Application of Computer Information Processing Technology in Virus Sequence Analysis ——Taking the COVID-19 as an Example

Xuwenqing Fan*

Harbin NO.3 High School

ABSTRACT: Since the outbreak of the COVID-19 in Wuhan in 2020, the virus has frequently mutated and spread extremely fast, leading to large-scale infections worldwide. At present, countries around the world are still in the stage of virus research. In terms of new crown prevention and treatment, vaccines that can effectively enhance immunity to COVID-19 have been developed. Taking the COVID-19 as an example, this paper compares the COVID-19 with other types of viruses, and discusses the application of computer technology in virus data analysis, sequence analysis and mutation detection. The purpose is to facilitate researchers to count the number of confirmed cases, the spread of the virus, and to detect the mutation trend of the virus in a timely manner, so as to make effective prevention and control, and to minimize the spread of the virus and the number of confirmed cases. Finally, it is concluded that R language, image processing technology, probability model and other computer technologies can help the virus sequence analysis to be more accurate and efficient.

1 INTRODUCTION

There are a large number of different types of viruses in nature, and with changes in temperature, environment, and geographic location, the sequences of viruses are becoming more and more complex, making virus sequence analysis more and more difficult. Viruses are divided into DNA viruses and RNA viruses. The genetic material of DNA viruses is composed of deoxyribonucleic acid, usually double-stranded. The replication of most DNA virus genomes occurs in the nucleus; the genetic material of RNA viruses, including ribonucleic acid, is mostly single-stranded structure, and the replication process of most RNA viruses occurs in the cytoplasm. The replication process of RNA virus is more complicated than that of DNA virus, so it is more difficult to treat. For example, because HIV is an RNA virus, it is more difficult to treat. More than two decades ago, Stephen Morse made the assumption that if each of the 50,000 vertebrate species carried about 200,000 different kinds of viruses, then it could be inferred that the world There are about 1 million vertebrate viruses. In the face of such a large number of viruses, the sequencing methods are limited. The existing virus sequence analysis methods include the SARS-CoV-2 whole genome sequencing platform, called CorvGenSurv (Coronavirus Genome Surveillance), and for chicken infectious bronchitis. S1 gene sequence analysis of virus variant strains and PCR technology for cloning and sequence analysis of the whole genome of porcine circovirus type 2 isolates,

This technology lays the foundation for future research on IBV genetically engineered vaccines and the establishment of molecular biology methods such as PCR, PCR-RFLP, etc. Virus sequencing is very important. Taking COVID-19 as an example, virus sequence analysis can help in the design of diagnostic kits. Inaccurate sequencing results may lead to low-specific primer design, affecting the sensitivity and specificity of probes. Both of them may reduce the accuracy of the test kit, greatly increase the false positives of the diagnosis, and cause a large number of misdiagnoses. If only the symptoms are known and there is no diagnostic kit, it may bring a large number of suspected cases and no further diagnosis can be made. In addition, virus sequence analysis can also assist in the study of virus homology traceability and mutation tracking. Help people effectively prevent and treat disease.

2 COMPUTER TECHNIQUES COMMONLY USED IN VIRUS SEQUENCE ANALYSIS

2.1 R language

There are many data analysis languages. R language is a powerful and widely used data analysis language. The virus sequence data has the characteristics of large quantity and complex arrangement, and the analysis function of R language can just operate the data. The R language is a language and operating environment for statistical analysis, graphing. Its advantages are reflected in the following points: First, R language has top statistical

*f1697248928@163.com

analysis capabilities, and almost any data analysis can be performed in R language. Second, the drawing function of the R language is powerful, especially in the visualization of complex data, and even the graphics that you cannot describe in language can be realized by the R language. Bioconductor, a genetic information analysis tool written in R language, has efficient basic analysis functions. For example, in the screening of new targets for cervical cancer, the R language-based survival rate correlation analysis technology was used to analyze the targeted therapy, drug development and pathogenic mechanism research for cervical cancer. This research can save time and effort and accurately screen potential cervical cancer patients and give targeted treatment.[1] In addition, the R language also plays a role in the analysis and identification of key genes in osteosarcoma. The R language limma package is used to perform differential gene expression analysis on the selected osteosarcoma data set, and the core gene regions of differentially expressed genes and Key genes, which were initially verified to play an important role in the progression of osteosarcoma.[2] At present, the application of this technology also plays a role in the analysis of the new coronavirus. It is necessary to know that the genetic analysis of the virus is very complicated. All the positive samples detected should be sequenced on the whole genome, and then the sequence will be analyzed to compare with the sequence of the early Wuhan cases. In contrast, the type of mutation was analyzed, such as the "delta" type, which was analyzed by gene sequencing. The virus sequence is then compared with the sequence in the database to check for the presence or absence of mutation sites, and to make homology comparison. The source and route of infection should also be compared with cases in other regions. This process takes 15-20 hours, and the R language makes complex viral gene analysis simple and time-saving.

2.2 Image Processing Technology

Image processing technology is a method and technology for removing noise, enhancing, restoring, segmenting, and extracting features of images through a computer. This technology has the following advantages: First, it is true and reliable, and the characteristics of things are more intuitively displayed through images. Second, it is fast and convenient to convert complex and huge data into images. Third, accuracy. Through image processing technology, information that cannot be extracted by the human eye can be observed. For example, by enlarging the chromosomes of cells, you can intuitively see a mutated gene. . Currently, image processing techniques are widely used, including for analyzing gene sequencing data. For example, the known gene variation data is integrated into the "visualization concept", and the image comparison technology is used to find the difference of genes, which greatly reduces the complexity of gene variation data analysis, reduces the analysis time, and makes the final analysis. Data is more intuitive. Nowadays, the detection of the COVID-19 also integrates image processing technologies such as CT scanning. CT is used to observe the lung characteristics of patients, and confirmed or suspected cases will be isolated. This method is more time-saving and accurate than nucleic acid detection. In addition, a large number of visual news reports have emerged in the current epidemic news reports. Using the images and tables of the introduction, the presentation forms are intuitive and rich, so that the audience can clearly understand the epidemic situation.

2.3 Probabilistic Models

A probabilistic model is a mathematical model used to describe the relationship between different random variables, usually describing the mutual non-deterministic probability relationship between one or more random variables. Probabilistic models have many good properties. They provide a simple method of visualizing probabilistic models, which is beneficial for designing and developing new models. Probabilistic model is used to represent complex reasoning and learning operations, and it can simplify mathematical expressions. Today, probabilistic models already play a role in biology. For example, in neonatal early-onset Streptococcus agalactiae, several important pathogenic mechanisms of this disease were screened, and then a probability prediction model of Logistic regression was established, and finally a method for treating this disease was found through polymerase chain reaction. Through the establishment of probabilistic models, the pathogenesis and prevention methods can be found in time, and timely clinical drug treatment can be carried out. At the same time, the number of newborns with this disease has been greatly reduced.[1] In addition, probabilistic models have long been used in predicting the mutation of lung cancer factors. Lung cancer is a common malignant tumor in the world, with extremely high morbidity and mortality, and the mutation rate is as high as 50%. Therefore, scientists combined PET/CT (Based on the metabolic parameters and clinicopathological characteristics of patients with increased glucose utilization in malignant tissues, a mathematical model was established to predict the mutation status of tumor factor receptors in patients with lung adenocarcinoma. The model can timely and effectively predict the gene mutation of tumors, so that medical scientists can give timely treatment plans in clinical practice.[2] It can be seen that the probabilistic model plays a role in predicting gene mutations. If this model is applied to the study of the mutation of the COVID-19, the mutation of the virus can be detected in a timely and effective manner, and relevant prevention plans can be given, which greatly reduces the number of infected people.

2.4 Machine Learning

Machine learning is the use of algorithms to analyze data, capture large amounts of experimental data from it, and make decisions. Machine learning has the following outstanding advantages: First, data generation, which can help people obtain very complex data. Second, prediction, machine learning is good at exploring the correlation between variables and making accurate predictions. With the increasing number and complexity of research

experiments in the biological field, and the increase in biological experimental data, machine learning methods are also playing an increasingly important role in biological sequence analysis. For example, in analyzing the sequence characteristics of the hepatitis B virus genome and analyzing the risk of liver cancer, machine learning plays a crucial role. After completing the genome construction of the hepatitis B virus using the corresponding computational methods, the machine learning model is used to analyze the data set. Prediction of liver cancer risk was then performed. After the application of machine learning, the disease risk can be more effectively predicted, which is helpful for early warning and treatment of liver cancer patients.[1] Therefore, if this method is applied to the genetic sequence analysis of the COVID-19 to detect the mutation of the virus, the prediction accuracy will be greatly improved.

3 APPLICATION

3.1 Analysis of COVID-19 data

Traditional virus data statistics and analysis are very complex, requiring researchers to manually record virus data and aggregate them uniformly. This method is complicated and time-consuming, and is prone to data omission, which affects the statistical results. With the continuous development of computer technology, more and more computer technologies can be perfectly applied to the statistical analysis of virus data. Taking the data analysis of the new coronavirus as an example, first, the similarity matching method is used to calculate the similarity of virus genes in different regions. degree of infection, find out the source of infection and the mode of transmission, and then establish a regression model to analyze the relationship between viruses in different regions. Combined with the K-Means spatial clustering analysis method, the spatial distribution characteristics of the epidemic points are classified according to the spatial affinity, so as to realize the spatial division of the epidemic area and determine the boundary of each epidemic point. Data analysis of suspected cases, confirmed cases, and the distance, location, and relationship of contacts of cases within the monitoring range will help to reveal the law of epidemic spread. Finally, the R language will be used to visualize the data with the information at hand, and map the epidemic situation of the whole province and the country, as well as the movement trajectories of the confirmed cases. Visually display the current virus data in the form of graphs and tables.

3.2 Sequence analysis of the virus

In the previous viral sequence analysis methods of viruses, it is necessary to perform segmental amplification and sequencing by RT-PRC to obtain the gene sequence, and then compare the deduced amino acid sequence of the gene determined by PRE with other viral gene sequences to construct a gene nucleus. Genetic evolutionary tree of nucleotide sequences. This method is very labor-intensive, highly technically demanding for researchers, and the sequencing results are not 100% accurate. Therefore, we can perform virus sequence analysis with the help of computer-based deep learning methods. Using this model, a large number of known virus sequences are trained by deep machine learning based on the convolutional neural network model, and the optimal virus sequence binary classifier is obtained. In addition to this, the model is trained and evaluated using a large, carefully selected data set containing hundreds of thousands of viral and prokaryotic sequences. Dividing the data set by time avoids overlap between training, validation, and testing data sets, and also helps to evaluate the ability of the method to predict future new viruses based on previously discovered virus sequences. In order to effectively correct the sampling bias and improve the accuracy of virus prediction, a large number of macro virus samples were also collected. Moreover, the model does not need to define the characteristics of the sequence in advance, and the model can autonomously learn various feature attributes required for virus prediction.

3.3 Virus mutation detection

In traditional virus mutation detection, medical personnel conduct manual detection through the characteristics, morphology, staining and biochemical characteristics of the virus. Estimated the mutated virus, but exhausted a lot of manpower and material resources. Taking the most common COVID-19 detection as an example, the most commonly used is the fluorescent PCR method. Finally, the change of the product amount is monitored by the change of the fluorescence intensity, thereby obtaining a fluorescence amplification curve. However, the test results are prone to errors due to improper storage or failure to submit for inspection in time. With the development of artificial intelligence, computer technology can be effectively applied to virus detection and mutation prediction. For example, a virus database is established from the perspective of virus gene and structure. On the basis of the frequency estimation of mutation sites, the difficulty of nucleotide mutation, the difficulty of amino acid substitution, the impact of mutation on protein secondary structure, and the occurrence of single amino acid mutation are analyzed. ACE2 and neutralizing antibody binding free energy change and other parameters to evaluate the variation in multiple dimensions, and comprehensively analyze the impact of known variation and potential virtual variation on the function of the virus. On this basis, the artificial intelligence classifier algorithm was used to effectively group the variant strains in terms of transmissibility and affinity for neutralizing antibodies, thereby realizing risk assessment and early warning based on virus sequences. Taking the COVID-19 as an example, scientists can arrange the existing COVID-19 gene types in chronological order based on the existing virus types, and analyze the similarities and differences of viruses in different periods. As well as the number, brand and time of vaccination for different patients, the probability distribution of infection, recovery or death, a decision tree model is established to finally predict the mutation

direction of the COVID-19.

4 CONCLUSIONS

This paper introduces the application technology of R language in virus gene sequence expression, the application technology of image processing technology in virus sequence image, the application technology of probability model for virus mutation prediction and other application technologies in virus sequence analysis, and introduces the application technology of virus sequence analysis. These techniques can be applied in the data statistics and analysis of the COVID-19, sequence analysis and virus mutation detection. Although all countries have made great progress in combating the new crown virus since the outbreak of the new crown virus, the new crown virus is highly contagious, mutates rapidly, and the direction of virus mutation is difficult to accurately predict. Therefore, in the future, we will continue to explore the application of computer image processing technology in the detection of COVID-19, and we can also explore the application of natural language processing technology in virus sequence classification and analysis, as well as the application of computer technology in virus spread and mutation. application.

REFERENCES

1. Nanjing Drum Tower Hospital. New target screening method and system for cervical cancer based on survival correlation analysis technique of R language: CN202011131091. 4[P]. 2021-01-22.
2. Zhiming Shan. Bioinformatics analysis and identification of key genes in osteosarcoma [D]. Zhengzhou university, 2021. DOI: 10.27466 /, dc nki. Gzzdu. 2021.003470..
3. Yingwei Wang, Yaoqiang Du, Xiaolin Miao, et al. Risk factors and drug resistance analysis of early-onset Streptococcus agalactiae infection in neonates (English) [J]. Journal of Zhejiang University-Science B(Biomedicine & Biotechnology), 2018, 12 (12) : 973-980.
4. Qianzhun Huang. Correlation between PET/CT metabolism and histopathological subtypes and EGFR gene mutation in patients with lung adenocarcinoma[D]. Guangdong Shantou University, 2019.
5. Chunfang Gao. A method for high-throughput analysis of RT/S region sequence features of hepatitis B virus genome with a machine learning model to predict the risk of liver cancer: CN202010411458.1[P]. 2020-09-01.