if the prediction is well enough unless cross-validate with PDB because some steps in protein folding are still remained to discover. That's why FM is still way behind TBM.

In the future, more details of the protein folding process are needed to reduce the fluctuation of modeling. For now we are still not clear about how domains assemble together dynamically. We have to know more about the intermediate steps within. Algorithms need more training and refinement, basically helped by the database. More advanced neural network is still remained to discover.

REFERENCES

- Amino Acid Index Database. Available from: <https://www.genome.jp/dbget/aaindex.html>.
- Anand, A., G. Pugalenti, and P.N. Suganthan, Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. *Journal of Theoretical Biology*, 2008. 253(2):
- Anfinsen, C.B., et al., Kinetics of formation of Native Ribonuclease during oxidation of reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 1961. 47(9): p. 1309-1314.
- Bateman, A., et al., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 2021. 49(D1): p. D480-D489.

- Bengio, Y., P. Simard, and P. Frasconi, Learning long-term dependencies with gradient descent is difficult. *Ieee Transactions on Neural Networks*, 1994. 5(2): p. 157-166.
- Bernardes, J.S. and C.E. Pedreira, A review of protein function prediction under machine learning perspective. *Recent patents on biotechnology*, 2013. 7(2): p. 122-41.
- Bonetta, R. and G. Valentino, Machine learning techniques for protein function prediction. *Proteins-Structure Function and Bioinformatics*, 2020. 88(3): p. 397-413.
- Cai, L.M., et al. Evolutionary computation techniques for multiple sequence alignment. in *2000 Congress on Evolutionary Computation (CEC2000)*. 2000. La Jolla, Ca.
- Chou, K.C. and C.T. Zhang, Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, 1995. 30(4): p. 275-349.
- Chou, K.-C., An insightful 20-year recollection since the birth of pseudo amino acid components. *Amino Acids*, 2020. 52(5): p. 847-847.
- Chou, K.C., Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins-Structure Function and Genetics*, 2001. 43(3): p. 246-255.
- Chou, K.-C., RETRACTION: WITHDRAWN: An insightful recollection for predicting protein subcellular locations in multi-label systems. *Genomics*, 2019.
- Comet, J.P. and J. Henry, Pairwise sequence alignment using a PROSITE pattern-derived similarity score. *Computers & Chemistry*, 2002. 26(5): p. 421-436.
- Dayhoff, M.O., W.C. Barker, and L.T. Hunt, Establishing homologies in protein sequences. *Methods in Enzymology*, 1983. 91: p. 524-545.
- Deng, H.Y., Y. Jia, and Y. Zhang, Protein structure prediction. *International Journal of Modern Physics B*, 2018. 32(18).
- Ding, S.F., et al., Evolutionary artificial neural networks: a review. *Artificial Intelligence Review*, 2013. 39(3): p. 251-260.
- Dubchak, I., et al., Prediction of protein-folding class using global description of amino-acid sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 1995. 92(1).
- Dubchak, I., et al., Recognition of a protein fold in the context of the SCOP classification. *Proteins-Structure Function and Bioinformatics*, 1999. 35(4): p. 401-407.
- Eddy, S.R., A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *Bmc Bioinformatics*, 2002. 3.
- Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 2004. 32(5): p. 1792-1797.
- Elmlund, D. and H. Elmlund, Cryogenic Electron Microscopy and Single-Particle Analysis, in *Annual Review of Biochemistry*, Vol 84, R.D. Kornberg, Editor. 2015. p. 499-517.
- Gondro, C. and B.P. Kinghorn, A simple genetic algorithm for multiple sequence alignment. *Genetics and Molecular Research*, 2007. 6(4): p. 964-982.
- Gonzalez-Lopez, F., et al. End-to-end prediction of protein-protein interaction based on embedding and recurrent neural networks. in *IEEE International Conference on Bioinformatics and Biomedicine*
- Hall, L.D., Nuclear magnetic resonance. *Advances in Carbohydrate Chemistry*, 1964. 19: p. 51-93.
- Hanson, J., et al., Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 2019. 35(14): p. 2403-2410.
- Henikoff, S. and J.G. Henikoff, Amino-acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 1992. 89(22): p. 10915-10919.
- Hornak, V., et al., Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics*, 2006. 65(3): p. 712-725.
- Ilonen, J., J.K. Kamarainen, and J. Lampinen, Differential evolution training algorithm for feed-forward neural networks. *Neural Processing Letters*, 2003. 17(1): p. 93-105.
- Jararweh, Y., et al., Improving the performance of the needleman-wunsch algorithm using parallelization and vectorization techniques. *Multimedia Tools and Applications*, 2019. 78(4): p. 3961-3977.
- Jumper, J., et al., Highly accurate protein structure prediction with AlphaFold. *Nature*: p. 12.
- Kabsch, W. and C. Sander, Dictionary of Protein Secondary Structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 1983. 22(12): p. 2577-2637.
- Kawashima, S. and M. Kanehisa, AAindex: Amino acid index database. *Nucleic Acids Research*, 2000. 28(1): p. 374-374.
- Kinich, L.N., et al., Evaluation of free modeling targets in CASP11 and ROLL. *Proteins-Structure Function and Bioinformatics*, 2016. 84: p. 51-66.
- Kurgan, L. and K. Chen, Prediction of protein structural class for the twilight zone sequences. *Biochemical and Biophysical Research Communications*, 2007. 357(2): p. 453-460.
- Landan, G. and D. Graur, Characterization of pairwise and multiple sequence alignment errors. *Gene*, 2009. 441(1-2): p. 141-147.
- Lecun, Y., et al., Gradient-based learning applied to document recognition. *Proceedings of the Ieee*, 1998. 86(11): p. 2278-2324.
- Lewicki, G. and G. Marino, Approximation by superpositions of a sigmoidal function. *Zeitschrift Fur Analysis Und Ihre Anwendungen*, 2003. 22(2): p. 463-470.
- Lin, C., et al., Hierarchical Classification of Protein Folds Using a Novel Ensemble Classifier. *Plos One*, 2013. 8(2).

- Lin, H. and Q.-Z. Li, Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components. *Journal of Computational Chemistry*, 2007. 28(9):
- Liu, M.-L., et al., Predicting Preference of Transcription Factors for Methylated DNA Using Sequence Information. *Molecular Therapy-Nucleic Acids*, 2020. 22: p. 1043-1050.
- Moult, J., et al., A LARGE-SCALE EXPERIMENT TO ASSESS PROTEIN-STRUCTURE PREDICTION METHODS. *Proteins-Structure Function and Genetics*, 1995. 23(3): p. R2-R4.
- Moult, J., et al., Critical assessment of methods of protein structure prediction - Round VII. *Proteins-Structure Function and Bioinformatics*, 2007. 69: p. 3-9.
- Nanni, L., S. Brahnam, and A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *Journal of Theoretica*
- Needleman, S.B. and C.D. Wunsch, A general method applicable to search for similarities in amino acid sequence of 2 proteins. *Journal of Molecular Biology*, 1970. 48(3): p. 443-453.
- Notredame, C., L. Holm, and D.G. Higgins, COFFEE: an objective function for multiple sequence alignments. *Bioinformatics (Oxford, England)*, 1998. 14(5): p. 407-22.
- Pei, J.M. and N.V. Grishin, PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, 2007. 23(7): p. 802-808.
- Pei, J.M., B.H. Kim, and N.V. Grishin, PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, 2008. 36(7): p. 2295-2300.
- PseAAC. Available from: <https://www.csbio.sjtu.edu.cn/bioinf/PseAAC/>.
- Qian, N. and T.J. Sejnowski, Predicting the secondary structure of globular-proteins using neural network models. *Journal of Molecular Biology*, 1988. 202(4): p. 865-884.
- Qin, Y., et al., Prediction of protein structural class based on Linear Predictive Coding of PSI-BLAST profiles. *Open Life Sciences*, 2015. 10(1): p. 529-536.
- Rumelhart, D.E., G.E. Hinton, and R.J. Williams, Learning representations by back-propagating errors. *Nature*, 1986. 323(6088): p. 533-536.
- Sadique, N., et al., Image-based effective feature generation for protein structural class and ligand binding prediction. *Peerj Computer Science*, 2020.
- Sanger, F. and H. Tuppy, The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 1951. 49(4): p. 463-481.
- Shen, H.-B. and K.-C. Chou, PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry*, 2008. 373(2): p. 386-388.
- Simpkin, A.J., et al., Evaluation of model refinement in CASP14. *Proteins-Structure Function and Bioinformatics*.
- Smyth, M.S. and J.H.J. Martin, x Ray crystallography. *Journal of Clinical Pathology-Molecular Pathology*, 2000. 53(1): p. 8-14.
- Wang, S., et al., Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *Plos Computational Biology*, 2017. 13(1): p. 34.
- Wang, S., S.Q. Sun, and J.B. Xu, Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins-Structure Function and Bioinformatics*, 2018. 86: p. 67-77.
- Xu, J.B. and S. Wang, Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins-Structure Function and Bioinformatics*, 2019. 87(12): p. 1069-1081.
- Ye, X., G. Wang, and S.F. Altschul, An assessment of substitution scores for protein profile-profile comparison. *Bioinformatics*, 2011. 27(24): p. 3356-3363.
- Yu, T., et al., Structural class tendency of polypeptide: A new conception in predicting protein structural class. *Physica a-Statistical Mechanics and Its Applications*, 2007. 386(1): p. 581-589.
- Zhang, T.-L., Y.-S. Ding, and K.-C. Chou, Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. *Journal of Theoretical Biology*, 2008. 250(1): p. 186-193.