Research on protein complex recognition method based on genetic algorithm

Zhiqiang Liu[1,*] and Yulan Tang[1]

[1] *Chongqing Sino Konjac Biotechnology Co., LTD, Chongqing, China*

Abstract

In the rapid development of social economy and science and technology, how to study the interaction between proteins to complete the law of life activities has been one of the hot topics in academic research. Since protein interaction network is a typical complex network, which directly presents a significant social structure, and the corresponding functional modules are regarded as protein complexes, the corresponding identification and research methods have a positive role in predicting protein functions and explaining specific biological processes. Therefore, after understanding the research status of protein interaction network at home and abroad, this paper clarifies the protein complex recognition algorithm based on genetic algorithm according to the basic concept of protein interaction network, and then verifies the application value of the algorithm through experimental research, so as to provide reference for the application of the algorithm in the new era.

Key words

genetic algorithm; Protein complex; Interaction; Select strategy; Crossover strategy

1. Introducion

Under the background of the comprehensive promotion of the human Genome Project, life science research has entered the post-genome era, and proteomics research is one of the key contents. As early as the early 1990s, some scholars proposed the basic concept of proteome, which refers to the collection of all proteins produced by a gene or variable RNA splicing and post-transcriptional modification of mRNA expressed by a cell or tissue. In other words, a collection of proteins expressed under certain conditions, at a certain moment and in a certain type of cell or organism. In the process of research and exploration, protein organization includes basic function, post-translational modification, protein identification and many other contents. Proteomics theory and related technologies can be used to accelerate the development of molecular medicine, truly serve

ICCIC 2024: International Conference on Computer and Intelligent Control, June 29–30, 2024, Kuala Lumpur, Malaysia

[*] Corresponding author.

liuzhiqiang@xidakonjac.onaliyun.com (Z. Liu); tangyulan@xidakonjac.onaliyun.com (Y. Tang)

0009 0000 8990509X (Z. Liu); 0009 0009 8290 7573 (Y. Tang)

human life and health, and provide reference for the design of drugs that act on specific target molecules. Protein is a biological macromolecule that occupies a special position in the biological body. It is composed of a variety of amino acids according to dehydration condensation according to a specific method. After winding and folding of polypeptide chains, polymer compounds with a certain spatial structure can be formed, which is the basic condition of all metabolic activities. Because proteins in cells cannot exist independently and must interact with other proteins to achieve biological functions, protein interactions regulate most biological processes at the cellular level, such as DNA replication and transfer, signaling, and cell cycle regulation. In order to easily and quickly discover complexes from protein interaction networks, some scholars proposed to use graph representation, protein interaction network, protein belongs to the vertex, the interaction between proteins belongs to the edge, the use of a variety of graph clustering algorithms, intelligent optimization algorithms and complex network theory for mining analysis, not only can better understand the topological structure of protein networks, It can also clarify the biological significance of the complex, which will eventually help predict unknown protein functions and human disease genes in the new era [1].

In the 21st century, some scholars first proposed topological structures found in networks. Complex recognition has attracted extensive attention in the fields of data mining, bioinformatics and complex networks. They are connected together in a relatively close way to accomplish specific functions together. It has been found that in the protein network, the complex has strong biological unity, and the functional similarity of the proteins contained in the complex exceeds the functional similarity of other protein pairs in the network. Some scholars have also proved that the density inside the complex is relatively high, which is related to the biological unity, so it is of great significance to study the complex recognition algorithm in the biological network. Current scholars have proposed a cluster-based protein complex discovery method, which uses points or clusters as seed nodes to expand, regards density as distance targets, and identifies the cluster as a protein complex when the density value of the cluster is higher than a certain threshold. Some scholars have proposed an intelligent optimization algorithm represented by heuristics, which has gradually formed a competitive complex discovery method in practice. Compared with clustering algorithm, it is mainly used to solve the problem of social structure discovery in complex networks such as social networks. For example, in 2007, some scholars used genetic algorithm for community mining for the first time, and proposed a single-path crossover operation suitable for string coding. In order to solve the insufficient application of faction filtering algorithm (DPM), some scholars introduced the protein complex scale of distance constraint recognition, and proposed the protein complex recognition algorithm DPM-DR based on group penetration and distance constraint, which is more accurate and effective in practice and more comprehensive in content [2.3]. This is also the beginning of the application of multi-objective evolutionary programming ideas in protein interaction networks. According to the existing research results, the most common compound discovery algorithm of protein interaction network is still clustering algorithm, but there are still many challenges in practical application research. Therefore, this project mainly discusses the complex discovery method in protein interaction network from the perspective of food science and technology research in China, taking the protein interaction network of Saccharomyces cerevisiae as an example [4].

1.1 Research status of protein complex

The recognition of protein complex is a hot topic in bioinformatics. The accurate recognition of protein complex can help human understand the process of living organism. With the advancement of traditional biological experiments, researchers have obtained a large amount of PPI data, which provides a basis for the identification of unknown protein complexes. Based on this, many methods for protein complex identification based on PPI data have emerged. In this paper, some representative methods for predicting protein complexes are selected, and these domestic and foreign research methods are summarized in this section. At present, there are four traditional methods to identify protein complexes in biological experiments: mass spectrometry, protein microarray, X-ray crystallography and nuclear magnetic resonance. Although the traditional prediction method has certain advantages, such as high accuracy, can identify the three-dimensional structure of the protein complex, can observe the atomic level of detail, but the high cost of the experiment, long cycle and other disadvantages pose challenges to the implementation of the experiment. Protein microarray technology requires expensive reactants and reagents; The method of identifying protein complex based on X-ray crystallography requires high quality crystal as the experimental medium, and there may be some radiation damage during the experiment. NMR, on the other hand, may be disturbed by low-quality signals and may produce spectral overlap. To avoid these problems, the identification of protein complex [5] by PPIN became a "parallel" to traditional prediction methods. Compared with the above six methods, other protein complex recognition algorithms are less used. The typical ones are RSGNM method, which constructs PPIN through Poisson distribution and defines a series of parameters, and determines whether a protein belongs to protein complex by adjusting the parameters. The SGNMF method combines the Laplace operator of PPIN and the Kullback-Lcibler divergence to find protein complexes . In addition, influenced by heuristic algorithms, some swarm intelligence optimization algorithms are also used to identify protein complexes. F-MCL method combines Markov method and Firefly FA method to mine protein complex. ICSC is an improved cuckoo search clustering algorithm similar to F-MCL method, which is used to detect protein complex in weighted dynamic PPIN. In ICSC, a dynamic PPIN needs to be constructed first. Then the complex core structure of the subnetwork is found, and the accessory structure of the protein complex is obtained by clustering, so as to identify the protein complex. Unlike ICSC, MNC uses a multi-network model to mine protein complexes, the authors argue that traditional graph clustering algorithms only identify protein complexes by a single network, ignoring information from many other networks, while multi-network clustering models mine protein complexes by relationships between domain interactions. At present, the mainstream key protein recognition algorithms can be divided into the following five categories:

1. Methods based on biological experiments

The most direct methods for identifying key proteins in biology-based experiments are conditional gene knockout, single gene knockout and RNA interference. Among them, gene knockout is a molecular biological technique that emerged after the 1980s. Based on the basic principle of DNA homologous recombination, gene knockout is realized by replacing the target gene fragment with the homologous fragment of gene. The principle of RNA interference to identify key proteins is that RNA induces rapid and specific degradation of homologous mRNA during evolution. Although the traditional method of identifying key proteins has high accuracy, it requires a lot of human, material and biological resources. With the development of high-throughput technology and the increasing amount of biological data, how to efficiently mine key proteins in massive PPins has become a research hotspot [6].

2. Method based on topology structure

Studies have shown that whether a protein is a key protein is related to the topological characteristics of protein nodes in PPIN. The method of identifying key proteins based on topology first needs to calculate the topological centrality of the protein nodes in the network. The larger the topology centrality value of nodes in the network, the higher the probability of node becoming a key protein. The "center-lethal" rule proposed by Jeong et al. shows that in PPIN, the greater the degree value of a node, the more important it is. A few years later, researchers used degree centrality (DC) to identify key proteins, with promising results. Since then, there have been many methods to identify key proteins by using topological centrality. Among them, the intermediate centrality (BC) is obtained by using the number of shortest paths passed by the node when calculating the topological centrality of a node. The information centrality (IC) calculates the length of the harmonic mean path to obtain the topological centrality value of the protein nodes in the network. Proximity centrality (CC) is to obtain the degree of intimacy between protein nodes and calculate the topological centrality of nodes. Subgraph centrality (SC) The number of simple paths taken by a node is calculated to determine topological centrality. Wang et al. proposed domain centrality (NC) based on edge aggregation coefficient to identify key proteins in 2011; In addition, eigenvector centrality (EC) and local mean junction centrality (LAC) are also used to identify key proteins. It is worth mentioning that Cytoscape is a software for biological network analysis and node centrality evaluation, making it easier to identify key proteins based on topological centrality. Although the identification of key proteins based on topological centrality is very accurate, the identification of key proteins will be greatly affected due to the large number of false positive data in the protein interaction data used [7].

2.Method

2.1 Basic Concepts

The protein interaction network can be viewed as a simple undirected graph, where node binding represents a collection of proteins and edits represent interaction binding between proteins. In this paper, the overlap score evaluation is used to measure the similarity between subgraphs. If subgraphs have more common nodes, the corresponding overlap score will be higher and higher. As one of the more common intelligent algorithms, genetic algorithm is a concept proposed by referring to biological genetic mechanism to find an optimal solution. The actual process is shown in Figure 1 below: [8.9]

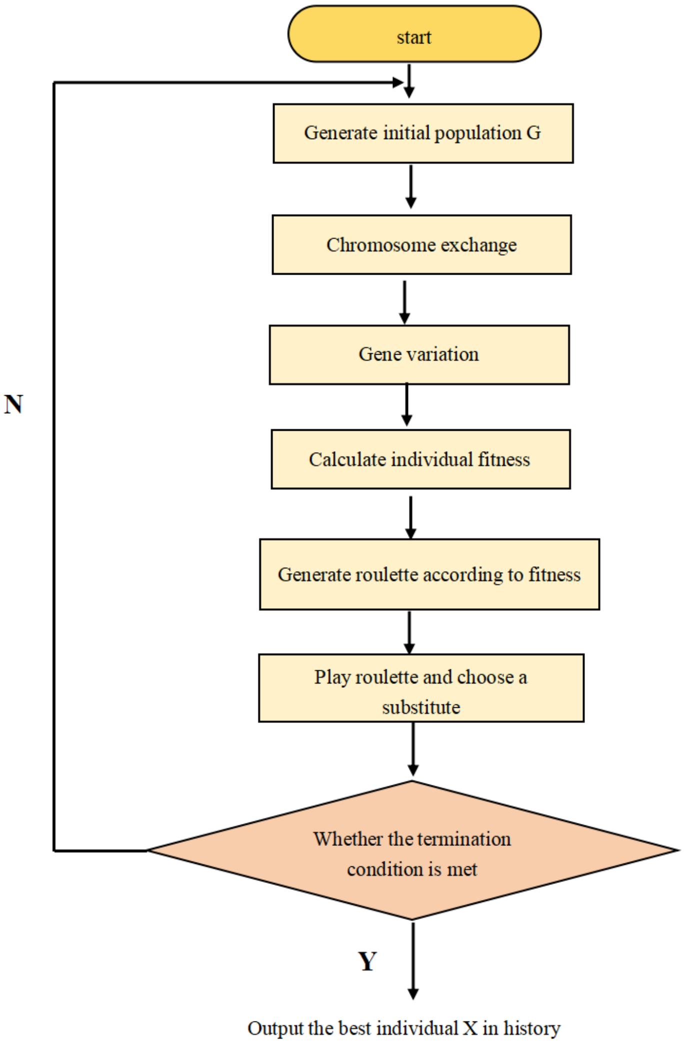
start
Generate initial population G
Chromosome exchange
Gene variation
Calculate individual fitness
Generate roulette according to fitness
Play roulette and choose a substitute
Whether the termination condition is met
N
Y
Output the best individual X in history

Figure 1: Flow chart of genetic algorithm.

2.2 Protein complex recognition algorithm based on genetic algorithm

The complex identification analysis is carried out in the protein interaction network, that is, the graph clustering of the corresponding network, and finally a cluster corresponding to a protein complex is obtained. In essence, graph clustering refers to a NP-hard combinatorial optimization method, which uses heuristic search method to obtain satisfactory results within a specified time, and the selection of seed nodes and cluster expansion process is the key. Genetic algorithm, as one of the more common algorithms in current application research, will use population search to select, cross, mutation and other basic operations on the current population, so as to generate a new generation of population, and use population evolution to reach or approach the optimal solution. In this paper, the graph clustering method GAGC with genetic algorithm as the core is used for complex discovery. The specific steps are as follows: [10.11]

First, objective function - individual fitness. As an important index of genetic algorithm, fitness is mainly used to evaluate the degree of superiority and inferiority of individuals in the population. The actual function value will affect the survival probability of each individual in the population, and will affect the performance and convergence speed of the algorithm. The algorithm studied in this paper takes F-measure as an individual fitness evaluation function, which includes two indicators: accuracy rate and recall rate. The former refers to the ratio between the compound number correctly identified by the algorithm and the compound number obtained by the algorithm, while the latter refers to the accuracy rate of identifying protein complexes in the standard database.

Second, the initial population generation process. In the research process, factor model algorithm (IPCA) is used to construct a set of feasible graph structure, which is regarded as the initial population. The node weights of the network are defined first, and the node with the greatest weight is selected as the seed node and expanded into clusters. Only the node with the greatest weight is selected as the seed node for cluster expansion, which cannot guarantee the diversity of the initial population. Therefore, the IPCA algorithm should be improved. In the process of selecting seed nodes, the roulette mechanism is added, and the greater the weight of nodes, the higher the probability of seed nodes. In this way, not only can the selection of seed nodes be guaranteed to have a greater weight, but also can reflect the diversity of choices, and finally obtain multiple feasible solutions as the initial population [12].

Third, choose a strategy. In order to demonstrate the survival of the fittest concept of genetic algorithm, high-quality individuals should be allowed to participate in evolution, and individuals in the population should be selected to produce the next generation of individuals. At this time, individuals with higher fitness should be selected for optimization. If the fitness value is selected directly according to the order from high to low, the diversity of the population will be reduced, and the genetic algorithm will be premature. Therefore, in the algorithm research, roulette is used to select the probability of becoming the next generation of individuals in direct proportion to the fitness of nodes. At the same time, in order to ensure that there are optimal individuals in each generation of population, if the fitness value of the optimal individual is lower than that of the previous generation of optimal individuals after a new population for next generation optimization is generated, So you replace the worst of the current population with the best of the previous generation.

Fourth, the crossover strategy. This step is the key content of genetic algorithm. If two chromosomes are directly selected to cross, there will be a large number of nodes without cluster belonging, which directly affects the clustering effect. In order to cross effectively, the genes in the two chromosomes are aligned according to the overlap score, so that the subgraph represented by the two alleles has more duplicate nodes. In this process, based on allelic cross operation, the connectivity of new subgraphs generated after gene cross can be ensured [13].

Fifth, the strategy of variation. In addition to inheriting the information of the parent's individual, the offspring produced by cross operation will also vary according to a certain probability, which fully reflects the diversity of biological heredity. In the process of mutation, a random mutation is first generated, corresponding to the gene on the chromosome [14].

3. Result analysis

3.1 Experimental design

In order to verify the application performance of the proposed algorithm, from the perspective of food science and technology in China, the yeast protein interaction network data was taken as the research object, and after obtaining the data set meeting the research requirements, the experimental analysis was carried out according to the process shown in Figure 2 below:

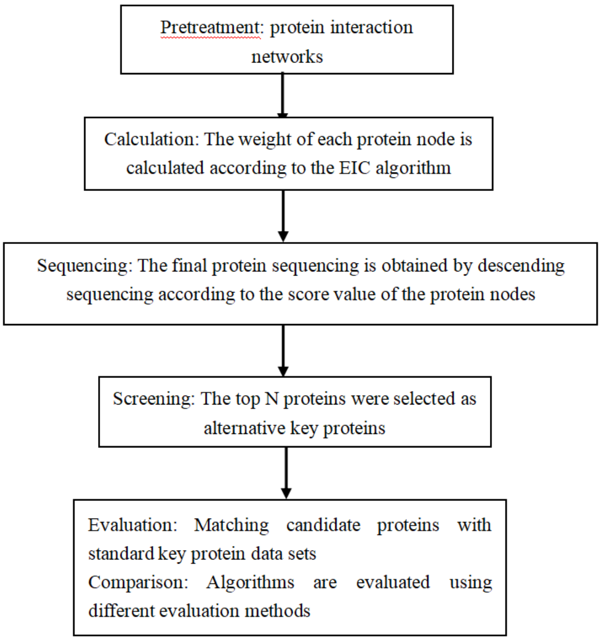
Pretreatment: protein interaction networks
Calculation: The weight of each protein node is calculated according to the EIC algorithm
Sequencing: The final protein sequencing is obtained by descending sequencing according to the score value of the protein nodes
Screening: The top N proteins were selected as alternative key proteins
Evaluation: Matching candidate proteins with standard key protein data sets
Comparison: Algorithms are evaluated using different evaluation methods

Figure 2: Experimental flow chart

Table 1

Data set and attribute analysis

Protein network	Protein Number of vertices	Network interaction Active edge number	Average degree	Key protein Qualitative quantity	Cluster coefficient
MIPS	4 546	12 319	5.42	1 016	0.0879
DIP	5 093	24 743	9.72	1 167	0.0973

In order to further explore the accuracy and effectiveness of the applied algorithm, the ranking results of corresponding protein nodes should be obtained after the node weights assigned by the descending order algorithm, and protein nodes of different sizes should be selected as alternative key proteins for comparison and analysis with standard key protein data. It is possible to specify the number of candidate key proteins that are truly recognized as key proteins.

3.2 Experimental Results

In the parameter setting and selection of the algorithm, experiments are conducted according to the above selected network set, the obtained node weights are processed according to the sorting-screening method, and the first 100, 200, 300, 400, 500 and 600 nodes in the sequence are selected as alternative key proteins. Thus, the number of true key proteins contained in the candidate key proteins can be determined, and finally the results shown in Table 2 below can be obtained:

Table 2

Parameter selection and comparison results

a	MIPS network						DIP network					
	100	200	300	400	500	600	100	200	300	400	500	600
0	48	82	123	163	213	255	46	82	115	158	201	252
0.1	52	97	130	178	222	267	69	103	144	186	236	278
0.2	61	105	147	186	230	275	77	132	181	217	260	301
0.3	66	118	163	199	239	278	83	145	193	241	283	324
0.4	65	126	172	209	249	288	82	145	199	253	300	349
0.5	67	127	180	226	265	298	81	149	203	257	305	349
0.6	67	127	186	234	279	312	79	147	207	259	302	347
0.7	68	131	187	238	281	320	75	147	208	260	304	352
0.8	**68**	**134**	**186**	**234**	**284**	**327**	**79**	**149**	**208**	**256**	**308**	**351**
0.9	67	133	184	229	279	331	75	149	209	262	305	349

After defining the algorithm flow, compared with the application performance of other algorithms, including DC, BC, EC, SC, LAC, NC, and UC, the final results show that the sensitivity, specificity, positive and negative prediction values of the proposed algorithm are significantly better than those of other algorithms, as shown in Table 3 below. In the future research and development of food science and technology, scholars should continue to combine other biological information, machine learning and deep learning algorithms to comprehensively improve the recognition efficiency and quality of key proteins, in order to provide a reference for the application of technical algorithms.

Table 3

Performance comparison results of different algorithms

Numeric set	method	SN	SP	PPV	NPV	F	ACC
MIPS	DC	0.254	0.803	0.291	0.772	0.271	0.671
	BC	0.197	0.796	0.278	0.716	0.231	0.629
	EC	0.139	0.773	0.163	0.738	0.150	0.620
	SC	0.138	0.773	0.162	0.739	0.149	0.620
	LAC	0.271	0.812	0.314	0.779	0.291	0.682
	NC	0.281	0.814	0.325	0.781	0.302	0.686
	UC	0.271	0.812	0.314	0.778	0.291	0.682
	EIC	0.426	0.853	0.480	0.824	0.452	0.750
DIP	DC	0.353	0.834	0.409	0.800	0.379	0.716
	BC	0.308	0.823	0.361	0.785	0.333	0.700
	EC	0.323	0.824	0.374	0.789	0.347	0.701
	SC	0.316	0.822	0.366	0.787	0.339	0.698
	LAC	0.405	0.852	0.472	0.815	0.436	0.743
	NC	0.400	0.850	0.463	0.813	0.428	0.739
	UC	0.391	0.850	0.458	0.811	0.422	0.737
	EIC	0.440	0.862	0.509	0.826	0.472	0.759

In this paper, protein networks are divided into 11 groups according to subcellular localization information. According to the above conclusions, we conclude that the distribution of key proteins in the organism is not uniform, and the number of key proteins is more in some regions. We expect to improve the recognition of complexes and, more importantly, key proteins by clustering protein networks by subcellular location.

In order to prove the effectiveness of Ess-LGS method, this section selects a variety of methods to compare ACC, Precision, Recall and F1 values, and compares the performance of different algorithms on the same data set. Table 4 shows the clustering effect of protein complexes in 11 subcellular locations in this paper.

Table 4

Experimental data for protein complexes at 11 subcellular locations

	Precision	Recall	F1 value	ACC
Endoplasmic	0.716	0.490	0.582	0.340
Cytoskeleton	0.590	0.770	0.670	0.330
Cytosol	0.532	0.648	0.584	0.219
Endosome	0.720	0.450	0.550	0.320
Extracellular	0.844	-	-	-
Golgi	0.868	0.564	0.676	0.342
Mitochondrion	0.701	0.644	0.740	0.317
Nucleus	0.636	0.590	0.640	0.374
Peroxisome	0.844	0.323	0.428	0.404
Plasma	-	0.733	0.784	0.294
Vacuole	0.794	0.500	0.608	0.316

4. Conclusion

To sum up, in the face of higher requirements for food safety and biomedicine in the new era, with the continuous improvement of relevant technical theories, the development of industry technology is facing new opportunities and challenges. Especially for the identification of key proteins, although scholars from various countries have proposed and applied a large number of methods to predict key proteins, how to improve the accuracy of recognition has been the main problem puzzling scholars. In this paper, a protein complex recognition method based on genetic algorithm is proposed, and its performance is obviously better than other algorithms, so it is worth popularizing and applying in clinical field.

Reference

[1] P. Lu, J. Wei, “Key protein recognition algorithm based on complex structure and network topology,” Journal of Lanzhou University of Technology, vol. 002, pp. 048, 2022.

[2] B. Zhong, J. Yang, Q. Yin, “Alloy process Optimization based on Neural Network and Genetic Algorithm,” Equipment Manufacturing Technology, vol. 7, pp. 272-273, 2023. (in Chinese)

[3] Y. Tang, D. Li, Q. He, “Expression recognition method based on genetic algorithm and integrated pruning,” Journal of Sichuan Light Chemical University: Natural Science Edition, vol. 36, no. 5, pp. 67-75, 2023. (in Chinese)

[4] J. Liang, Y. Zhang, D. Chun, "Frozen soil constitutive model based on GA and PSO algorithm parameter identification," Journal of ice frozen soil, vol. 46 no. 1, pp. 235-246, 2024. DOI: 10.7522 / j.i SSN. 1000-0240.2024.0020.

[5] R. Feng, Z. Chen, S. Yi, "Identification of maize varieties by SVM based on Bayesian optimization," Spectroscopy and Spectral Analysis, vol. 006, pp. 042, 2022.

[6] R. Li, Y. Chen, H. Wu, et al., "Analysis of damage and failure of reinforced concrete shed cave laid with sand-EPE composite cushion under rockfall impact," Engineering Mechanics, vol. 41, pp. 1-16, 2024.

[7] J. Zhang, C. Zhong, "Research progress of protein complex and functional module prediction algorithms based on protein interaction networks," Guangxi Science, vol. 002, pp. 029, 2022.

[8] Y. Ge, X. Zhao, "Two-step protein complex prediction method based on graph embedding and cuckoo search," Journal of Ocean University of China (Natural Science Edition), vol. 007, pp. 052, 2022.

[9] Z. He, H. Cao, F. Zhao, "Simulation of 10 kV Arrester deterioration state recognition based on Hybrid Genetic Algorithm," Microcomputer Applications, vol. 11, pp. 69-72, 2023.

[10] W. Chen, "Design of automatic facial expression image recognition system based on improved genetic algorithm," Automation Technology and Application, vol. 42, no. 5, pp. 35-39, 2023. (in Chinese)

[11] S. Lin, "Identification of sizing slice spun material based on Genetic algorithm and BP neural network," Light Industry Science and Technology, vol. 6, pp. 125-129, 2023. (in Chinese)

[12] Q. Zhang, "Handwritten digit recognition based on BP neural network optimization based on genetic algorithm," Information and Computer, vol. 35, no. 7, pp. 75-77, 2023. (in Chinese)

[13] H. Huang, T. Wu, W. Wang, et al., "Protein complex structure prediction: methods and advances," Chinese Journal of Synthetic Biology, vol. 4, no. 3, pp. 507-523, 2023. (in Chinese)

[14] L. Liang, W. Cao, L. Li, et al., "Research progress of Protein-polysaccharide non-covalent and covalent complex embedding active ingredients," Food Science, vol. 44, no. 21, pp. 368-385, 2023.