Scholar Publishing Group
Academic Journal of Environmental Biology
https://doi.org/10.38007/AJEB.2021.020204
ISSN 2789-8407 Vol. 2, Issue 2: 30-38

Bioinformatics Data Analysis of Space Environment under the Background of Artificial Intelligence

Additya Kumar*

Binghamton University State University of New York, The United States of America

**corresponding author*

Keywords: Artificial Intelligence, Space Environment, Bioinformatics Data, Data Analysis Methods

Abstract: Bioinformatics (BI) addresses academic problems by using information technology to collect, store, organize, and index biological data, and to present, analyze, and integrate problems through biological data. It applies the fundamentals of computer science and information technology to make large, diverse and complex scientific data understandable and help realize its potential. The main purpose of this paper is to study the analysis methods of space environment BI data based on the background of artificial intelligence (AI). This paper combines big data (BD), discusses the ecological value of technology, puts forward the ecological value of BD technology, applies BD technology to the ecological field, points out the problems and reasons existing in the application of BD technology in the ecological environment, and finds solutions in the process of practice. In order to give full play to the ecological value of BD technology. Experiments show that the improved algorithm DGAN-VAE has good performance and can work well on images and gene sequences in the field of biological information.

1. Introduction

With the advent of the era of BD and the continuous development of Internet technology and information technology, compared with the past, BD is constantly changing people's thinking and looking at the world with a new and holistic way of thinking. Since the reform and opening up, my country's economy has maintained a medium-to-high speed growth and its scale has continued to expand. While gradually meeting people's material needs, it also pays a huge cost. The world's environmental problems are obvious, and it is not only humans who solve problems through emerging technologies [1-2].

In a related study, Jamshidi et al. proposed a response to the virus through AI (AI) [3]. Several deep learning (DL) methods have been shown to achieve this, including generative adversarial

networks (GAN), extreme learning machines (ELM), and long/short-term memory (LSTM). It describes an integrated BI approach in which different aspects of continuous information from structured and unstructured data sources are brought together to form a user-friendly platform for physicians and researchers. Bagabir et al.'s work using AI to identify genome sequences, develop COVID-19 drugs and vaccines, and recognize the advantages and challenges of using such technologies [4]. A non-systematic review of methods was also performed. can help to rapidly identify variants of interest (VOCs) as delta strains and Omicron. Also, many drugs are applied with the help of AI.

Variational Autoencoders (VAEs) can handle discrete data, while their loss functions can constrain the results produced by the model. Therefore, this algorithm is introduced to improve the generative adversarial network, and a generative adversarial network fusion model based on variational autoencoding is proposed: DGAN-VAE. A new neural network structure is designed for image and sequence data, and the results are compared with other improved models on the public dataset MNIST. The key technical issues and feasibility of Generative Adversarial Networks in the field of BI are explored.

2. Design Research

2.1. Research Advantages

The research of BI is a very broad topic, and it is a very valuable research direction from the characteristics of gene sequence, the microstructure of the principle to the health status of the individual. All these studies have one thing in common: they all need to analyze massive gene sequencing data or pathological data [5-6]. The hidden potential value of these massive data sets cannot be compared with small data sets. Compared with previous small data analysis, it has greater advantages:

(1) Reduce errors. Large sample analysis can significantly improve the accuracy of analysis results and reduce sampling errors caused by small samples.

Comprehensive information. For monomers such as individuals or gene fragments, comprehensive and objective information records can reduce the misleading of subjective evaluations and improve the accuracy of monomer analysis.

(2) Reflect the essence of things. For a certain feature or state of biological information, only by analyzing a large amount of biological data can we know whether the feature or state is a special phenomenon or a general phenomenon.

(3) Get the data rules. According to the central limit theorem, the larger the sample size, the more objective laws in the data can be analyzed.

While obtaining these advantages, there are also many problems. For example, in the face of massive biological data, there are major challenges from hardware devices to analysis techniques [7-8].

2.2. Biological Database

At present, BI has penetrated into various branches of computer science, generating a large amount of biological data. Therefore, the current database is not only a storage medium, but also integrates various functions such as data analysis, data demand and data storage [9-10].

The development of database can be divided into two stages: primary database and secondary database. The original database was developed to digitize the information in the published literature,

i.e. to create a bibliographic index in the database. With the rapid development of molecular biology, experiments have generated a large amount of data, which is difficult to record through paper media; that is, databases that record protein and nucleic acid sequences through information technology are born, which are called primary databases. After this, a large amount of data does not need to be expressed on paper media, greatly increasing the speed of development of BI. It should be noted that the primary database does not represent any biological knowledge, but consists of a large amount of redundant experimental data mixed with various degrees of error. The secondary database is based on the primary database to further synthesize biological knowledge and information, so as to facilitate the further analysis and exchange of data by researchers; that is, a biological database for theoretical analysis of data for a specific target in the primary database [11-12].

2.3. Problems Existing in BD Technology in the Ecological Field

With the continuous development of BD technology, various government departments actively meet new challenges and use BD to change the way of work. While bringing some convenience, there are still many shortcomings [13-14], mainly including:

(1) The utilization rate of BD technology in the ecological field is low

Judging from the current situation of data collection, the scope of the basic collection facilities now involves all-round three-dimensional monitoring of the sky and the earth, the BD collection technology is only simple storage, and the hardware and software are still in the research and development stage, which has not reached the level of intelligence. The economic development in the north and the south is unbalanced, and there are differences in the application of BD. Due to economic constraints, the pace of infrastructure construction and innovation lags behind, and it is impossible to carry out data interconnection between provinces and cities. The utilization rate of BD in the national ecological field grows slowly. Management is difficult.

In addition, in terms of environmental pollution control, more than 5,900 monitoring points for air quality, acid rain, sand and dust storms, surface water, groundwater, and soil have been built, and they are stored in real-time categories every day. The lag in data storage will weaken the advantages of BD technology, requiring intelligent technology to solve storage problems. At the same time, the update, upgrade and replacement of infrastructure equipment requires a lot of human and financial support. UAV monitoring is only used in emergency situations, and does not make it a normal monitoring work. The low utilization of monitoring facilities makes it difficult to ensure the comprehensiveness of my country's ecosystem monitoring data, which brings difficulties to the BD technology to process and analyze data, leads to one-sided data analysis, and brings difficulties to monitoring management and evaluation.

BD technology has been tried in water conservancy, atmosphere, vegetation, and desert monitoring, and is generally in its infancy. The collected data resources are huge but of low quality, which cannot effectively analyze and process the data. At the same time, the quality of data determines that the decision-making ability is weak. The combination of ecological fields such as water conservancy and BD is not deep enough, and a benign linkage mechanism has not been formed, resulting in extremely low utilization rate of infrastructure, so that the value of BD technology in the ecological field cannot be exerted [15] -16].

(2) The technological transformation rate of BD technology in the ecological field is low

Compared with national technology, there is a clear gap between scientific and technological innovation in my country's ecological field, and it is still at the level of follow-up. Environmental

technology is mainly promoted in China, with less export abroad, and it is in its infancy. For example, focusing on the pollution control work in the Liaohe River Basin, the special Liaohe River Water Project has evaluated more than 5,000 pollution control technologies during the "Twelfth Five-Year Plan" period, and passed more than 200 practical technologies, but only 45 have been industrialized and applied. Regarding the construction of a water environment safety monitoring network in the Liaohe River Basin, the network monitors the total amount of water pollution discharge, water ecology, and water environment in real time, and implements mobile phones to check water quality in a timely manner, and conduct effective supervision; on the establishment of a provincial-level BD center, to build an intelligent platform for integrated water environment management, to break through the multi-technology acquisition and transmission technology, and play a good demonstration effect to other regions. Combining the above research with the collection function of BD technology, it is necessary to further study and apply the integration and analysis functions of BD technology, break through the exploration stage of applied research in my country's ecological field, continuously improve the transformation of scientific and technological achievements, and promote the construction of BD technology in the entire ecosystem [17-18].

2.4. Generative Adversarial Networks

The core idea of GAN is to train the network model through a zero-sum game, so that the generator gradually learns the probability distribution of real data. The training process uses the minimax algorithm:

$$L_{GAN} = \min_{G} \max_{D} \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[1-\log D(G(z))] \tag{1}$$

The objective optimization function is the core part of the production confrontation network. Through optimization, after deriving GAN in

$$D_G^*(x) = \frac{p_{x\sim data}(x)}{p_{x\sim data}(x) + p_{z\sim G(z)}(x)} \tag{2}$$

To find the optimal solution of the discriminator D, in the whole GAN framework, the main purpose is to obtain PG(x;θ)≈Pdata(x). Therefore, it can be obtained that D*G(x)=1/2 the optimal solution.

The way of generating adversarial network training is: first fix the generator G, optimize the discriminator D through the real data xr=x and the fake data xf=G(z), so that D can distinguish xr and xf as much as possible; then do not optimize D, Optimizing for G is optimal when PG(x;θ)≈ Pdata(x).

3. Experimental Study

3.1. BD Technology Processing Flow

From the perspective of the four characteristics of BD, there are many ways to collect data and there are various types of data, so the processing methods will also be different. (1) Data collection is the first and basic step of BD technology processing. The first characteristic of BD is "quantity", that is, the data sources are many, the quantity is large, and the types are complex, so the methods and means to obtain large amounts of data are also crucial. The current way of collecting data can be collected by sensors, radio frequency identification, search engine technology and so on. In the

process of collection, not only relevant data, but also some seemingly unrelated data must be collected, which is to provide powerful and accurate data information support for subsequent data analysis.

(2) Data processing and integration are relatively difficult links. When the first stage is over, "cleaning" and "denoising" processing are required, and the processed data is integrated and stored. If this step is omitted and data analysis is performed directly, it will increase the difficulty of data operation. In this regard, the collected data is first filtered, and the useless information is "cleaned" and "denoised" by clustering and association analysis methods to prevent interference with the final result. ② Then integrate and correct the filtered data, and store it in a special database by classification and grading, which improves the efficiency of data query and simplifies the difficulty of data analysis.

(3) Data analysis is the core of data processing. After the completion of the previous stage, the original data is analyzed and summarized, and the BD technology further excavates the internal correlation of the data according to customer needs, reflecting the potential value of the data. The era of BD cannot be limited to simple data analysis methods in the past. Therefore, Google took the lead in proposing the concept of "cloud computing". With the Internet as the center, everyone can use and obtain a large amount of resources and data, providing effective solutions for BD processing and analysis.

(4) Data interpretation is very important. The most valuable thing for users is the interpretation of analysis results. With the increase in the amount of data, simple words cannot clearly explain the data results to users, causing confusion to users. Therefore, the emergence of data visualization technology solves the problem. It can intuitively present the results to users in simple and easy-to-understand forms such as images and tables, which is convenient for users to understand.

3.2. Problem Analysis and Process

In the field of BI, discovering the underlying laws in data is a basic requirement for studying life sciences. In this field, finding unknown information or states is the most important thing in deciphering the laws of life. For example: looking for loop structures in gene sequences, searching for unknown protein secondary structures in protein synthesis, finding molecular structures that meet certain characteristics for drug design, etc. In addition, in the field of biomedicine, exploring the possible conditions of lesions according to the core characteristics of lesions, and generating pictures that match the characteristics of the disease based on existing medical images, thereby reducing misdiagnosis and overtreatment, are not machine learning of the type of discriminant mode. way can be solved.

This type of problem requires that the model algorithm needs to have a certain creative ability, and can produce some "exceptional" results that people have not yet specifically recognized. There are very few researchers and research materials currently exploring this question. Therefore, in order to explore whether the DGAN-VAE fusion model proposed in this paper can be useful in the field of biological information, this paper starts from the two problems of generating medical images and generating gene sequences, and uses generative adversarial network technology to explore the essential laws of biological information data. Way. Thereby making a little contribution to more BI problems.

Here is a technical framework diagram, as shown in Figure 1. Compared with the general flowchart, the biggest difference lies in the construction of the network model and the definition of the loss function. This paper sets two different problems: image processing and sequence processing.

As far as the current technology is concerned, there are relatively excellent deep learning technologies in each of the two sub-fields. GAN or DGAN-VAE is only equivalent to a container, which can place different neural networks and combine them into suitable models to solve different problems. As mentioned in the previous chapters, there are only minima and minima objective optimization functions in the generative adversarial network. For different networks, the required loss functions can be added based on actual needs. For example, the loss function of DGAN-VAE adds some additional loss constraints in addition to the basic objective function of GAN.

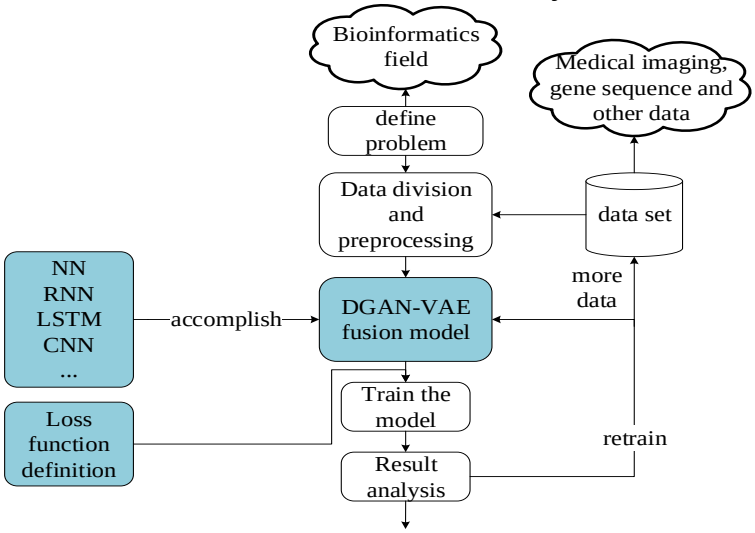

Figure 1. Technical framework for BI

4. Experiment Analysis

4.1. NLL Value Analysis

When evaluating the model, use the hyperparameter settings that were used to train MNIST. The model is first trained on DGAN-VAE using the TRAIN subset in the MNIST dataset. After the model is trained, DGAN-VAE is tested with 10,000 pieces of data in MNIST. During the test, the hidden variable z is not sampled from a normal distribution, but only the process of encoding and decoding the Encoder and the Decoder. Then calculate the NLL value of each image through NLL, and finally take the average as the final comparison result. The model comparison results are shown in Table 1, in which the results in black fonts are the results of the model in this paper.

Table 1. NLL result test on handwritten digit collection

Method	NLLTest
VAE(8-dim)	98.6
VAE(32-dim)	87.5
Nornalizingflows	85.10
VAE+HVI	88.3
VAB(8-dim)	84.0
DiscreteVAE	81.01
VAE+IAF	80.08
DGAN-VAE	**83.37**

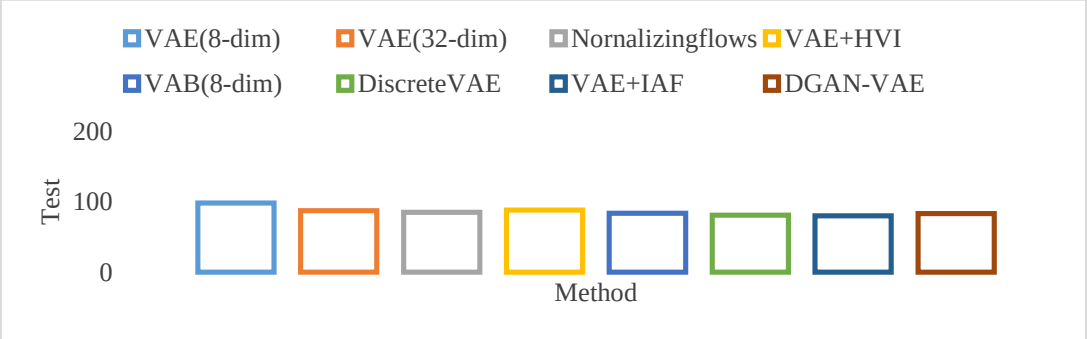

Figure 2. Analysis of NLL results test results on a collection of handwritten digits

In Figure 2, the NLL value of the model in this paper is lower than that of the first five models, because in the VAE and the improved VAE model, the model DGAN-VAE in this paper forms a mutual constraint on the loss function through the fusion of the GAN model and the VAE model. But the NLL value of DGAN-VAE is higher than the latter two models. After analysis, it is found that the constraint of image generation in the VAE model is the L2 method. There are strict pixel constraints between the two. However, in DGAN-VAE, the L2 constraint is only part of the influence of the entire model, and does not limit the model capability of GAN. According to the test results, it can be concluded that the DGAN-VAE model not only has the learning ability of GAN, but also the entire model is constrained by the VAE model.

4.2. Results Comparison

Use the SVM classifier for classification, with the help of the SVM software package LIBSVM, where the main parameters in the SVM are obtained by the grid search method, and the kernel function selects RBF, which is the same as the KNN cross-validation method. The overall accuracy of the feature extraction method using LIBSVM classification is shown in Table 2.

Table 2. Comparison of results of different methods on two datasets (%)

	Dataset A	Dataset B
Grayscale histogram	91.35	95.78
Grayscale Co-occurrence Matrix	92.78	97.60
Dual-tree complex wavelet	98.50	99.29

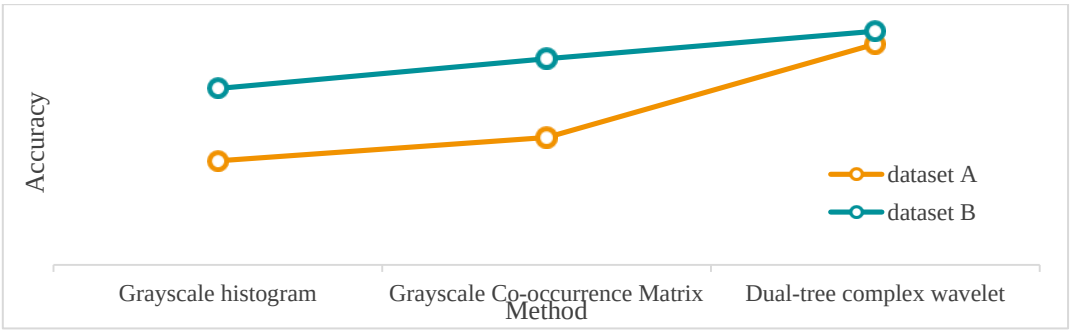

Figure 3. Comparative analysis of the results of different methods on the two datasets

As can be seen from Figure 3, using the SVM classifier classification, compared with the previous two feature representation methods, the dual-tree complex wavelet features are 7.15% and 5.72% higher in dataset A, and higher in dataset B. Out of 3.51%, 1.69%.

5. Conclusion

BI is an evolving modern scientific field created by the intersection of biology, computer science and information technology to support the storage, organization and retrieval of biological data. The field of BI has enabled rapid advances in the fields of genomics, proteomics and systems biology. In short, BI was developed in response to the increasingly complex data types and relationships in biological research, and it addresses the need to manage and interpret biological information. Thus, BI solves biological problems by using information technology to collect, store, organize, and index biological data, presenting, analyzing, and integrating problems with biological data. It applies the principles of computer science and information technology to make large, diverse and complex life science data easier to understand and help to realize its full potential.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Prasad B, Richter P, Wilhelm M F, et al. How the space environment influences organisms: an astrobiological perspective and review. International Journal of Astrobiology, 2021, 20(2):159-177. https://doi.org/10.1017/S1473550421000057

[2] Rai P, Sengupta D, Majumdar A . SelfE: Gene Selection via Self-Expression for Single-Cell Data. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1.

[3] Jamshidi M B, Lalbakhsh A, Talla J, et al. AI and COVID-19: Deep Learning Approaches for Diagnosis and Treatment. IEEE Access, 2020, PP(99):1-1.

[4] Bagabir S A, Ibrahim N K, Bagabir H A, et al. Covid-19 and AI: Genome sequencing, drug development and vaccine discovery. Journal of Infection and Public Health, 2021, 15(2):289-296.

[5] Mmn A, Ss B, Vnnc D, et al. Report of the 1st African Enteric Viruses Genome Initiative (AEVGI) Data and BI Workshop on whole-genome analysis of some African rotavirus strains held in Bloemfontein, South Africa - ScienceDirect. Vaccine, 2020, 38(34):5402-5407.

[6] Ebadfardzadeh J, Kazemi M, Aghazadeh A, et al. Employing BI analysis to identify hub genes and microRNAs involved in colorectal cancer. Medical Oncology, 2021, 38(9):1-8. https://doi.org/10.1007/s12032-021-01543-5

[7] Mandal K, Sarmah R, Bhattacharyya D K . POPBic: Pathway-based Order Preserving Biclustering algorithm towards the analysis of gene expression data. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1.

[8] Paula G, Farias C . A competency question-oriented approach for the transformation of semi-structured BI data into linked open data. Engineering Applications of AI, 2020, 90(Apr.):103495.1-103495.13.

[9] Ascension A M, Arauzo-Bravo M J . BigMPI4py: Python module for parallelization of BD objects discloses germ layer specific DNA demethylation motifs. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1. https://doi.org/10.1109/TCBB.2020.3043979

[10] Mallick K, Mallik S, Bandyopadhyay S, et al. A Novel Graph Topology based GO-Similarity Measure for Signature Detection from Multi-Omics Data and its Application to Other Problems. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1.

[11] Genty J X, Amin M R, Shaw N D, et al. Sparse Deconvolution of Pulsatile Growth Hormone Secretion in Adolescents. IEEE/ACM Transactions on Computational Biology and BI, 2021, PP(99):1-1. https://doi.org/10.1109/TCBB.2021.3088437

[12] Penchovsky R, Pavlova N, Kaloudas D . RSwitch: a novel BI database on riboswitches as antibacterial drug targets. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1.

[13] Paula G, Farias C . A competency question-oriented approach for the transformation of semi-structured BI data into linked open data. Engineering Applications of AI, 2020, 90(Apr.):103495.1-103495.13. https://doi.org/10.1016/j.engappai.2020.103495

[14] Pursalim M, Kwoh C K . An Efficient Multiresolution Clustering for Motif Discovery in Complex Networks. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1.

[15] Sarkar A, Mishra P, Kahveci T . Data perturbation and recovery of time series gene expression data. IEEE/ACM Transactions on Computational Biology and BI, 2021, PP(99):1-1.

[16] Walve R, Puglisi S J, Salmela L . Space-Efficient Indexing of Spaced Seeds for Accurate Overlap Computation of Raw Optical Mapping Data. IEEE/ACM Transactions on Computational Biology and BI, 2021, PP(99):1-1. https://doi.org/10.1109/TCBB.2021.3085086

[17] Er F, Goularas D . Predicting the prognosis of MCI patients using longitudinal MRI data. IEEE/ACM Transactions on Computational Biology and BI, 2020, PP(99):1-1.

[18] Musliji Z S, Pollozhani A K, Lisichkov K, et al. Comparative analysis of genes associated with obesity in humans using bioinformatic data and tools. Balkan Journal of Medical Genetics, 2021, 24(1):35-40. https://doi.org/10.2478/bjmg-2021-0012