The Framework of Bioinformatics Platform for Gastric Cancer Gene Chip Based on Clustering Analysis Algorithm

Fu'an Wang

Pingdingshan University Medical School, Pingdingshan 467000, China

Abstract

At present, high-performance computers have been widely applied in meteorology, ocean current analysis, military and other fields. In the study of biological gene chips, genomic alignment can also be supported by the operational efficiency of high-performance computers. For this purpose, this study proposes the construction of a bioinformatics platform for gastric cancer gene chips, builds on high performance computers as the operating conditions, integrates with the hierarchical clustering analysis algorithm to support the timeliness and authenticity of the results, promotes the data expression of more genes in the gastric cancer gene chip detection, analyzes the massive information generated by the genome data to achieve the optimal results of parallel computing, and adopts parallel computing to reduce the analysis time and improve the detection efficiency of gastric cancer gene chip data processing.

Keywords: Gene Chip, Biological Information, Hierarchical Clustering Algorithm, Gastric Cancer Gene, System Framework.

1. RESEARCH BACKGROUND

1.1 Literature review

The continuous progress of computer technology has provided support for the technology development of biological information and gene chip. The study on the bearing mode and the transformation method of biological information can afford more accurate data support for the medical field, achieve the optimal mode for sharing information resources, and realize the practical effect of multi-party consultation (Zhang et al., 2017). The pattern of manifestation of gastric cancer gene chip (GCGC) in biological information and the exploration of the mode of information transmission are also the key directions in the fields of medicine and computer technology. The research goal is to summarize the formation mechanism of gastric cancer gene and generalize the production conditions of the relevant factors in order to provide a data reference for DNA database (Cheng et al., 2011). Relatively speaking, gene chips hold the advantages of shorter time and higher detection effect in data processing functions. Thus, how to use these advantages to create a bioinformatics platform for the GCGC is also the most critical academic research direction.

1.2 Research objective

During the biological information transmission of biological gene chips, the data information and types of gastric cancer genomes need to be sorted according to the order of their genomes (Xiong et al., 2015). The sequencing principle of the gene chip is the hybrid sequencing method. Specifically, nucleic acid sequence determination is conducted by hybridizing a group of nucleic acid probes with known sequences, and the target nucleotide probe with a known sequence is immobilized on the surface of a substrate (Wang et al., 2012). When the solution contains fluorescently labeled nucleic acid sequence, TATGCAATCTAG, and it matches with the nucleic acid probe at the corresponding position of the gene chip, a set of fully complementary probe sequences are obtained by determining the probe position with the strongest fluorescence intensity, and the sequence of the target nucleic acid is accordingly recombined. Furthermore, the data volume of the GCGC computing conditions is large. For this reason, this study selects the hierarchical clustering analysis algorithm to create a better computing environment and effectiveness for the GCGC bioinformatics platform.
2. THE OVERVIEW OF GCGC TECHNOLOGY

2.1 HPC and its related technologies

High-performance computing (HPC), also known as parallel computing, specifically refers to completing load processing and data analysis of tremendous amount of information on high performance computers. Under the existing hardware conditions, the advantage of parallel computing can be applied to reduce the effect of physical limits on the arithmetic constraints, thus completing the serial computing task at a lower time cost (Jin et al., 2017). The related information processing and calculation of gastric cancer genes are featured with the maximum computation burden of massive data information. If simply taking the computer operation mode can not achieve the basic conditions of fast computing, in the application process of HPC, it is necessary to integrate the MPI messaging interface technology, an information transfer application programming interface that includes protocol and semantic descriptions and indicates how to perform its features in various implementations. MPI technology possesses high performance, large scale, portability and other features and can provide HPC with higher support for the processing operation of data information, so as to achieve the actual effect of providing operational efficiency.

2.2 Gene chips and biological information

Biological information plays a vital role in the survival and reproduction of living organisms and covers a wide range. In addition to genetic materials, electrical nerve impulses and hormones, sound, odour, colour and behaviour of organisms contain information, have an impact on the individuals and groups of organisms, and are inseparable from the survival and evolution of living beings. The feature of biometric information is that a minimal amount of energy and material consumption can exert a significant biological effect. It is more intuitive and practical to describe the formation mechanism of gastric genes by means of biological information (Dong et al., 2011). Gastric cancer, as one of the most prevalent incidences in the current clinical medicine, ranks the first among all gastrointestinal malignancies. The means of identifying its etiology and pathogenesis and seeking an early diagnosis and effective drug targets is particularly important. The current study suggests that gastric cancer is a complex multi-factor and multi-stage process involving a substantial amount of structure changes and abnormal expression of related tumour genes. Gene chip technology analyzes tens of thousands of gene expression profiles in a fast, accurate and high-throughput way, locates the advantages of differentially expressed genes, and demonstrates its application value in the biomedical field such as cancer. Thus, the GCGC is applied to record its genetic characteristics and summarize its biological information, which is the technology foundation of constructing the information sharing mechanism for the GCGC bioinformatics platform. To this end, the study on the GCGC also focuses on the related direction of information transmission methods.

3. DATA PROCESSING MODEL FOR THE GCGC BASED ON CLUSTER ANALYSIS

3.1 Data processing of the GCGC

In the process of analyzing and processing data information, the GCGC should take the detailed process of standard streamlining or information categorization as a basis for analyzing data, and information with biological significance is then extracted from the categorized data (Liu et al., 2016). Under normal conditions, the DNA matrix with double fluorescence staining is utilized to arrange biological information and the GCGC is taken as the reference gene or sample gene. The constraint conditions of biological information are distinguished by fluorescent markers of different colours, and gene expression forms and quantitative criteria under different conditions are differentiated on basis of the comparison between sample gene and reference gene.

3.2 Data processing classification of gene chip

First, data processing should be conducted to the chip image information, and the images of the GCGC are scanned by the optical scanner. Meanwhile, the basic operation of the extracted sample points provides a reference for the biological sample of data information in order to summarize the statistical analysis results. At this point, the chip storage conditions of the related data information are obtained and quantitative indexes are provided for the contents of multiple sets of data (Liang et al., 2016). Secondly, the data of gastric cancer gene should be pretreated. Low-quality information exists in the preprocessing. To achieve the accuracy of data information, filtering and filtering must be performed according to the type of data and its main operation objects are negative data and zero.
In the end, through the standardization of gastric cancer gene data, the final effect of gene chip information collection can be achieved. The goal of the operation is to eliminate the deviation of gene expression caused by the technical defects of the experiment, to promote the consistency of the information of multiple samples with the experimental data, and to consequently achieve the effect of controlling the output standard of biological information. The final result of data standardization provides the operating conditions and basis for the standardization of inter-chip data and parallel experimental data.

3.3 Operational model for data standardization

In designing the bioinformatics platform of GCGC, it is necessary to standardize a variety of information materials to support system identification and unified deployment (Li et al., 2013). There are three major types of data processing standardization, namely inter-IC data standardization, parallel test data standardization and on-chip data standardization.

First, the inter-chip data standardization is to adjust multiple sets of experimental data information into a unified index, take the average algorithm to obtain the index type of the inter-group data, and set the scalar of the information collection sample within its unified index. The index computing model for R/G data set is:

$$\log(R/G) = \log_2(R/G) - \text{mean}_a$$

(1)

In this formula, mean, represents the average value of the data samples in each group. When this average value is taken as the sample parameter of the evaluation data index and meets the unified quantitative standard, the output conditions with a high consistency can be created for the inter-group data.

Secondly, the standardization of parallel test data is based on the evaluation on the data difference of the same sample in multiple tests. When test data illustrates a large gap between the output values for a few times, it is necessary to calculate the average value by means of parallel test data standardization, so as to achieve the unity of the test data (Diao et al, 2011). The computing model is:

$$\log(R/G) = \left[ \frac{\sum_{i=1}^{n} \log(R/G)^i} {n} \right]$$

(2)

In this formula, n represents the number of experiments that have acquired valid experimental data information and i stands for the optimal experimental results sequence within the number of experiments. In the course of its operation, the completion degree of information acquisition conditions of parallel experimental data can be gradually analyzed, and furthermore, the average value of the constraints can be defined as a reference data type, thereby improving the precision and the standardization result of the calculated value.

In the end, on-chip data standardization refers to the index sample consistency of the biological information stored in the GCGC, and its operation objective is to reduce the error value of the individual chips in the storage unit so as to achieve the best effect of uniformly outputting the sample parameters. Local data weighting algorithm can be used, and its regression model is:

$$\log(R/G) = \log_2(R/G)^T - \log_2(\sqrt{R*G})$$

(3)

3.4 Cluster analysis model

Cluster analysis aims to collect data for classification on the basis of similarity. Clustering originates from a number of fields, including mathematics, computer science, statistics, biology, economics, etc. In different application fields, numerous clustering techniques have been developed and used to describe the data, measure the similarity between different data sources, and classify data sources into different clusters. The quantification standards of gene chip technology targeted for gastric cancer gene can be analyzed by the operational mode of clustering analysis. To this end, this study applies the clustering analysis algorithm to calculate the expression data of gene chip, divides the genotypes by function, and categorizes the expression direction (Yang et al., 2015). The hierarchical clustering analysis model applies the gene chip clustering algorithm at the earliest. This type of algorithm establishes the phylogenetic tree by calculating the correlation coefficient between any two groups of
genes and comparing the weights corresponding to gene positions so as to explore the spatial distance between genes. The correlation coefficients of gastric cancer genes are obtained and a distance matrix is generated. The computing model is:

\[ U(x, y) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i - x_{\min}}{\Theta_x} \right) \left( \frac{y_i - y_{\max}}{\Theta_y} \right) \] (4)

By substituting the previous standardized data information into this formula, the spatial coordinates of the desired gene sample, \((\Theta_x, \Theta_y)\), can be derived from the average value. Also, the weighted average algorithm is applied to verify the differences in the expression of the operation results. In this way, data information can be obtained and applied to the storage and transmission of gastric cancer gene data and accordingly fulfill the expected platform information operation standard.

4. SYSTEM FRAMEWORK OF GCGC BIOINFORMATICS PLATFORM

4.1 Module partition of platform function

In view of the module partition of platform function, this study divides the calculation module into two sets of functions, namely, the standardized calculation function and the clustering calculation function. Corresponding to the two steps of gene chip data processing—standardization and clustering, the module functions are completed by the compute nodes. Respectively, data and information are sent and exchanged in turn during the same period of time, and no data is repeatedly exchanged among the compute nodes, which is provided to its functional modules. Furthermore, the compute nodes should be guaranteed to communicate with each other only one time to avoid double counting, which lays a high data processing foundation for clustering calculations (Ma and Zhang, 2014). The functions of matrix division and collection module are taken as the condition of the completion degree of node operation, and its data types are applied in the parallel computing of distributed computing nodes, in order to eventually achieve the goal of collecting and organizing gastric cancer genome information, transmitting the target matrix information and returning to the user layer. Figure 1 demonstrates the module partition of system functions.

![Module partition of platform function](image)

**Figure 1.** The Main Frame Structure of the Platform Functional Module Partition

4.3 System architecture process

The essence of gene chip data processing is large-scale matrix operation. Based on the actual needs of the objective of GCGC data processing, the processing results should be assigned to the correlation coefficient values and the source microarray positions are searched according to the source data to further ensure that two groups of source data values have a reference value (Wu and Yan, 2015). In the process, all source data should be integrated into a one-dimensional array order, and number labels are used to mark the position of source microarray. The spatial
positions of the genomes in its sequence are respectively recorded. For example, five tissue data of G₁ is at the first row of the sequence and G₂ is at the second row. By parity of reasoning, the one-dimensional array identifier is identified by the sequence number of the gene. The design idea is that the host node gives one-dimensional serial numbers to the data in the source microarray and then distributes to compute nodes in proper order. Meanwhile, parallel computing of normalization and clustering is performed on the data distributed to the local at each compute node. In the end, the result of calculation and the location of corresponding source data are sent to the host node that organizes the data into a target matrix and feedbacks the data information to users, which achieves the effective operation of the biological information sharing mechanism of the GCGC data and supports the synchronous updates of platform information.

4.4 Platform operating system process

The operating system process of the GCGC bioinformatics platform adopts the node sequence 0 as the host node, takes charge of data preprocessing of the source microarray after the data is filtered, and then distributes it to the compute node based on the tissue chip sequence. Since the data information in n organizations of data standardization is basically independently calculated, it can be divided into n stages of operating unit. After the data is normalized at each node, node 0 reclaims the data in sequence to generate a target matrix. The cluster analysis operation model is employed to compute the nodes whose independence is stronger than column independence. Node 0 follows the matrix type transpose and distributes the matrix by row to multiple compute nodes again (Han et al., 2013). In the end, node 0 collects the data information generated by each compute node to serve as the corresponding gene sequence number of the final result. In each compute node, a uniform result is obtained in terms of the entire process for the overall data processing and sent with the standardization. In addition, the operating efficiency after the completion of data normalization most supports the basic conditions where the nodes retrieve and send. Based on the new post-transpose distribution order of node 0 in the matrix, the sending and receiving process as well as the redistributed genomic data type are taken as the budget model category. Afterwards, the data information of other nodes is identified in the data sent or received by the compute nodes, while ensuring that only one communication result exists between every two nodes and avoiding the negative effects of double counting (Tang et al., 2010). In the meantime, the pairwise relationship of the local data should be calculated. The calculated result is sent to node 0 together with the corresponding gastric cancer gene number. Node 0 combines the results into a final matrix to generate the desired data information result. Figure 2 illustrates the flow structure of platform operating systems.

Figure 2. Process Structure of the Platform Operating System
5. CONCLUSION

At present, a number of reports focus on the application of gene chip technology in detecting the related gene expression of tumour such as gastric cancer. Relevant studies indicate that the operation results of gene chips are more authentic and objective than the statistical analysis results. This study conducts no experimental processing on the operating data. However, only from the perspective of the computational efficiency of cluster analysis, its operation effect is sufficient to support the unified computing at different time nodes and can avoid the adverse effects of repeated operations. For this reason, the hierarchical clustering analysis algorithm can be applied to the bioinformatics expression of the GCGC, thereby improving the actual operation effect of the biological information platform and achieving the expected demand for evaluation information sharing.

REFERENCES