Applications of bioinformatics in the post-genome era

Yuming Tang *

School of foreign languages, China Pharmaceutical University, Nanjing, China Pharmaceutical University, China

* Corresponding Author Email: 2417705019@qq.com

Abstract. Bioinformatics uses various informatics technical tools to study biological data that are difficult to process by general methods, to predict and compare biological information, etc. In the post-genomic era, the application of bioinformatics is highlighted and the methods of information processing have been expanded through innovation based on the original ones. In the post-genomic era, bioinformatics has attracted tremendous interest from medical and industrial communities worldwide, and it studies various biological genome sequences, single nucleotide polymorphism analysis, comparative genomics and protein structures, playing a profound role in biology, medicine, agriculture and information technology.

Keywords: Bioinformatics, Applications, Prospects.

1. Introduction

When did we discover the biological secrets of our bodies? DNA sequencing technology reveals the mystery of the human genome to us and brought about a dramatic increase in the amount of data in the DNA database, in the era of "big data", bioinformatics plays a huge role in life science research through statistical calculation and analysis of the data, the acquisition and interpretation of genomic information [1]. For instance, advances in the single-cell transcriptome have allowed researchers to discover new cell types, study complex cell differentiation [2], trace developmental processes, and improve understanding of human disease, in single-cell RNA sequencing, data amplification, and data loss will interfere with data analysis of RNA sequencing [3], in that case, noise reduction technology is needed for sparse single-cell RNA sequencing. Moreover, Mathematics also plays an important role in bioinformatics, statistics is a fundamental basis of it [4], probability theory and stochastic process theory, operational research, and optimization theory are also practiced in applications [5], such as hidden Markov chain model. In the post-genome era, the work is to study the structure of protein products and find ways to cure diseases such as cancer and AIDS [6]. How will bioinformatics be applied with its expanded techniques and methods?

Bioinformatics covers database technology, molecular modeling technology, mathematical statistical method, pattern recognition technology, and many other methods to be applied in these aspects: gene sequencing, protein structure prediction [7], and single nucleotide polymorphism.

2. Gene Sequencing

TIn the sequencing of genomes, bioinformatics has found practical applications. It goes through various stages of data collection and storage, data processing, and analysis. We take metagenomics as an example, from the storage of big data to the pre-processing of information, and finally information analysis to form a theoretical basis. In the analysis of microbial community, which is an important application area of metagenomics, 16S rRNA gene amplification is the most commonly used sequencing technology, 16S rRNA is a component of the 30S subunit in the ribosome of prokaryotes, including a conserved region and a variable region [8]. The conserved region reflects the genetic relationship between species, while the variable region reflects the differences between species [9], the specificity of the variable region can reflect the characteristic nucleotide sequences of different microorganisms, to analyze the microbial information [10]. It is worth noting that there are nine variable regions of v1-v9 in the16s rRNA gene sequence, which have different lengths, none

of the variable regions can accurately classify all kinds of bacteria clearly, and the different region also has a distinctive influence on the analysis of microbial community structure. The experimental process of sequencing is divided into the extraction of sample DNA, PCR amplification of the specified region, and final sequencing [11].

For instance, in the comparison of microbial diversity in the milk of two different breeds of camels, researchers take a total DNA test, measuring the concentration, purity, and quality of the DNA, and use it as a template for PCR amplification of the bacteria's v3-v4 variable regions, finally, they use sequencing technology to measure microbial 16s sRNA sequence, compare and analyze community structure and diversity, and apply abundance grade curves to reflect the richness and uniformity of the species contained [12], where OTU is a hypothesized unit in phylogenetic analysis or population genetic studies, generally, 97 percent similarity is clustered into one OTU, and each OTU corresponds to one species, a representative sequence is selected from the OUT of each cluster for subsequent analysis. After cluster analysis, OTU will be annotated to obtain specific species classification.

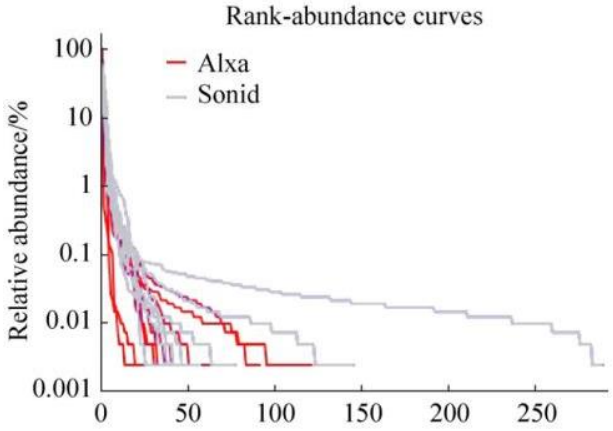

Figure 1. OTU level rank [12]

According to the figure, the curve of the second sample corresponding to Alxa dropped significantly and had a small span, which was reflected by the large abundance difference and low uniformity of OTU, indicating that the bacteria had a single composition and obvious dominant strains, while the curve characteristics of the ninth sample corresponding to Sonid were contrary to the former, indicating that the species were rich in composition and the dominant strains were not obvious.

In addition, Alpha diversity was applied in the context of sequencing results to microbial abundance and evenness in the samples. Shannon, Simpson, ACE, and Chao 1 index specifically reflect the richness and diversity of microorganisms.

In the first figure, the Simpson index, as a value greater than 0 and less than 1, represents the selection of two individuals from a community to determine the probability that they belong to the same species. It can be seen that the high value of group I indicates low richness and evenness and poor diversity. For convenience, sometimes Simpson's index of Diversity 1-D or Simpson's index of Diversity 1/D is used to solve the problem that the data is inconsistent with intuition, here, we refer to the original concept to analyze the value, instead of using these two versions of the formula.

In the second figure, Chao 1 index is used to estimate the number of OTUs in the community by the chao1 algorithm. The large chart indicates a large total number of species. Combined with the first figure, we deduce that group L is characterized by a large number of species, good richness and uniformity, and high diversity.

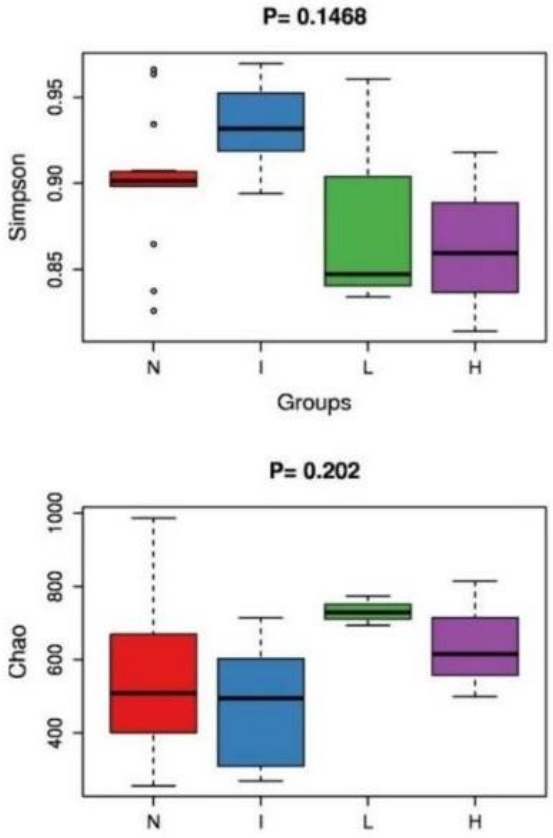

Figure 2. Sequencing results of microbial Alpha diversity [29]

\Beta diversity represents the difference in species composition between communities at different sites, which is commonly assessed by the distance between species communities. PCA analysis is one of the effective methods.

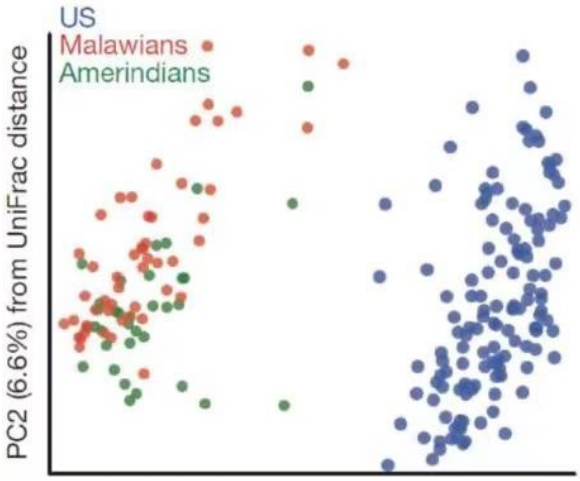

Figure 3. Differences in species composition between communities at different sites by beta diversity sequencing [29]

In the figure, three points of different colors correspond to a specific sample, the samples of the same color belong to the same group. For instance, blue belong to the US, it can be seen that the distance between samples of the same color is concentrated, indicating that there is a little discrepancy in community composition in the US group. However, it is not excluded that some individual samples with different colors have a small discrepancy with other groups, which is substantiated by many samples of Malawians and Amerindians that are close to each other.

Finally, the species with significant differences between groups are included in the results.

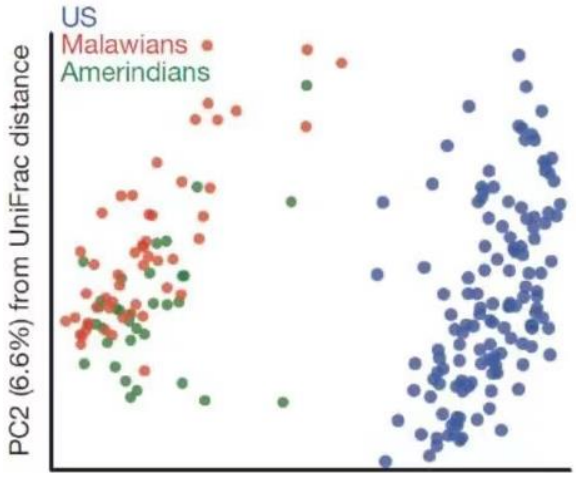

Figure 4. Species ring map [29]

Cladogram is an evolutionary branching diagram drawn according to the analysis structure, in which different circles represent various classification levels and each node represents a species. For example, a large node represents a high species abundance. Red means a significant difference, yellow means the opposite.

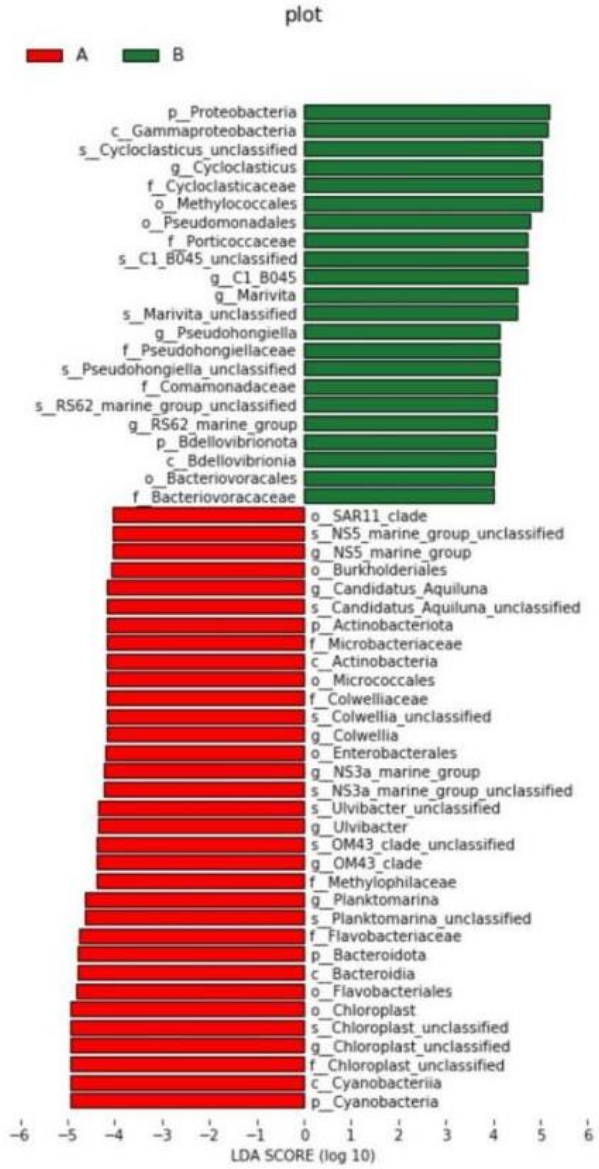

Figure 5. Evolutionary branching diagram [29]

Besides, the column chart demonstrates significantly different species. The color represents the group in which different species are enriched, and the column length describe the size of the LDA score, and the influence degree of species between different groups.

Through the understanding of the procedure steps of 16s rRNA gene sequencing technology and the analysis of examples, we can see the effective role of bioinformatics in gene sequencing, especially focusing on data reprocessing and analysis, concentrating species information on some hypothetical units, using images for analysis, and according to the corresponding data information of images and color information to determine the abundance, diversity, and discrepancy between biological population, which enables us to obtain more genetic information conveniently[13]. In addition, facing the current situation of a high incidence of infectious diseases, rapid identification of pathogens is of great significance for clinical diagnosis and treatment, control of disease transmission and cost reduction, the use of gene sequencing technology, such as metagenomic sequencing, breaks the restriction of isolation and pure culture of microorganisms, reduces the time from sampling to case report within 24 hours, and improves the probability of successful detection of pathogens. All these are practical and effective applications of bioinformatics gene sequencing in life.

3. Protein Structure Prediction

In the field of protein research, bioinformatics plays an important role. It focuses on the prediction of the spatial structure of proteins, which has a guiding significance for understanding proteins, enhancing protein engineering, and giving structure to drug molecule design. Considering the slow progress of X-ray crystallographic truncation and multi-flavor NMR spectroscopy in the study of the three-dimensional structure of proteins, bioinformatics mainly used homology modeling and Ab initio. When we know the first-order sequence of proteins, how do we get their tertiary structure? In contrast to accurate measurements, the analysis of the predictive structure can also obtain a lot of information, such as predicting the mutation sites of proteins. The structure of a protein is determined only by the amino acid sequence, knowing the primary sequence can theoretically deduce secondary and tertiary structure [14]. On account of the conservative nature of tertiary structure in evolution, assuming that half of the amino acid sequences of two proteins are identical, about ninety percent of alpha carbon atoms are within 3 angstroms of each other. Under these conditions, we can use homology modeling techniques to predict the secondary and tertiary [15]. Different from homology modeling, abinitio uses optimization algorithms to search for global lowest energy conformational solutions in conformational space directly based on protein physical or knowledge energy models according to the Anfinsen rule, which declares the natural structure of small water-soluble proteins under physiological conditions only depends on their amnio acid sequence. Although the homology modeling method has high accuracy when it can find highly reliable templates, it is difficult to achieve good results when homologous protein cannot be found. For ab initio prediction method, it is divided into two categories, the first method simulates the folding process of protein and optimizes the predetermined energy function to the lowest state from random conformation, but it takes a lot of time to simulate protein folding, and it is difficult to achieve high accuracy when proteins are long in length. Secondly, the algorithm presented by rosetta cannot measure the mass of all conformations because it only uses a single energy function for optimization, in that case, it is not entirely accurate. In order to improve the predictive performance of protein structure, we need to pay attention to several aspects, 1. Reduce the redundancy of information in the protein sequence, 2. Change the prediction method to improve the prediction accuracy, 3. Design efficient algorithms, 4. Deal with the folding phenomenon of protein structure [16].

Now we refer to MELD-DNA as an example, a new computational method for predicting the structure of protein-DNA complexes. First, the prediction of protein structure and the interaction between protein and nucleic acid helps to decipher gene expression, genome repair, and storage mechanisms [17]. In biology, the binding of transcription factor protein and DNA has received special attention on account of its biological function [18]. To understand the interaction between

transcription factor protein and DNA, it is necessary to master what structure a specific protein-DNA use, in addition, understanding the structure of the complex assists to predict the relative binding affinity of different sequences, but there was a lack of experimental structural data and calculation methods for this complex [19].

MELD-DNA method uses Bayesian inference to combine molecular dynamics simulations with general knowledge and experimental information. It also applies the necessary constraints to prevent DNA and proteins from unrolling at high temperatures. Different from the usual protein structure prediction methods in the past, it poses general knowledge in terms of ambiguous data to drive protein-DNA structure prediction and assumes protein structure, combined DNA sequence, and DNA binding domain. First, using chimeras, individuals with different genetic traits chimeric or mixed performance, to generate a smooth rotating trapezoidal spiral structure of B-DNA and dummy particles at the N1 position of each purine base, this is different from protein approaching through the major groove, which is the spiral groove on the surface of twining B-DNA double heli.

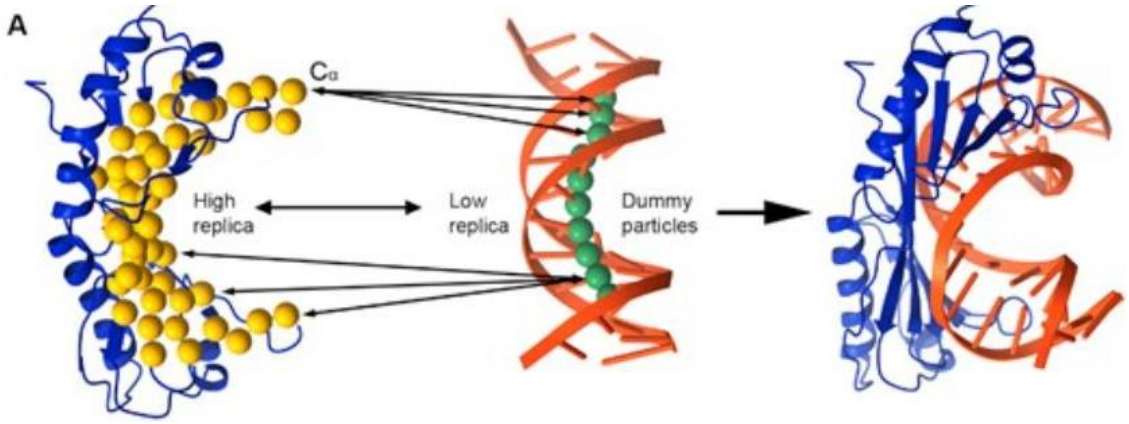

Figure 6. The spiral groove on the surface of twining B-DNA double heli [20]

The distance between protein(blue) and DNA (orange) is at least 30 Å at the beginning to ensure that the system has never been in a binding state, interaction between Ca atoms(yellow) and dummy particles(green) generates simulation information.

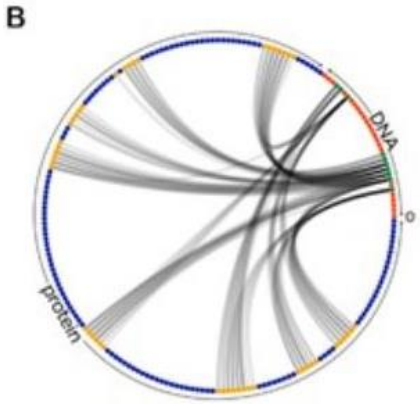

Figure 7. Interactions between proteins, DNA, Ca atoms, and pseudo particles [20]

A list of possible contacts is generated, however, due to the existence of noisy data, only some of them can be satisfied. Use a circular chart to show the data, each residual is a point on the circle, the black line refers to information used to guide binding. The meld simulation only meets a small part of all possibilities at any given time in the simulation.

Second, there is a peculiarity in the treatment of the lack of conformation, the experiment does not interfere much, this is different from many improved experimental methods, for instance, a method of using multiple energy functions to expand the search space for the lack of conformational evaluation caused by Rosetta's use of a single energy function. The experiment relies on the force

field to sample the most likely conformation according to the existing data. See the schematic of the MELD Bayesian inference approach.

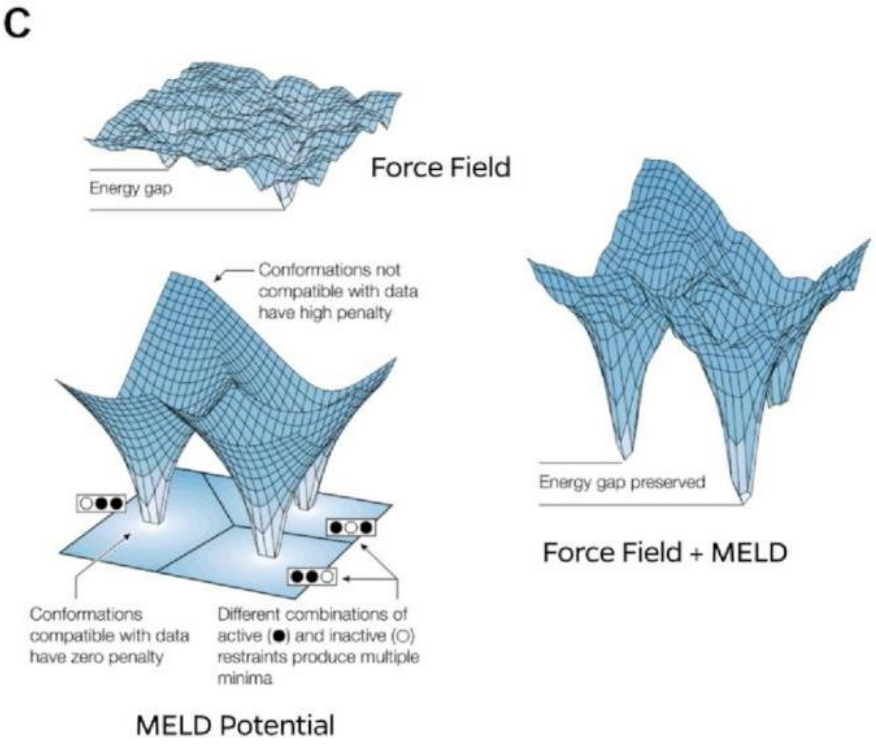

Figure 8. The schematic of the MELD Bayesian inference approach

Chose some protein-DNA system to apply this approach.

Table 1. Protein-DNA systems simulated in this study, along with their DNA sequence and PDB IDs

System	DNA sequence	PDB	Reference
Nuclear intron-encoded homing endonuclease 1-Ppoi	TGACTCTCTTAAGAGAGTCA	1A74	(51)
Hyperthermophile chromosomal protein Sac7d	GCGATCGC	1AZP	(52)
9-Cisretinoic acid receptor	TAGGTCAAAGGTCAG	1BY4	(53)
Human papillomavirus type-18 E2	CAACCGAATTCGGTTG	1JJ4	(54)
Phage 434 Cro	AGTACAAACTTTCTTGTAT	3CRO	(55)
Fungal transcription factor Put3	CGGGAAGCCAACTCCG	1ZME	(56)
Murine Creb Bzip-Cre complex	CTTGGCTGACGTCAGCCAAG	1DH3	(57)
P22 C2 repressor	CATTTAAGATATCTTAAATA	2R1J	(58)
Human Tbp core domain	CTGCTATAAAAGGCTG	1CDW	(59)
Gcn4 leucine zipper	TTCCTATGACTCATCCAGTT	1YSA	(60)
Gcn4 leucine zipper	TGGAGATGACGTCATCTCC	2DGC	(61)
Glucocorticoid eeceptor	TCAGAACATGATGTTCTCA	1R4R	(62)
Glucocorticoid eeceptor	CCAGAACATCGATGTTCTG	1R40	(62)
EcoRV restriction endonuclease	CGGGATATCCC	1BGB	(63)
EcoRV restriction endonuclease	AAAGAATTCTT	2BOD	(64)

Now look at the results.

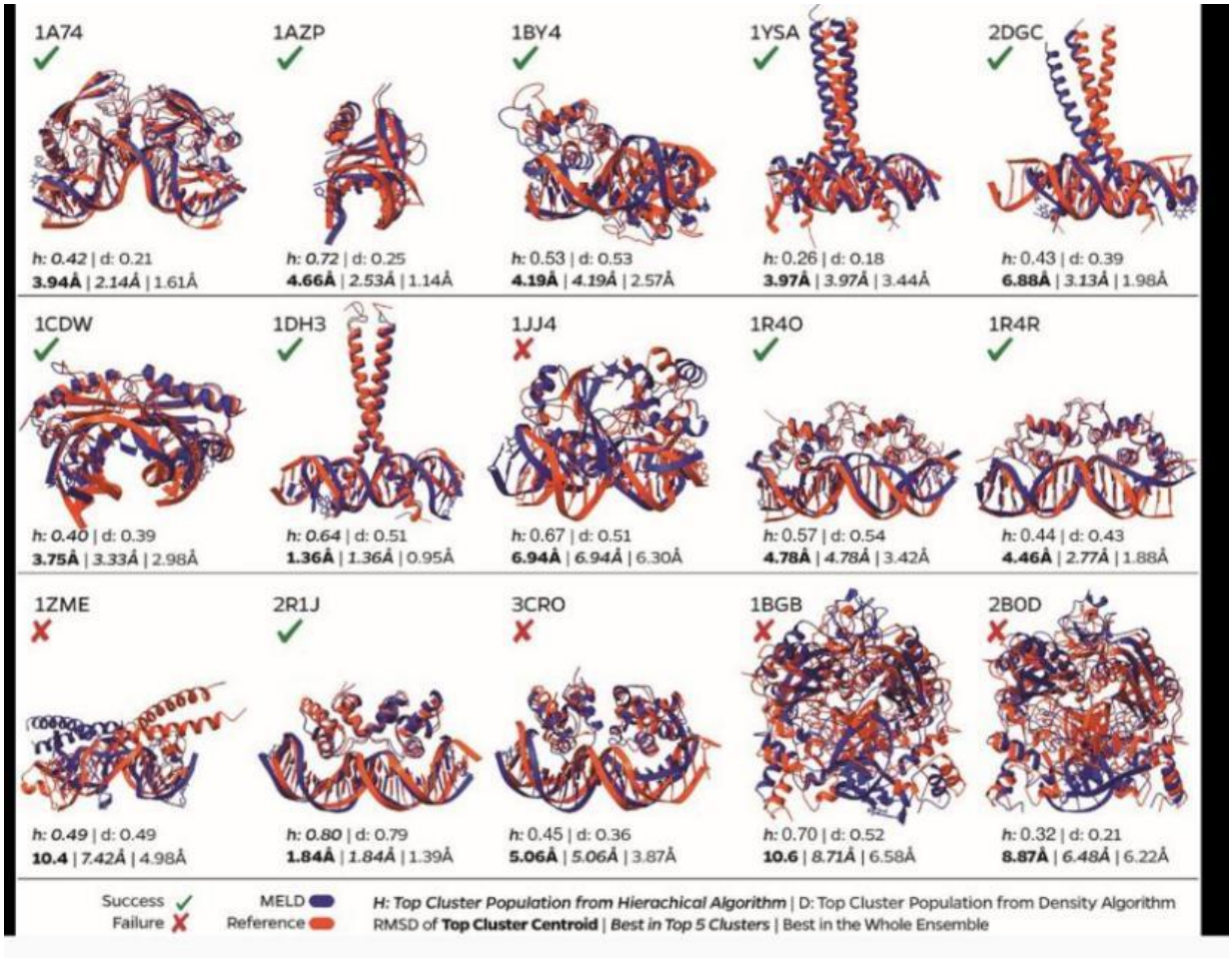

Figure 9. Results of some protein-DNA systems using the MELD Bayesian inference method [20]

The picture shows the superposition of the optimal cluster and the experimental structure in the first five clusters simulated by MELD, if the conformation less than 5 A is founded in the first five clusters, the prediction is successful.

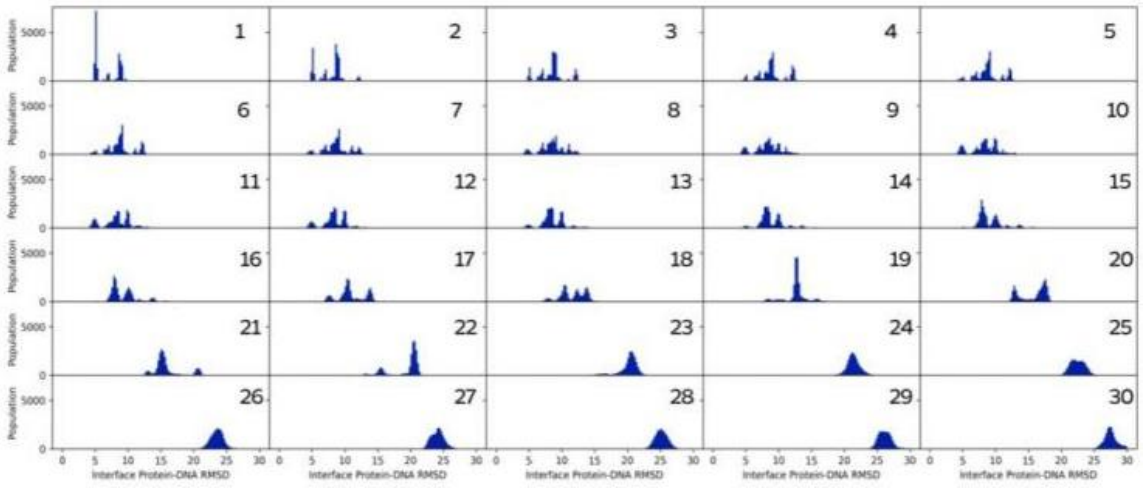

Figure 10. The superposition of the optimal cluster and the experimental structure in the first five clusters simulated by MELD [20]

This figure shows the structural distribution of one of 15 complexes in the meld method sample at the replica.

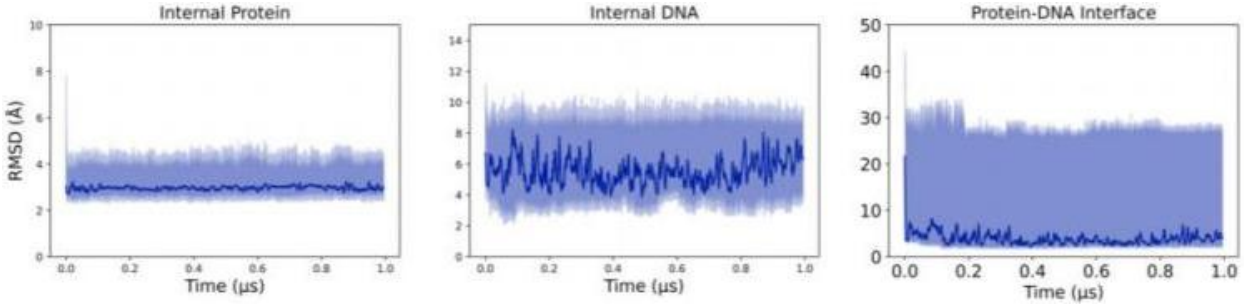

Figure 11. The structural distribution of one of 15 complexes in the meld method sample at the replica [20]

The figure demonstrates that the DNA and protein ensembles approach the holo conformation as the native binding mode is sampled. Even though the protein and DNA have been restricted before [20].

In conclusion, the above results prove that this method can sample the native state and help us to identify the native state with high confidence without knowing the actual structure, the MELD method has been used to predict protein structure, small molecules, and protein complexes in the past, while the MELD-DNA method not only expands the framework, but also proves that it can be used for general protein-DNA structure, and it is widely used in this field. This method samples a variety of binding patterns, is sensitive to the configuration of DNA sequence, and is different from the general method in the processing of conformation. It can be seen as the expansion of the protein structure prediction method, and introduces new sight in the supplement of the DNA structure database and the interaction between protein and DNA [21].

4. Single Nucleotide Polymorphism

Finally, bioinformatics is applied to the analysis of single nucleotide polymorphism, which is the nucleic acid sequence polymorphism caused by the change of a single nucleotide. According to its position in the gene, single nucleotide polymorphism is divided into Coding-region SNPs, cSNPs, Perigenic SNPs, pSNPs, and intergenic SNPS, iSNPs. Snp is widely present in the human genome [22]. Generally. There are few cases of snp with three or four allele polymorphisms, in that case, snp is considered a secondary allele, which is only composed of one or two bases. As a third-generation DNA genetic marker, snp is applied in many fields such as molecular, forensic material evidence testing, disease diagnosis, and treatment [23]. Single Nucleotide Polymorphisms, an important basis for genetic variation, are of great significance to population genetics, pharmacogenomics, and disease gene research. Let's analyze the specific detection methods and their main applications in the post-genome era [24].

The first is the sequencing method, Sanger sequencing is a classic method of DNA sequence analysis, which can directly obtain nucleic acid sequence information, and is an important standard of SNP detection. Sanger sequencing can find unknown SNP sites and determine the mutation type and mutation location of SNP, demonstrating the characteristics of direct and effective. Sanger sequencing is based on the method of termination of dideoxy ribonucleotides (ddNTP), which starts at a fixed point and terminates randomly at a specific base, then, a series of nucleic acid fragments are separated by electrophoresis to obtain the visible base sequence, thus obtaining the nucleotide. The diagram below shows how Sanger sequencing works [25].

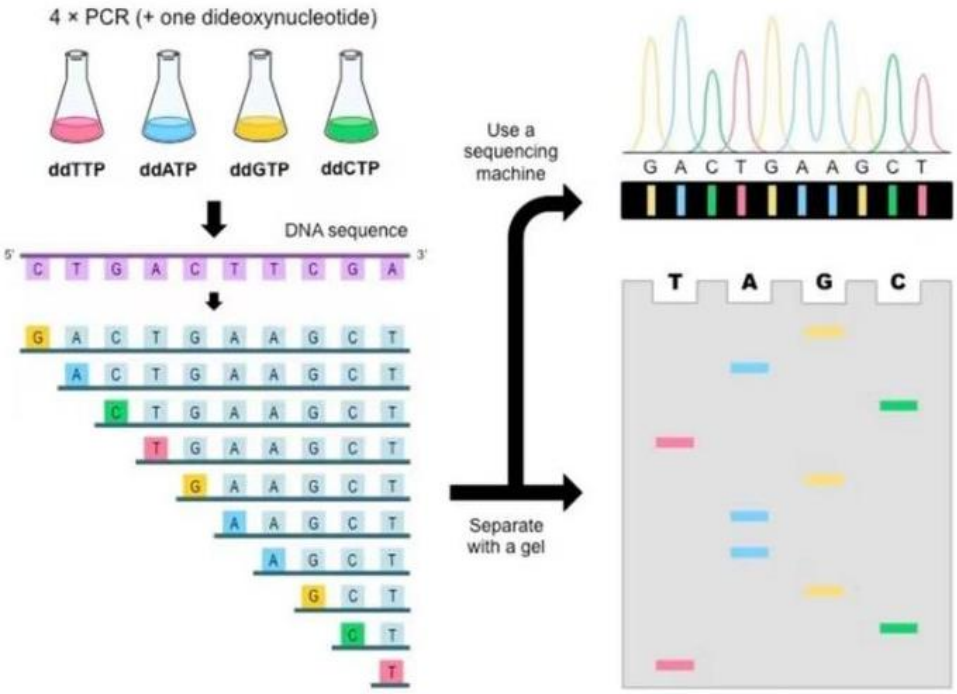

Figure 12. Sanger sequencing principle of operation [20]

Now, take the application of Sanger sequencing in the detection of drug-resistant mutations in patients with hepatitis B as an example to explain the specific operation of Sanger sequencing. HBV is a blood-borne hepadnaviral, which causes chronic hepatitis, liver cancer, and other diseases. First, extract DNA, in this experiment, researchers used an HBV-DNA extraction kit, followed the protocol provided by the kit, amplified and purified the HBV-DNA, then carried out PCR sequencing and purified the sequencing products, one percent AGAR electrophoresis was used to observe the sequencing status, the product was added into the 96-well plate and sequenced. Finally, data collection and analysis were carried out to obtain the mutation of the hepatitis B drug resistance gene and explore the occurrence of nucleotide resistance [26].

Currently, except the Sanger sequencing method, which directly obtains nucleic acid information, other methods are designed based on known SNP sites, such as the TaqMan probe method, snapshot method, etc. At present, with the rapid development of gene chips, high-throughput sequencing, and omics big data technology, SNP detection cost has been greatly reduced. Gene chip refers to a large number of oligonucleotides targeted at SNP that are fixed on a surface in high density, thus forming a multiple oligonucleotide dimensional array. The target sequence is fixed according to the hybridization characteristics of nucleic acid fragments, and SNP sites are determined by scanning.

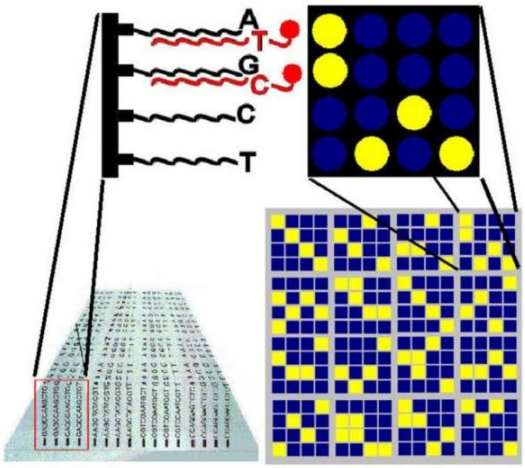

Figure 13. Gene Chip sites SNP [31]

This figure shows the detection principle of gene chip, which has high detection and can be automated, many enterprises, such as llumina and Thermo Fisher, are engaged in the production of gene chips. We take the application of gene chips in the construction of island cotton fingerprints as an example, the researcher developed high-density SNP(Cottonsnp80k) on their own and selected six SNP sites to develop specific SNP products, ten sea island cotton materials were selected for verification analysis based on electrophoretic detection and consistency of SNP polymorphism and chip typing result.

Table 2. SNP loci information and their confirmation results

SNP marker	Chr	Position (bp)	Unmatched No.	Primer (5'-3')
TM6482	A03	12431407	0	F: GGCCCCCAATTACACCCCACAAG R: ATCGATCCCAGGTCTTCCCCTATGC
TM10688	A05	14284527	0	F: GTTAGGTAAAACGTTCGTTGAGATGACGAATG R: GAACCATTTGCAATTTAAAAATTTATAGTGATAAGA
TM15125	A06	65978280	0	F: ATCTCATTTCTATCTATTTATTGCATTGACATGGA R: CCTGTAGGTTTCAAAGGGTTGCAAGTTAGG
TM13666	A06	1879153	0	F: GCAGTCAAAAAATAACAGTTAAAGGCGTAGGGATA R: GGGCTTTGGATCTACTTAATTTCACTCCACTG
TM20004	A07	26468617	0	F: CAACGTCTTAATTTGTTATGTTCTTAGAGTATTGCA R: ATGGCACGTTGACAACTTGACTCTTTCAC
TM20245	A07	39991651	0	F: CATCAAACGACGACAAGTGGGTAACAAGA R: AGAAGTATGTTTACCACACAACCGAGGTTATCAAC
TM53439	D03	1828343	0	F: CCAACGACGATCAGGGTAATCTTGTCTCATA R: TCAATAGCAGGAGAGTTGGTTCCAGTGATTATT
TM55633	D04	4698154	1	F: AAGTAGTTCGTAGTATGAGAATAGCAACGAAGATAG R: TCAAAAACTTTCATTCAAGTCACTAAATCATTCAAA

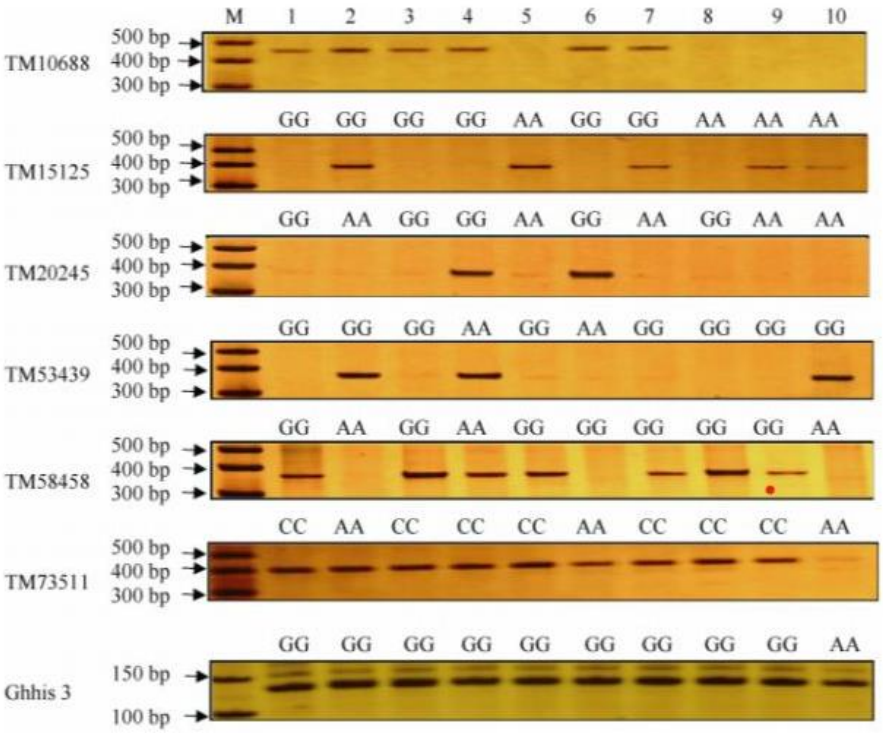

Figure 14. Verification of SNPs from chip genotyping by SNP -PCR analysis [27]

The results showed that the results of chip analysis were highly consistent with SNP-PCR, which proved the availability of SNP sites and the accuracy of gene chips based on chip SNP typing [27].

SNP detection relies on progressive methods, which is of practical significance for the diagnosis of genetic diseases and corresponding drugs [28]. Currently, coronavirus, as an RNA virus, is constantly mutating in transmission, and its toxicity and infectivity are enhanced, which is closely related to SNP sites, there will be greater demand for the development of SNP detection in the future.

5. Conclusion

In conclusion, bioinformatics uses a variety of informatics technology means to study biological data that are difficult to be processed by general methods, to predict and compare biological information, etc. In the post-genome era, the applications of biological information have been highlighted, and the methods of processing information have also been expanded on the original basis through innovation. For example, in gene sequencing, metagenomics can rapidly predict pathogens, which is the key to clinical treatment. In protein structure prediction, MELD-DNA method expands the original protein structure prediction method, while SNP detection is also of great significance for understanding and treating coronavirus. In the post-genome era, bioinformatics arouses huge interest in the medical and industrial sectors around the world, it studies various biological genome sequences, single nucleotide polymorphism analysis, comparative genomics, and protein structure, plays a profound role in the field of biology, medicine, agriculture, and information technology.

References

[1] Gao Yuan. Bioinformatics development in the post-genomic era [J]. China Science and Technology Information, 2009, (10): 225 - 226.

[2] Van Zundert G.C.P., Rodrigues J.P.G.L.M., Trellet M., Schmitz C., Kastritis P. L., Karaca E., Melquiond A.S.J., Dijk M., de Vries S.J., Bonvin A.M.J.J The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J. Mol. Biol. 2016; 428: 720 - 725.

[3] Chen Ming. Bioinformatics in the post-genomic era [J]. Bioinformatics, 2004, (02): 29 - 34.

[4] Buniello A., MacArthur J.A.L., Cerezo M., Harris L.W., Hayhurst J., Malangone C., McMahon A., Morales J., Mountjoy E., Sollis E. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019; 47: D1005. D1012.

[5] Gapsys V., de Groot B.L. Alchemical free energy calculations for nucleotide mutations in protein–DNA complexes. J. Chem. Theory Comput. 2017; 13:6275 - 6289.

[6] Rashkin S.R., Graff R.E., Kachuri L., Thai K.K., Alexeeff S.E., Blatchins M.A., Cavazos T.B., Corley D.A., Emami N.C., Hoffman J.D. et al. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat. Commun. 2020; 11: 4423.

[7] Grant B.J., Rodrigues A.P.C., ElSawy K.M., McCammon J.A., Caves L.S.D. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006; 22: 2695 - 2696.

[8] Chen Wenjuan, Ganchi, Zhao Ruike, Mo Qian, Cao Qing. Clinical value of PCR / 16 sRNA combined with nucleotide sequencing method [J]. Advances in Modern biomedicine, 2018,18 (12): 2289 - 2293.

[9] Ning Weize, Xu Jun. ITS and 16sRNA gene libraries for the microbial composition spectrum of P. Jinhua tea [J]. Chinese Tea, 2018, 40 (05): 63 - 66.

[10] Wang Guoyang, Wang Jing, Zhao Liping, Zhang Xiaojun, Zhang Menghui. 16 Comparative analysis of the two variable regions of the S rRNA gene reflecting the gut flora diversity and the ability to identify species [J]. Chinese Journal of Microecology, 2013, 25 (09): 1005 - 1009.

[11] Liu Shuangshuang, Tiyun, Qi Lin, Liu Fenghui, Wang Lei. 16 similarities of the evolutionary relationship between the variable region of S rRNA gene and full-length sequence [J]. Journal of Zhengzhou University (Science edition), 2022, 54 (01): 19 - 24.

[12] Zona, Eli, Hosna, Jimitu. Comparing the microbial diversity of naturally fermented acid camel milk between Sunet and Alxa camels based on 16S rRNA gene sequence analysis [J]. Journal of Microbiology, 2019, 59 (10): 1948 - 1959.

[13] Auton A., Abecasis G.R., Altshuler D.M., Durbin R.M., Abecasis G.R., Bentley D.R., Chakravarti A., Clark A.G., Donnelly P., Eichler E.E. et al. A global reference for human genetic variation. Nature. 2015; 526:68.74.

[14] Wang Fang, Li Hongjin, Li Huyang. Exploring the methods of protein structure prediction [J]. Modern Information Technology, 2022, 6 (18): 122 - 125.

[15] Liu Jilong, Xiao Zhixiong, Cao Yang. Analysis of the suitability of homology modeling methods for predicting protein mutational structures [J]. Journal of Sichuan University (Natural Science Edition), 2017, 54 (03): 658 - 664.

[16] Wang Yulin, Zhang Biao, Shen Hongbin. Multi-objective optimized protein 3 D structure prediction [J]. Journal of Jiangsu University of Science and Technology (Natural Science edition), 2021, 35 (04): 66 - 74.

[17] Moult J., Fidelis K., Kryshtafovych A., Schwede T., Tramontano A. Critical assessment of methods of protein structure prediction: progress and new directions in round XI. Proteins Struct. Funct. Bioinform. 2016; 84: 4. 14.

[18] Dai H., Umarov R., Kuwahara H., Li Y., Song L., Gao X. Sequence2Vec: a novel embedding approach for modeling transcription factor binding affinity landscape. Bioinformatics. 2017; 33:3575.3583.

[19] Honorato R.V., Koukos P.I., Jiménez-García B., Tsaregorodtsev A., Verlato M., Giachetti A., Rosato A., Bonvin A. M. J. J. Structural biology in the clouds: the WeNMR-EOSC ecosystem. Front. Mol. Biosci. 2021; 8: 729513.

[20] Esmaeeli Reza, Bauzá Antonio, Perez Alberto. Structural predictions of protein-DNA binding: MELD - DNA. [J]. Nucleic acids research, 2023,

[21] Zhou Z., Dong S. Protein. DNA interactions: a novel approach to improve the fluorescence stability of DNA/Ag nanoclusters. Nanoscale. 2014; 7: 1296.1300.

[22] Liu Jiqiang, Hao Xiaodong, Wu Lina, Liao Shiying, Feng Yifang, Mi Shirong, Liu Shen, Liu Jian, Zhang Longchao. Application of whole genome SNP typing in livestock genetic breeding [J]. Journal of Animal Husbandry and Veterinary Medicine, 2022, 53 (12): 4123 - 4137.

[23] Campolongo M. J., Tan S.J., Xu J., Luo D. DNA nanomedicine: engineering DNA as a polymer for therapeutic and diagnostic applications. Adv. Drug Deliv. Rev. 2010; 62: 606.616.

[24] Wei Jie, Zhang Xinyan, Wang Hong, Zhao Lan, Liu Wei, Li Huan, Fu Rui, Qiao Han, Zhao Meng, Xiang Xinhua, Yue Bingfei. Evaluation of ability validation results for detection of single nucleotide polymorphism markers in experimental mouse nucleic acid samples [J]. Laboratory animals and Comparative Medicine, 2022, 42 (06): 505 - 510.

[25] Zhou Qingsu, Meng Fei, Zhu Liming, Jiang Xiaofen, Zhu Fangchao, Pan Jie. Application of PCR-Sanger sequencing method in the individualized treatment of H. pylori infection [J]. Zhejiang Medical, 2021, 43 (01): 37 - 41 + 119.

[26] Ma Liang, Yu Yang, Zhang Jingying, Cong Xiao, Liu Qian, Yang Hui, Cao Yongtong. Application of Sanger sequencing method in detection of resistant mutations in patients with hepatitis B [J]. International Journal of Laboratory Medicine, 2019, 40 (01): 1-3 + 7.

[27] Li Lechen, Zhu Guozhong, Su Xiujuan, Guo Wangzhen. Screening and evaluation of SNP core sites suitable for sea island cotton fingerprint construction [J]. Journal of Crop Sciences, 2019, 45 (05): 647 - 655.

[28] Zheng J., Erzurumluoglu A.M., Elsworth B.L., Kemp J.P., Howe L., Haycock P.C., Hemani G., Tansey K., Laurin C. Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics. 2017; 33: 272 - 279.

[29] https://zhuanlan.zhihu.com/p/534068667.

[30] http://www.accbio.com.cn/service-22.html.

[31] https://mp.weixin.qq.com/s?__biz=MzA4NjIwNzEyMQ==&mid=2651905192&idx=2&sn=50c435654a09608d646cc5100de6cd6f&chksm=84284356b35fca4089479c5424e59d3d67e08e9327f60cfb3d64c28a89aaab0bc162a2bb590e&scene=27.