Cellular and Molecular Biology
E-ISSN : 1165-158X / P-ISSN : 0145-5680
www.cellmolbiol.org
C M B Association Publisher

Original Research

Design of biological gene information collection system based on data mining technology

Yuanjun Wu[1*], Xiaotong Mu[2], Danial Kahrizi[3]

[1] School of Information Yungui, Anhui Finance & Trade Vocational College, Hefei230601, China
[2] Management and Finance, University of Sussex, BrightonBN19RH, UK
[3] Faculty of Agricultural Science and Engineering, Razi University, Kermanshah, Iran

*Correspondence to: rscwuyj@163.com

Received March 15, 2020; **Accepted** June 27, 2020; **Published** October 31, 2020

Doi: http://dx.doi.org/10.14715/cmb/2020.66.7.15

Abstract: At present, bioinformatics research focuses on the development from the accumulation of biological data to the integration and processing of biological data. This paper designs a bio gene information collection system based on data mining technology. In the system, the information of gene web analysis database, data mining model database and gene chip database is transferred to gene algorithm tool library, which can extract, transform and load the biological gene information, transfer the collected and processed biological gene information to gene general chip and web database analysis logic, and pass it to gene expression spectrum chip/data mining module through API function GUI, through the data mining module GUI feedback to the system users. The system hardware stores the biochip information in the virtual chip set model through the gene expression spectrum data analysis model uses the gene expression similarity analysis model to analyze the expression similarity of the biological gene information, and stores the information in the gene chip database; through the multi-layer structure model, constructs the web genome biochip including the application layer, the data processing layer and the representation layer. The information analysis module analyzes the biological gene information and stores the information in the gene web analysis database. The system software adopts the method of automatic collection of biological gene data based on the web to realize the collection of biological gene information, and gives the main implementation technology of the system. The experimental results show that the system can effectively collect biological gene information, and has high accuracy and anti-noise performance.

Key words: Data mining; Biological gene; Information collection; Gene expression profile; Web analysis; Virtual chipset.

Introduction

With the initial completion of the human genome project and some other model biological genome projects, the focus of bioinformatics research has shifted from the accumulation of biological data to the integration of biological data. Therefore, the construction of the bioinformatics analysis system and its data mining have become a research hotspot in the field of bioinformatics (1-3). However, due to the diversity of biological data and the complexity of its analysis and application, there is no general construction model that can meet the development needs of the biological information analysis system, especially the rapid evolution of biological genome and the rapid growth of data, which put forward higher requirements for data update and mining (4,5). In 1964, Davies pioneered the research of protein structure prediction; in 1970, Needleman and Wunsch published two sequence comparison algorithms which are widely valued; in 1974, Ratner first used theoretical methods to process and analyze the molecular genetic regulatory system; in 1975, Pipas and MC Mahon first proposed the use of computer technology to predict RNA secondary structure; with the emergence of a large number of biological data analysis technologies after 1976, Science published a review on Computational Molecular Biology in volume 209 in 1980.

The data mining algorithm tool refers to a specific data mining algorithm and algorithm necessary data conversion preprocessing work, which is a complete algorithm code unit. An algorithm tool may be used in different analysis models. In the algorithm design, flexible parameter settings need to be provided. In addition to using the configuration tool to set the default value for the analysis model and reduce the work of the user, the user is also allowed to customize the threshold of each algorithm and other parameters. With the development of data mining technology for more than ten years, algorithm research has become mature, and its research focus has shifted from algorithm research to applied research (6). In various application fields, different industries, data mining has been more and more in-depth application, and vigorous development combined with the characteristics of various fields, such as research results in CRM, e-commerce, financial securities and other fields can be seen everywhere in people's lives (7).

The biological information analysis module includes a biological database and analysis tools. Generally speaking, the biological information analysis module refers to a unified platform for users to retrieve, analyze and mine biological information by collecting relevant biological data (including nucleic acid sequence, protein sequence, protein structure or knowledge base), generating a corresponding biological database, developing and integrating relevant analysis tools (8).

Based on the above shortcomings and the creation of predecessors, this paper designs a bio gene information collection system based on data mining technology, which combines data mining technology with bioinformatics research to realize the effective collection and analysis of bio gene information, among which data mining technology mainly includes gene expression spectrum, web analysis and gene algorithm tool library ETL. (A) Gene expression profile chip is an important gene chip. The analysis client of the gene expression profile chip data mining module provides the interface of analysis function, where the user completes the data analysis process. According to the data analysis of gene expression profile, this paper abstracts several analysis models, such as gene expression similarity analysis, gene expression co-occurrence analysis, gene expression path analysis, etc. These analysis models are implemented by one or more data mining algorithms. Through the analysis of the virtual chip set model, a local ETL tool selects one or more individual chips in the existing chip set, extracts and transforms the data through data normalization and data cleaning, and forms the target data for data mining and analysis. (B) The system uses the web biological information analysis model based on multi-layer architecture to study biological information. Aiming at the problems of data management, integration and application of biological information analysis model, on the basis of summarizing the general workflow of the biological information analysis model, a general biological information analysis model based on multi-layer architecture is proposed. This model is based on the general three-layer architecture model. On the basis of this, the data processing layer is added to solve the format transformation, processing, integration, update and other problems of biological data (9). In order to solve the problem of automatic download and update of biological information data in the automatic acquisition of biological data based on the Web, a feasible processing scheme based on network agent is proposed, and the algorithm design and the effect of biological gene information collection are described in detail.

The system gene expression profile data analysis model analyzes the expression similarity of biological gene information, and stores the information in the gene chip database; the web genome biological information analysis module analyzes the biological gene information, and stores the information in the gene web analysis database. After the ETL tool library extracts, transforms and loads the information of gene web analysis database, data mining model database and gene chip database, it feeds back this information to gene expression profile chip/data mining module GUI and feeds back the biological gene information collection results to system users through GUI module.

Materials and Methods

The overall structure of bio gene information collection system based on data mining technology

The structure of database and user management mode (Figure 1) and system structure (Figure 2) were used and they have been explained.

The overall structure of the system is composed of gene web analysis library, data mining model library, gene chip database, data mining algorithm tool library, gene expression general chip data mining module GUI, API function, gene general chip and web database analysis logic.

The data mining algorithm tool refers to a specific data mining algorithm and algorithm necessary data conversion preprocessing work, which is a complete algorithm code unit. An algorithm tool may be used in different analysis models. In the algorithm design, flexible parameter settings need to be provided. In addition to setting the default value for the analysis model with the configuration tool to reduce the user's work, the user is also allowed to customize the threshold of each algorithm and other parameters (10).

ETL tools are data extraction, transformation and load tools. Database and user management client mainly completes the extraction and transformation of external data, as well as the management of gene chip database, data mining model database, gene web analysis database and system user management.

The expression data of gene expression microarray is only a small part of the experimental data, and there is also experimental design, chip design, sample information, hybridization, image information and so on. Generally, they are data files and other database files converted from excel tables, which are generated by chip image processing software or from a laboratory information management system (LIMS). Small labs simply use image processing results.

ETL tools convert these data into database files of the system database schema.

The analysis client of the gene expression profile chip data mining module provides the interface of analysis function, where the user completes the data analysis process. Web service is an interface that can be accessed through the network using XML messages, and it is a distributed environment composed of loosely coupled components. It uses standard Internet protocols (HTTP, SMTP, FTP, etc.) to solve distributed computing based on the Internet, not limited to the intranet. As long as the system supports these network standards, it supports web services.

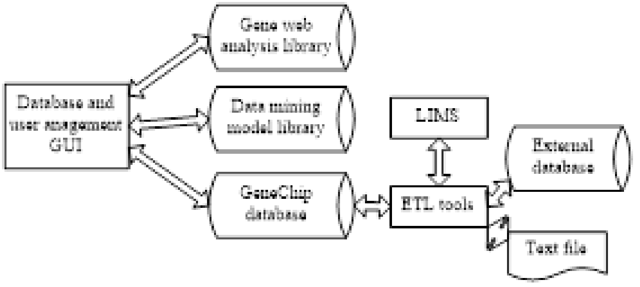
Gene web analysis library
Database and user management GUI
Data mining model library
LIMS
External database
GeneChip database
ETL tools
Text file

Figure 1. Structure of database and user management mode.

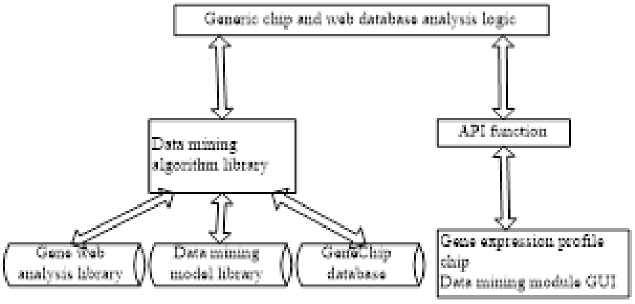
Generic chip and web database analysis logic
Data mining algorithm library
API function
Gene web analysis library
Data mining model library
GeneChip database
Gene expression profile chip
Data mining module GUI

Figure 2. System structure.

Data analysis model of the gene expression profile

According to the expression profile data analysis, the system abstracts several analysis models according to the specific analysis expectation, such as gene expression similarity analysis, gene expression co-occurrence analysis, gene expression path analysis, etc. These analysis models are implemented by one or more data mining algorithms (11,12). The analysis model is an independent specific analysis function, which can be realized by one algorithm, or a processing flow composed of several or several algorithms, including the setting of result visualization, encapsulated as a processing module of a specific analysis task. Different algorithms can be used to complete an analysis function. Different algorithms may have different degrees of excellence. It is necessary to compare them and output the optimal results.

Virtual chipset model

This analysis model is equivalent to a local ETL tool, which selects one or more individual chips in the existing chip set, extracts and transforms data through data normalization and data cleaning, and forms the target data of data mining analysis, which is called virtual chip set. Because its further function is to treat all chips in the database as a large chip when the experimental data is sufficient, select the appropriate gene expression under the corresponding experimental conditions to simulate a chip or chipset (the sequence of multiple chips), extract and simulate the information contained in the existing chip experiment by mining the existing experimental data (13). The new experimental data can greatly reduce the experimental links, accelerate the experimental speed, make full use of the previous experiments, and reduce the research cost.

Similarity analysis model of gene expression

The similarity of sequence, expression and other measurable data often implies the similarity of function, which is one of the important guiding ideas of functional genomics research. According to the similarity of expression changes, we classify the genes in the experiment and deduce the gene function according to the similarity of expression, which is the clue for further experimental verification. The results showed that the expression of similar gene clusters belonged to the same type, such as signal transduction and wound healing. Expression similarity analysis can also be used as the basis of other knowledge inference, such as gene coregulation. Clustering results showed that gene clusters expressed similar genes under various experimental conditions, which could be used for gene function inference, disease prediction, guiding gene chip design, etc. Among them, different clustering algorithms have different sensitivity to the biological data of different species, and the best algorithm can be selected according to the excellent index and other parameters obtained from data experiments.

Analysis module of genomic bio-information based on web

The rapid development of the Internet and Intranet has a huge impact on the operation mode of application modules, from C/S mode to B/S mode, from two layers to three layers, from centralized to distributed. The current general distributed three-tier framework model is shown in Figure 3. The multi-layer architecture is to refine the middle layer of the three-layer framework model, that is, the business logic layer, introduce the concept of components and middleware for structural layering, or according to the function of the business. Based on the analysis of the workflow of the general bioinformatics analysis platform, this paper proposes a multi-layer structure model (BIOCMSM) for building a bioinformatics analysis module. It is mainly to refine the data layer and add a data processing layer to solve the format conversion and data processing of biological data (14). The data management module mainly realizes the functions of automatic download, submission, update and management of biological sequence data; the data application module mainly realizes the functions of release, sharing, retrieval and analysis of biological sequence data. The web-based genomic bio-information analysis module is shown in Figure 4. Each layer of this module framework model is relatively independent, and it can be built on different servers within the LAN. At the same time, this hierarchical division is not physical, but a logical division. Different layers may be implemented in the same physical server, and the same layer may be implemented by different physical servers.

Application layer

The application layer is the traditional business logic layer. Many applications and components silently realize the main functions of the system here. They are hid-

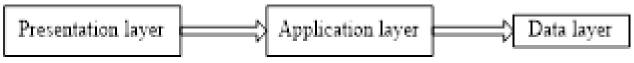

Figure 3. Three-story frame model.

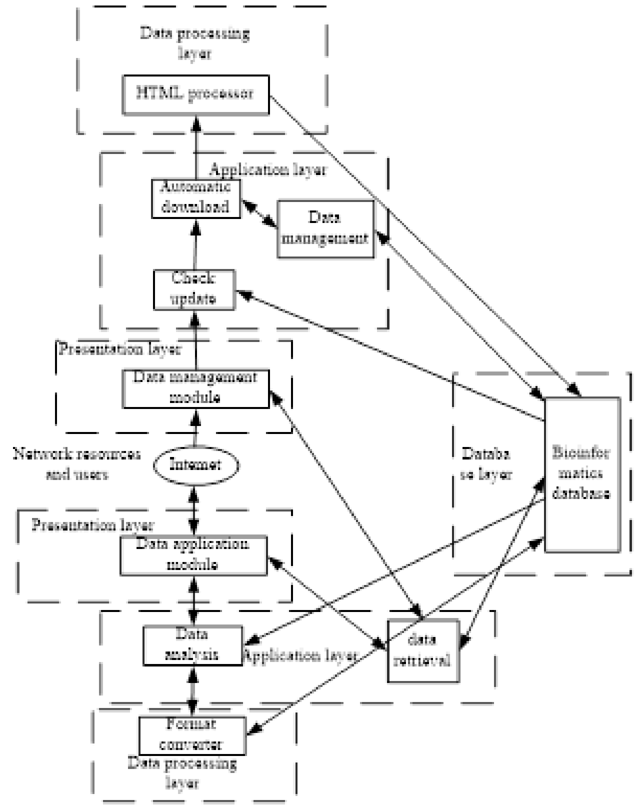

Figure 4. Multi-layer model of bioinformatics analysis platform (biocmsm).

den from the external boundary. They can choose CORBA, DCOM, web services and other ways to realize software reuse and data reuse according to their needs.

CORBA is an organization of OMG (OMG E-sports club, China network media, full name: OH MY GOD). This paper proposes a technical specification to describe object interoperability in a distributed heterogeneous environment. It is mainly divided into three levels: object request agent, public object service and public facilities. The characteristics and disadvantages of CORBA are huge and complex, and the update of technology and standards is relatively slow.

COM is a set of component object interface standards based on the Microsoft Windows platform, which is the foundation of OLE and active X system. DCOM is developed from COM to adapt to the development of distributed computing. It refers to the active X interface of distributed component network application program. Its support for OS platforms other than Windows is not ideal. For example, EBI uses CORBA and Java RMI to publish EMBL sequence data.

The application layer is the traditional business logic layer. Many applications and components silently realize the main functions of the system here. They are hidden from the external boundary. They can choose CORBA, DCOM, web services and other ways to realize software reuse and data reuse according to their needs.

Data processing layer

Data integration is the main problem to be solved in the data processing layer. In this paper, we advocate building a local database (including some file storage systems) with data warehouse as the main method and virtual local database as the consideration. The source of data is the related first-level database or another second-level database. The data processing layer is responsible for parsing, filtering and integrating the data of local interest into the local database, and providing corresponding mechanisms to achieve dynamic update (15). In this way, we try to learn from each other's strengths and make full use of them. As shown in Figure 5, the data processing layer consists of an HTML processor and a format converter. The HTML processor mainly extracts the data in flat-file format from the HTML page automatically downloaded by the application layer and submits it to the database of flat file; and divides the data in a flat-file format to obtain the information of each item needed in the RDB data table and submits it to the database of RDB. The main purpose of the format converter is to provide the data format needed by each application in the upper application layer. (See Figure 6 for the design of key module HTML processor).

Presentation layer

The presentation layer consists of a data management interface and a data application interface. Data management: provide data update, data retrieval and data management interfaces for users to realize the interaction between users and local system; data application: generate different user interfaces according to different user requirements, and realize the interaction between users and local system through web-based user access interface C/S mode is convenient for background data management, and data security is easy to guarantee. B/S mode is simpler for the sharing and utilization of biological information resources. Through a plug-in, Java applet, active X, JavaScript and other technologies, we can make up for the lack of HTML standard functions.

System software design

Main implementation technology and system characteristics

Main implementation technology: according to the development direction of data mining, the system establishes three-tier architecture, adopts component technology, supports PMML (predictive model markup language), XML (Extensible Markup Language) standards, and combines data mining technology with bioinformatics research.

After more than ten years of development, the research of data mining has changed from the research of algorithm and general data mining tools to the combination of application and started the research of data mining standardization. Now the W3C has the data mining language standard PMML. Similarly, there are many kinds of biological information. Most of the data are stored in a heterogeneous biological database. The network characteristics of biological information are difficult to reflect. Standardization is one of the urgent problems in bioinformatics, and it is the basis of many problems XML is becoming the standard computer language of bioinformatics. The system supports the XML standard in the data source and adopts XML format in the mining process and result representation.

In the specific implementation, Java is chosen as the development language, and EJB (enterprise JVA bean) technology is used in the implementation of the data mining algorithm and analysis model. Java is robust, safe, easy to use and so on. Because of its good cross-platform ability, it has been favored by more and more software developers. Among them, EJB Technology provides good support for the multi-layer architecture of this paper. EJB Technology combines object technology and component technology and conforms to the

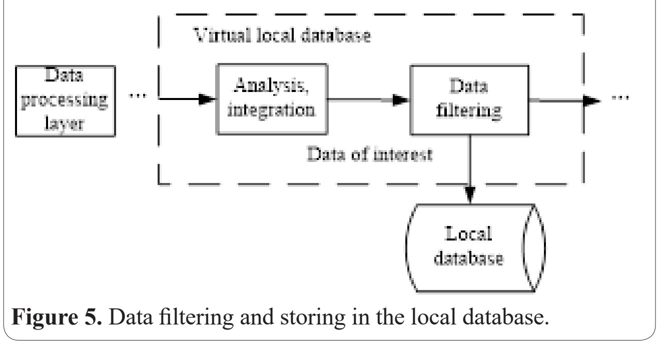

Figure 5. Data filtering and storing in the local database.

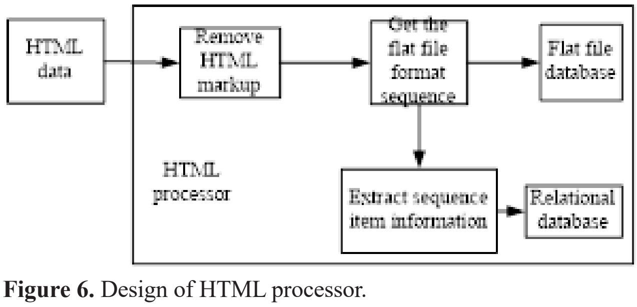

Figure 6. Design of HTML processor.

new programming method component programming method. EJB can be distributed in the network and run across computers. It can be called through the network. Users don't need to know about the internal components. They only need to know the interface parameters and configure the components to complete the required work.

Because of these characteristics of EJB Technology, the function of the biological gene information collection system is easy to expand. In the process of gene chip research, the system function can be continuously enriched and improved, the existing algorithm can be updated, the interface definition can be maintained and replaced, the system can introduce a new and better algorithm, and complete the needs of biological data processing. In the analysis logic layer, according to the expanding application field of gene chip, a new analysis model is built, and a special analysis platform for different applications is differentiated. In the process of collecting and analyzing biological gene information (16), the automatic download and update of biological information data are faced, and a feasible processing scheme based on a web agent is given. The algorithm design and implementation of the scheme are described in detail. In the practical application of automatic acquisition of biological data, good results have been achieved. At the same time, the technology is not limited to a certain field of biotechnology research and has good versatility (17).

According to the requirements of the project application unit, the system uses IBM DB2 as the database development platform and connects the database through JDBC. JDBC makes it easy for developers to connect to the database and use SQL for database operation. The combination of Java and JDBC can realize one-time writing and can be used everywhere.

System characteristics

Through the selection of system architecture and microarray analysis model design and implementation technology, the system mainly has the following characteristics:

(A) Multi-tier architecture. The biological gene information collection system adopts multi-layer system structure and component technology, and realizes the separation of algorithm and application at the user level; it adopts Java language and EJB Technology, which is platform-independent, distributed and scalable, which provides convenience for system expansion and secondary development This architecture well supports Mr. Luo Jingchu's discussion on three kinds of people in bioinformatics research, that is, the division of work and cooperation among biological researchers, algorithm researchers and software researchers.

(B) Not related to the platform. Due to the good cross-platform characteristics of Java, the system is independent of the platform and can run in either Windows or Linux environments. At the same time, Java multithreading technology provides a certain degree of parallelism to speed up the operation of the algorithm.

(C) System scalability and entity independence. The system adopts multi-layer architecture, and algorithm and analysis become two levels of concepts and entities. In the implementation, using java EJB Technology, each algorithm tool and analysis model is an independent distribution unit. In the data mining algorithm tool layer and analysis logic layer, new entities can be added and embedded into the system seamlessly.

(D) Rich analysis function. The system provides a variety of expression spectrum analysis functions, covering typical expression spectrum analysis expectations. For each analysis model, the corresponding visualization method is used to intuitively display the analysis results (18), which is convenient for users to understand and apply.

The workflow of biological information analysis based on web

In order to construct the analysis module of genomic biological information, we should not only solve the problems of biological data integration and application integration but also fully consider the difficulties of data updating caused by the rapid evolution of genes. Combined with the above analysis, this paper presents the general workflow of a biological information analysis module (as shown in Figure 7): use network resources to achieve biological data management, transfer biological information database through biological data management, conduct biological data analysis and feedback to network users.

Automatic collection of biological data based on web

(A) Relevant technologies for automatic acquisition of biological data

A feasible solution is to let computer programs replace people to query and download data. While using browsers to access the Internet, there are also some special network users working on the Internet. These users are Internet programs. There are many kinds of Internet programs (Figure 8) that perform different special functions. For example, Google, a famous search engine, uses a spider program to traverse the web site to create and maintain a large website database. In contrast, Intuit's financial software, quicken, use the aggregator program to view users' multiple financial and credit account information.

The agent program is another kind of Internet program which is often used. It can set a series of keywords, scan specific information sources and find specific information that users are interested in. The manual query often requires users to constantly select and judge and respond to the results, while the agent program automatically completes such a process. It uses a series of predetermined ways to judge instead of users, shuttle

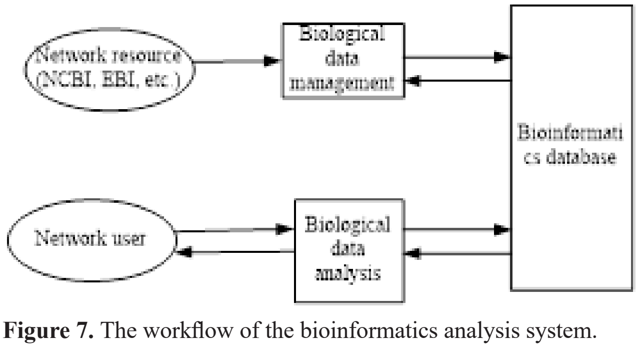
Network resource (NCBI, EBI, etc.)
Biological data management
Bioinformatics database
Network user
Biological data analysis

Figure 7. The workflow of the bioinformatics analysis system.

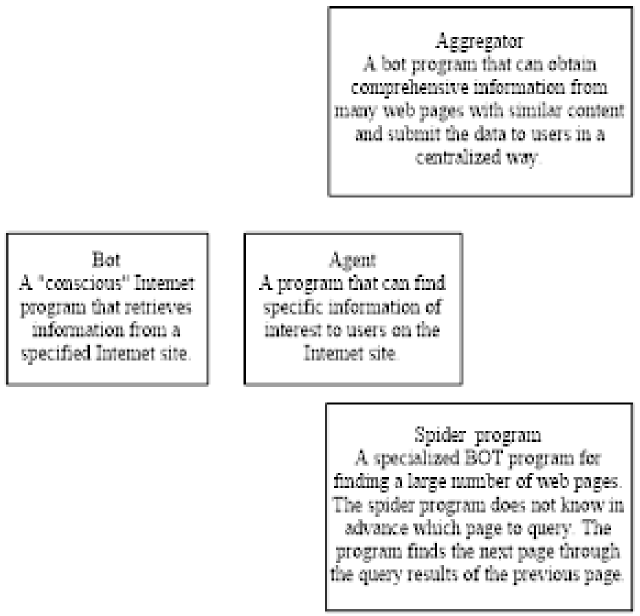
Aggregator
A bot program that can obtain comprehensive information from many web pages with similar content and submit the data to users in a centralized way.
Bot
A "conscious" Internet program that retrieves information from a specified Internet site.
Agent
A program that can find specific information of interest to users on the Internet site.
Spider program
A specialized BOT program for finding a large number of web pages. The spider program does not know in advance which page to query. The program finds the next page through the query results of the previous page.

Figure 8. BOT Program, spider program, aggregator program and agent program.

and collects the required information on the first level database website.

Most agents rely on the subtleties of a site-specific web interface, and changes in the format of data publishing or the insertion of other information can cause them to fail to work. Fortunately, most public bioinformatics level 1 databases maintain one or several fixed data formats that may not change in a few years. For example, NCBI has detailed and strict regulations on the data publishing format of GenBank, the nucleic acid sequence database maintains: each sequence entry consists of fields, some of which are divided into several subfields; each field starts with an identifier, followed by a specific description of the field, see Table 1; the web publishing of entries starts with the identifier "locus" and uses the double slash "//" End tag; the primary identifier starts at column 1, the secondary identifier starts at line 3, the property table specifier starts at line 5, and so on (see ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt).

Obviously, GenBank's data publishing is not designed for computer reading. This chapter mainly designs a complete set of computer programs to realize automatic sequence downloading, and analyzes, transforms and extracts its records, that is, the check update, automatic download and HTML processor three modules in BIOCMSM.

(B) Algorithm design

At present, we use the nucleic acid sequence database and protein sequence database supported by NCBI as the data source of a professional secondary database. NCBI is one of the main life science information service institutions in the world. Every day, a large number of sequence data from relevant laboratories and sequencing institutions enter its database, and maintain the data exchange and update with other databases (such as EMBL, DDBJ, etc.), so it collects all the current open nucleic acid and protein sequences. In addition, NCBI data resources can be freely used (on the premise of non-profit education and scientific research), so there is no user verification problem (19). The functional structure of the whole algorithm mainly includes three parts: check and update, automatic download and HTML processor. The check and update module mainly realizes the comparison between the local database and NCBI database and prompts whether the local database needs to be updated. The automatic download module is mainly used to download HTML documents corresponding to new sequences. The HTML processor mainly analyzes and processes the downloaded HTML documents, and extracts valuable information to the bioinformatics database. The main flow of the algorithm is shown in Figure 9.

Results

Bioinformatics sequence retrieval

The biological information sequence retrieval interface of the system is shown in Figure 10. It can be seen that the sequence retrieval includes two parts: simple retrieval and advanced retrieval. The simple retrieval is mainly based on the sequence GI number, gene type and other conditions, for example, NDV genome contains six kinds of genes, such as F, L, M, N, HN,

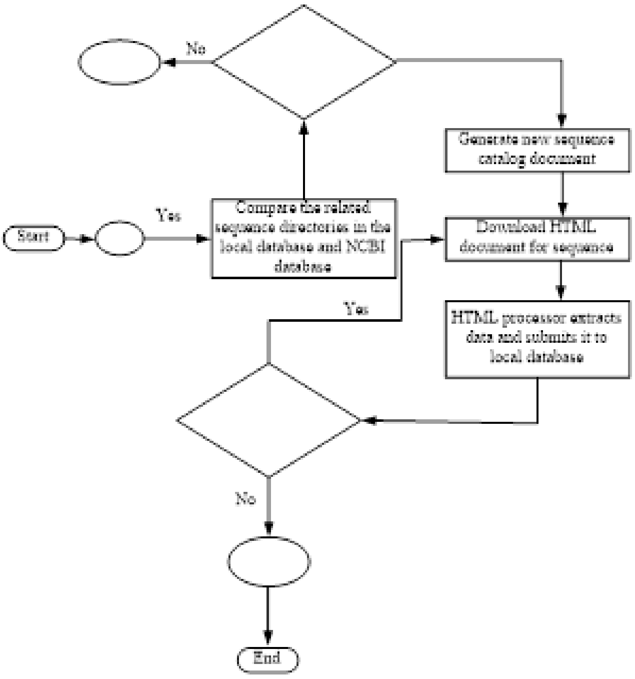
No
Generate new sequence catalog document
Start
Yes
Compare the related sequence directories in the local database and NCBI database
Download HTML document for sequence
Yes
HTML processor extracts data and submits it to local database
No
End

Figure 9. Flow chart of automatic acquisition of biological data.

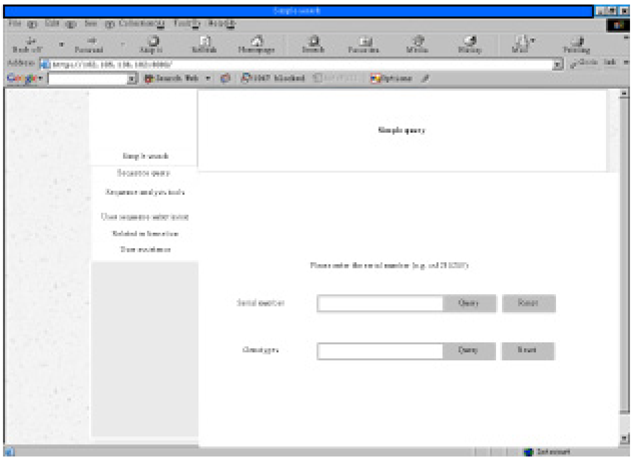

Figure 10. Bio information sequence retrieval interface.

Table 1. Main fields of GenBank database entries.

Identifier	Locus	Definition	Accession Version	Keywords	Source Organism	Reference	Authors Title	Journal	Medline Comment	Features	Base Count Origin
Content	Sequence name, length, molecular type, sequence category, modification date	Simple definition	Serial number serial version number	Related words of sequence	Scientific name and taxonomic location of species origin	Reference number	Title of the author's reference	Journal name of the reference publication	Reference MEDLINE citation code comments on sequences	Sequence property table (children)	Nucleotide number statistical sequence

P; the advanced retrieval provides the composite query of subsequence and gene type, for example, the query of F gene sequence contains the sequence of similar sequence “ATGAGGTAA”.

Biological gene collection results

In this paper, the collection system characterizes the cap gene chip, collects and stores the hybridization results of the cap gene chip (Figure 11), and effectively realizes the gene collection of the cap gene chip.

Acquisition performance analysis

Analyze the results of 10 different biological gene information collected by the system under different signal-to-noise ratio, and describe them with Table 2, Table 3, Table 4 and Table 5 respectively. It can be seen from the results in the four tables that, with the increase of signal-to-noise ratio, the error rate of biological gene information collection results of the system in this paper increases gradually, and the increased range is not large, always less than 2%. At the same time, it can be seen from Table 6 that the average error of the system in this paper is the lowest compared with other systems, and the average error rate of the system in this paper is always lower than 1.9%. It shows that this system can effectively collect biological gene information, and has high accuracy and anti-noise performance.

Discussion

With the implementation of large-scale scientific projects such as the human genome project and human brain project, how to deal with mass data is an urgent

Table 2. Collection results of biological gene information (10 dB).

Gene name	Error rate (%)
Lung cancer tissue gene	1.05
Prostate cancer-specific genes	1.03
Hypertension related genes	1.02
Guanling cattle myosin heavy chain 1 gene 5	1.01
Ginkgo whole gene	0.81
Osteomodulin gene of fat lambs Heishan sheep	0.54
TYRP1 gene of raccoon dog	0.96
CMV melon isolate coat protein gene	1.04
Tobacco carbonic anhydrase (CA) gene	1.06
MITF-M gene of blue fox	0.69

Table 3. Collection results of biological gene information (30 dB).

Gene name	Error rate (%)
Lung cancer tissue gene	1.07
Prostate cancer-specific genes	1.05
Hypertension related genes	1.05
Guanling cattle myosin heavy chain 1 gene 5	1.03
Ginkgo whole gene	0.85
Osteomodulin gene of fat lambs Heishan sheep	0.71
TYRP1 gene of raccoon dog	0.99
CMV melon isolate coat protein gene	1.07
Tobacco carbonic anhydrase (CA) gene	1.09
MITF-M gene of blue fox	0.89

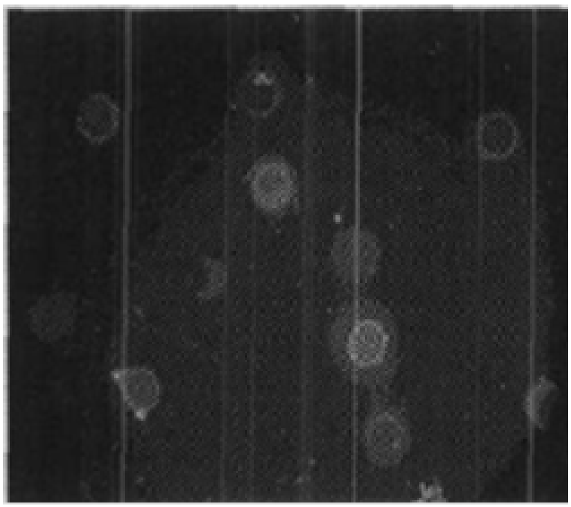

Figure 11. Detection results of gene chip.

Table 4. Collection results of biological gene information (50dB).

Gene name	Error rate (%)
Lung cancer tissue gene	1.09
Prostate cancer-specific genes	1.09
Hypertension related genes	1.09
Guanling cattle myosin heavy chain 1 gene 5	1.06
Ginkgo whole gene	1.10
Osteomodulin gene of fat lambs Heishan sheep	1.31
TYRP1 gene of raccoon dog	1.02
CMV melon isolate coat protein gene	1.14
Tobacco carbonic anhydrase (CA) gene	1.14
MITF-M gene of blue fox	1.34

Table 5. Collection results of biological gene information (70dB).

Gene name	Error rate (%)
Lung cancer tissue gene	1.13
Prostate cancer-specific genes	1.12
Hypertension related genes	1.15
Guanling cattle myosin heavy chain 1 gene 5	1.14
Ginkgo whole gene	1.16
Osteomodulin gene of fat lambs Heishan sheep	1.89
TYRP1 gene of raccoon dog	1.14
CMV melon isolate coat protein gene	1.31
Tobacco carbonic anhydrase (CA) gene	1.26
MITF-M gene of blue fox	1.68

Table 6. Collection results of biological gene information.

Gene name	Average error rate (%)
Lung cancer tissue gene	1.09
Prostate cancer-specific genes	1.07
Hypertension related genes	1.08
Guanling cattle myosin heavy chain 1 gene 5	1.06
Ginkgo whole gene	0.98
Osteomodulin gene of fat lambs Heishan sheep	1.11
TYRP1 gene of raccoon dog	1.03
CMV melon isolate coat protein gene	1.14
Tobacco carbonic anhydrase (CA) gene	1.14
MITF-M gene of blue fox	1.15

problem, which also drives a huge mass storage market. The key problem is how to design a special mass storage technology for biological information. Due to the lag of technology development, the effective utilization rate of biological information resources is very low, which seriously affects the utilization of biological information. Information access has become a challenging problem in biology, but also a challenge to computer science. In the post-genomic era, biologists often need to find data from public databases through the web and build their specific systems or theme-oriented databases to support their complex analysis and knowledge discovery. Therefore, it is an important problem to access, extract, transform and integrate these heterogeneous Web data. Some large biological databases often use their own data storage formats and data operation programs, such as the SRS system and Entrez system, which integrate various types of databases, provide a unified interface and query method, and realize information sharing. The development ideas of these systems are worth learning, but they are too complex in data management, update and maintenance, and they are not targeted. The development is based on specific data models, such as ASN. 1 (Abstract Syntax Notation One) adopted by Entrez, which is an informative description and conversion language selected by NCBI about ten years ago from the telecommunications industry, with limited expressiveness and lack of tools. It cannot meet the requirements of the construction of the special bioinformatics system (20-23).

Aiming at the above problems, this paper designed a bio-gene information collection system based on data mining technology. In this paper, the data mining based biological gene information collection system was designed, which uses web and gene expression profile as the tool of biological gene information collection and analysis. Through the gene expression profile data analysis model, the biochip information was stored in the virtual chip set model, the information contained in the existing chip experiments was mined, and the new experimental data was extracted and simulated from the existing experimental data, which can greatly reduce the experimental links, accelerate the experimental speed, make full use of the previous experiments, and reduce the research cost. Through gene expression similarity analysis model to analyze the gene expression rules under different conditions, according to the similarity of expression changes, we classify the genes in the experiment and deduce the gene function according to the similarity of expression as a clue for further experimental verification. We can conclude that different clustering algorithms have different sensitivity to biological data of different species. The best algorithm is chosen by the good index and other parameters. The application layer is the traditional business logic layer. Many applications and components silently realize the main functions of the system here. They are hidden from the outside. According to the needs, we can choose CORBA, DCOM, web services and other ways to achieve software reuse and data reuse. Data layer: the source of data is the related first-level database or other second-level databases. The data processing layer is responsible for analyzing, filtering and integrating the data of local interest into the local database, and providing the corresponding mechanism to achieve dynamic update. In this way, we try to learn from each other's strengths and make full use of them. Presentation layer: it mainly consists of a data management interface and data application interface. Data management: provide data update, data retrieval and data management interfaces for users to realize the interaction between users and local system; data application: generate different user interfaces according to different user requirements, and realize the interaction between users and local system through web-based user access interface. Through the search of the biological information sequence of the web, the collection system of this paper collects, transfers and explains the results of hybridization of gene chip and 10 kinds of different biological gene information. This system can effectively collect biological gene information, with high accuracy and noise resistance.

Data mining technology has the advantages of a wide range of use and accurate data mining. At the same time, biological gene collection is the main research direction in the field of the biological gene. The combination of the two forms an efficient and accurate biological gene collection method. Based on the data mining technology, the biological gene information collection system designed in this paper takes the web and gene expression spectrum as the basis of biological gene information collection and analysis, and uses the gene algorithm tool library ETL to extract, transform and load the information of gene web analysis database, data mining model database and gene chip database, and uses the gene expression spectrum chip/data mining module GUI for the system. The results of the collection of biological gene information from household feedback. Using data mining technology is the innovation of this system, which is different from other biological gene information collection systems, and effectively improves the accuracy and anti-noise of biological gene information collection.

Acknowledgments

Key Projects of Natural Science Research of Colleges and Universities in Anhui Province, 2016 (KJ2016A010); Key Projects of Natural Science Research of Connotation Enhancement Plan in Anhui Finance and Trade Vocational College, 2017.

References

1. Ju YS, Tae MS, Dong CS. Data mining of web-based documents on social networking sites that included suicide-related words among Korean adolescents. J Adolesc Health 2016; 59: 668-673.

2.Thomas C, Elisabeth C, Francesco I. GDSCtools for mining pharmacogenomic interactions in cancer. Bioinformatics 2017; 34: 1226-1228.

3. You XW, Yao T, Xing QZ. NOSEP: nonoverlapping sequence pattern mining with gap constraints. IEEE Trans Cybern 2017; 48(10): 2809-2822.

4. Zheng WH, Mao ZL, Christos C. Schema theory-based data engineering in gene expression programming for big data analytics. IEEE Trans Evol Comput 2017. 22(5):792-804.

5. Victor WC, Raymond KW, Chi HC.The design of a cloud-based tracker platform based on system-of-systems service architecture. Inf Syst Front 2017; 19: 1-17.

6. Wang L, Chen Q, Gao H. Framework of fault trace for smart subs-

tation based on big data mining technology. Autom Electric Power Syst 2018; 42: 84-91.

7. Mei YD. The application of data warehouse model based on the integration of emerging technologies in business data mining. J Comput Theor Nanosci 2016; 13(12): 9581-9585.

8. Lázaro B, René C, Raudel H. On the design of hardware-software architectures for frequent itemsets mining on data streams. J Intell Inf Syst 2017; 50: 1-26.

9. Mohamed I, Afdel K, Mustapha B. Distributed intrusion detection system for cloud environments based on data mining techniques. Procedia Computer Sci 2018; 127: 35-41.

10. Hong BX, Hai XW, Ming ZQ. In silico drug repositioning for the treatment of Alzheimer's disease using molecular docking and gene expression data. RSC Adv 2016; 6: 98080-98090.

11. Hai XX, Fei YY, Sheng H. Erratum: Bayes performance of batch data mining based on functional dependencies. Int J Pattern Recog Artif Intell 2019; 33(03): 1959011.

12. Bin Y, Qiran W, Xue MW. On extraction of cancer informative genes and gene expression data mining. J Bionanosci 2016; 10: 293-299(7).

13. Bao, HX, Ting, L. Physical health data mining of college students based on DRF algorithm. Wirel Pers Commun 2018; 102: 1-11.

14. Lu ZX, Li JN, Li X. Research on classification and recognition method of ionospheric phase contamination. J China Acad Electron Inform Technol 2016; 11: 503-509.

15. Yuan H, Ding XP, Wang BR. Research on effect on dc-dc converter caused by parasitic resistance. J Power Supply 2016; 14: 86-94.

16. Yao ZM, Pan F, Yu XH. Energy-saving and emission-reduction display platform design of PV power plant based on ZigBee wireless acquisition. Chin J Power Sources 2016; 40: 1993-1996.

17. Luo NH, Tao JY. Liu JQ. Research on heterogeneous monitoring between business systems based on big data outlier mining algorithm. Autom Instrumen 2019; 179-182.

18.Tian BH, Yi L. Wang XW. Construction of phage display of antigenic variability gene fragments library of bird flu virus. J Jilin Univ (Science Edition): 2019; 989-996.

19.Wu HL, Ren XY. Specific data mining algorithm based on fuzzy constraint database. Comput Simul 2016; 33: 240-243.

20. Talat F, Wang K. Comparative Bioinformatics Analysis of the Chloroplast Genomes of a Wild Diploid Gossypium and Two Cultivated Allotetraploid Species. Iran J Biotechnol 2015; 13(3): 47-56.

21. Son J, Jeong H, Lee E, No S, Park D, Chung H. Identification of specific gene expression after exposure to low dose ionizing radiation revealed through integrative analysis of cDNA microarray data and the interactome. Int J Radiat Res 2019; 17 (1) :15-23.

22. Kuai H, Zhong N. The Extensible Data-Brain Model: Architecture, Applications and Directions. J Comput Sci 2020: 101103.

23. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019; 6(1):54.