Indian Journal of Science and Technology, *Vol 8(15), IPL043, July 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Parallel Processing Scheme for Minimizing Computational and Communication Cost of Bioinformatics Data

Yoon-Su Jeong*

Division of Information and Communication Convergence Engineering, Mokwon University, Republic of Korea; bukmunro@mokwon.ac.kr

Abstract

With the completion of the Human Genome Project, volumes of genomic data being routinely produced are far exceeding the ability for humans to digest and identify underlying phenomena. Bioinformatics plays a significant role in interpreting large amounts of genomic data. It helps scientists develop new treatments for genetic diseases on the basis of genomic data. Databases are essential for bioinformatics research and applications, so hundreds of databases in bioinformatics and inconsistent terminology and data formats from a variety of data sources must be efficiently managed. This paper proposes a parallel processing scheme that helps the understanding of bioinformatics data in heterogeneous network environments. The proposed scheme creates a hierarchical representation of bioinformatics datasets in Hadoop using the fuzzy relation theory. The internal relations of bioinformatics data are computed via the fuzzy relational product and the external relations are computed via data exchanges among network nodes. The proposed scheme reduces the computational cost for analyzing, correlating and visualizing bioinformatics data by considering only their internal and external relations, irrespective of their types, functionalities, and characteristics. Hadoop employed in the proposed scheme allows distributed storage and parallel processing of huge volumes of data, speeding up processing and communications in general. In addition, the proposed scheme adopts Apache Hive to improve the analysis of distributed bioinformatics data in Hadoop.

Keywords: Bioinformatics, Fuzzy Relation, Hadoop, Heterogeneous Environment, Visualization

1. Introduction

This template, modified in MS Word 2003 and saved as "Word 97-2003 & 6.0/95 – RTF" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers.

Bioinformatics, an umbrella term for biotechnology and information technology, develops computer-based methods and tools that facilitate understanding of the molecular mechanisms of life on Earth, largely by analyzing and correlating genomic information[1–3]. For example, there are methods that link biological data with techniques for information storage, distribution, and analysis to support multiple areas of research. As an interdisciplinary field of science, bioinformatics combines biology, computer science, statistics, mathematics, and engineering to study and process biological data. Bioinformatics serves as the basis for the study of biological evolution and development, genetics, and medicine. Thus, it plays an important role in the growth of biotechnology and pharmaceutical industries[4,5].

The human genome is thought to harbor up to 100,000 genes, and every single gene is made up of thousands, even hundreds of thousands, of chemical bases. In order to understand the genetic basis of hereditary diseases and pinpoint the mutations for such complex diseases, one must process and analyze the astronomical numbers of genetic and genomic data. Bioinformatics technologies like gene prediction tools and sequencing databases are essential in dealing with such huge volumes of data. In recent years, biochip technology that accelerates the analysis of genetic information has also emerged[6–8].

*Author for correspondence

A biochip is a collection of miniaturized test sites arranged on a solid substrate that permits many tests to be performed at the same time. With the planted DNA strands, it can perform thousands of biological reactions, such as decoding genes, in a few seconds. The biochip facilitates screening, diagnosis, and monitoring of diseases.

Many industries and institutions around the world have been working on the development of genomics databases for the preparation of the commercialization of biochips. Some companies have already begun to make a profit by providing gene information services. In Korea, the number of companies that enter this market is growing. The Korean government is carrying out the 21st century strategic industry promotion project that includes the field of bioinformatics. A large-scale development plan for this field has been set up[9–11].

Despite huge volumes of available bioinformatics information, the information that meets the requirements of an individual (society or group) with a specific purpose is scarce. This indicates that identifying meaningful information from large amounts of genetic and genomics data is a challenging task and more effective and sensitive analysis technologies are needed.

While the knowledge gained from the sequencing of the human genome via bioinformatics is expected to change our lives, it entails privacy and security issues. Some of these issues are violations of medical and personal privacy, medical stereotyping of individuals, families, or the entire population, potential discrimination based on medical or genetic data, and a monopoly on medical research and drugs by large companies. Successfully addressing these issues is essential to further progress in the field[12–14].

Bioinformatics data are generally large in volume and often stored in heterogeneous environments. To process and analyze such data requires a large computational power of the server machine—it may take many days of CPU time on large-memory, multiprocessor computers. In addition, there is an increasing need for real-time availability, i.e., data must be available at all times in a global setting. Hence, efficient mechanisms and algorithms for processing bioinformatics data are an important area of bioinformatics research[15,16].

This paper proposes a parallel processing scheme that helps researchers understand and glean insight from bioinformatics data in heterogeneous networks. The proposed scheme uses the fuzzy relation theory to represent bioinformatics data in a hierarchy. The proposed scheme computes the internal relations of bioinformatics data via the fuzzy relational product and the external relations via bioinformatics data exchanges among network nodes. For visualization, the proposed scheme uses the computed internal and external relations irrespective of the types, functionalities, and characteristics of the data, thus decreasing processing and communication overhead. The proposed scheme employs Hadoop to enable distributed storage and parallel processing of huge volumes of bioinformatics data. In addition, Hive is used to enhance the analysis of distributed bioinformatics data in Hadoop.
The rest of this paper is organized as follows:

- Section 2 introduces Hadoop and bioinformatics and describes existing visualization schemes.
- Section 3 presents the proposed visualization scheme for Hadoop-based bioinformatics data.
- Section 4 evaluates the performance of the proposed scheme. Finally, Section 5 concludes the paper.

2. Related Work

2.1 Hadoop

Hadoop, a well-known Big Data technology, is an open source framework for distributed storing and processing of large datasets[1]. The two core components of Hadoop are HDFS, a distributed file system, and MapReduce, a library for parallel processing of large distributed datasets[4,5]. HDFS and MapReduce were originally inspired by technologies created inside Google, i.e., Google File System (GFS) and Google MapReduce, respectively[2,3].

HDFS is a distributed, scalable file system that stores large files across multiple machines. HDFS divides a file into smaller blocks, and stores these blocks in various slave nodes in the Hadoop cluster. A typical block size is 64 MB. To store a file in HDFS, the file is chopped up into 64 MB chunks and the remainder smaller than the 64 MB block size is stored as it is, not occupying the full block space.

As shown in Figure 1, a Hadoop cluster consists of a single namenode (master), a secondary namenode, and multiple datanodes (slaves)[5,6].

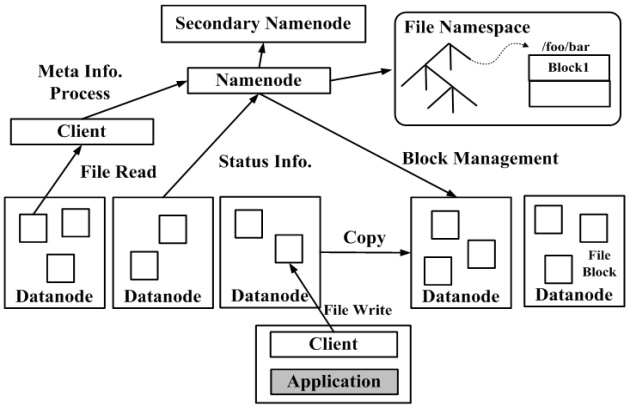

Figure 1 Typical HDFS architecture.

The namenode is the single point for storage and management of all metadata (i.e., the information about file names, directories, permissions, etc.). It is a master server that maintains the file system namespace and regulates access to files by clients. The namenode distributes data blocks to datanodes in the cluster and stores this mapping information. In addition, the namenode keeps track of which blocks need to be replicated and initiates replication whenever necessary.

Datanodes are where blocks of application data are actually stored. The datanode is responsible for serving read/write requests from clients. The datanode performs block creation, deletion, and replication upon instruction from the namenode.

The namenode periodically receives a Heartbeat and a Blockreport from each of the datanodes in the cluster. Receipt of a Heartbeat implies that the datanode is functioning normally. A Blockreport contains the list of all data blocks residing in the datanode, providing an up-to-date view of the block locations to the namenode.

2.2 Bioinformatics

Bioinformatics, the convergence of the biotechnology and information technology, appeared in early 1990s. Bioinformatics seeks to use computers and software for the collection, classification, storage, and analysis of biological information. It is an interdisciplinary field that combines biology, computer science, statistics, mathematics, and engineering. Bioinformatics entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve problems arising from the management and analysis of biological data. Nowadays, bioinformatics is applied to a variety of fields including molecular genetics, genomics, proteomics, medicine, and biotechnology[2,5].

In the post-genomic era that concentrates on harvesting the fruits hidden in the genomic text, bioinformatics is particularly important. It was predicted that humans had about 100,000 genes and each gene is made up of thousands, even hundreds of thousands, of chemical bases. Researchers have to process and analyze an astronomical scale of genome data sets to gain biological insights whose deciphering can be the basis for dramatic scientific and economic success. Bioinformatics technologies such as gene prediction tools and sequencing databases are essential in dealing with such huge volumes of data. In addition, the biochip that facilitates the test and analysis of genetic information has gained a lot of attention lately[4,10,15].

Biochips are miniaturized laboratories that can perform hundreds or thousands of simultaneous biochemical reactions, achieving higher throughput and speed. Physically, a biochip is a square of glass, no larger than a fingernail, covered with millions of strands of DNA. Biochips are likely to have an increasing impact on genetic diagnostics, drug discovery, and basic research applications. With biochip products on the market, it may become possible for ordinary people to diagnose their genetic disorders or diseases.

A large number of companies and institutions all over the world have been striving to develop genomics and bioinformatics databases in connection with the commercialization and maturation of biochips technologies. Some companies have already begun to make a profit by providing chargeable bioinformatics data services. In Korea, the number of companies that enter this market is growing. The Korean government has set up a grand plan to foster the field of bioinformatics, as part of a national effort to gain competitive advantages[7,16].

2.3 Visualization Schemes

Existing visual representation schemes that conceptualize the structure for complex data sets can be categorized into three approaches: node-link (graph-based)[6], matrix (table-based)[7] and hybrid of these two[8,9].

The node-link approach is effective for graphically illustrating the overall network structure, but different nodes and edges might appear to be overlapped in the display. This readability problem deteriorates as the relationships between nodes increases. Various post-processing techniques for refinement, such as sampling, filtering, and clustering, were suggested to address this problem. The complexity of the node-link diagrams in its entirety and high computational costs are drawbacks of this approach.

The matrix approach does not suffer from the readability problem of the node-link approach. Matrix-based visualizations have advantages over node-link diagrams including the ability to create compact graph representations and the ability to remove edge overlapping. A drawback is that the generated matrix is often sparse—the data density is so low that the resulting representations are not particularly insightful and there is a lot of wasted space. In addition, tracking flows and identifying paths are not straightforward as a matrix view is not an easily understandable format for human readers[6,7].

The hybrid approach enhances the performance of the matrix approach by attaching the edge information produced by the node-link approach to the output of the matrix approach. The disadvantages of this approach are the complexity of the edge information attached to the matrix and the difficulty of understanding the relationships among nodes in the matrix.

These three visualization approaches address some of the problems associated with visualizing large datasets and are optimized for efficient sorting and querying of data in networks. However, they have a common problem that the amount of computations increases as the network size increases.

3. Introduction

The proposed visual representation scheme allows the parallel processing of heterogeneous bioinformatics data using traditional bioinformatics tools, or more specifically, sequence database search tools, like BLAST and FASTA.

3.1 Overview

The proposed scheme aims to visualize bioinformatics data in heterogeneous networks while minimizing computational and communication costs. The proposed scheme computes the internal relations of members in a dataset via the fuzzy relational product and their external relations via data exchanges among network nodes, and uses the computed internal and external relations to visualize the dataset. In the proposed scheme, bioinformatics data are stored in Hadoop so distributed storage and parallel processing are possible. In addition, the proposed scheme utilizes Hive, a data warehouse system for Hadoop facilitating querying and managing large datasets residing in distributed storage, to enhance the performance of analysis.

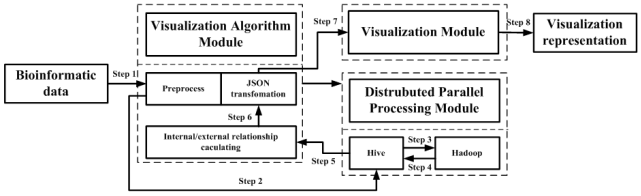

Figure 2 Architecture of the proposed scheme.

Figure 2 shows the component structure of the proposed visualization scheme that takes bioinformatics data as input and produces a hierarchical representation of the input data. The proposed scheme is composed of three main modules: the Algorithm module, Distributed Parallel Processing module, and Visualization module. The Algorithm module takes bioinformatics data and sends the pre-processed input data to the Hive component in the Distributed Parallel Processing module. This data is distributed and stored in Hadoop and then processed in parallel. The processed data is transformed into the JSON format, a lightweight data interchange format. The JSON files are passed to the Visualization module that produces a hierarchical representation of the input bioinformatics data.

3.2 Notations

The terms used in the proposed scheme is that mn denotes the number of bioinformatics data elements that are delivered by a network node to another node. tmn is the total number of bioinformatics data elements that are delivered between network nodes. rmn is the number of already-used bioinformatics data elements that are delivered again by the referring node to other nodes. unrm is the number of nodes that refers to the already-used bioinformatics data elements. Di is the total number

of bioinformatics data elements that are received during the ith day. tdme is the total number of days during which bioinformatics data elements are delivered. tnme is the number of nouns (textual terms) included in the bioinformatics data elements delivered between two nodes.

3.3 Definition of Bioinformatics Data

The proposed scheme uses the fuzzy relational product to define bioinformatics data stored in Hadoop. Bioinformatics data in Hadoop is processed by the namenode and datanodes. A fuzzy implication operator is applied to the computation of fuzzy relational product, i.e., [0,1]×[0,1]→[0,1]. Among the panoply of multi-valued implication operators, the Kleene-Diense operator is used, as shown in Equation (1).

$$a \rightarrow b = (1\text{-}a) \vee b = \max(1\text{-}a, b),\ a=0\sim1,\ b=0\sim1 \quad (1)$$

The definitions of bioinformatics data defined with the fuzzy relational product are as follows.

Definition 1: Let → be a fuzzy implication operator, U be a universal bioinformatics data set, and B be a fuzzy set of U. The membership function, denoted as $\mu_{\cdot B}$, of the fuzzy power set of B is defined as follows.

$$\mu_{\cdot B} A = x \in \cup(\mu_A X \rightarrow \mu_b x) \quad (2)$$

Definition 2: Suppose that there are three finite universal sets of bioinformatics data, U_1, U_2, and U_3. Equation (3) represents the three sets with the fuzzy relational product. Note that the fuzzy relational product indicates the degree of relation of a data member with another. For example, provided that data member a belongs to U_1 and data member c belongs to U_3, the computed fuzzy relational product indicates the degree of relation of a with c.

$$(R \Delta S)_{ik} = \frac{1}{Nj} \sum (R_{ij} \rightarrow S_{jk}) \quad (3)$$

In the equation above, R is a fuzzy relation from U_1 to U_2, and S is a fuzzy relation from U_2 to U_3. i denotes the members (elements) of U_1, k denotes the members of U_3, and j denotes the members of U_2. N_j denotes the number of j.

Definition 3: The fuzzy implication operator is affected by the category (either 'foreset' or 'afterset') of a given bioinformatics data member. Foreset Sc is a fuzzy subset of U_2 consisting of $y \in U_2$ that are related with c, $c \in U_3$. The membership function is $Sc\ (y)=S(y, c)$. Afterset aR is a fuzzy subset of U_2 consisting of $y \in U_2$ that are related with a, $a \in U_1$. The membership function is $\mu_a R\ (y)=\mu_R(a,y)$. Equation (4) represents the subset relation of aR on Sc based on the fuzzy implication of the membership of elements in a set, i.e., $y \in aR$ and $y \in Sc$.

$$\pi_m(aR \subseteq Sc) = \frac{1}{N_{u_2}} \sum_{y \in u_2} (\mu_{aR}(y) \rightarrow \mu_{Sc}(y)) \quad (4)$$

In the equation above, π_m denotes a function that indicates the average degree of relation of a with c based on $R \Delta S$.

3.4 Bioinformatics Data Relations of Internal and External

3.4.1 Internal Relations

To compute the internal relations of bioinformatics data in Hadoop, the proposed scheme uses a matrix in which the frequency of use of the elements (members) in a bioinformatics data set by the network nodes is quantified. This matrix allows measuring the relation of inclusion of the bioinformatics data in different nodes. With the matrix and the Kleene-Diense fuzzy implication operator, the fuzzy relational product is computed to measure the internal relation, as shown in Equation (5). If node n_i has many bioinformatics data members, denoted by x, that has a small membership value, denoted by $\mu_{ni}(x)$, the internal relation score close to 1 can be produced irrespective of the relation of inclusion $n_i \subseteq n_j$. To avoid this, Equation (5) takes into account the relation of inclusion between nodes that use bioinformatics data elements.

$$ir(n_i, n_j) = \mu_{m,\beta}(n_i \subseteq n_j) = (R^T \Delta\ \beta^R)_{ij} = \frac{1}{|n_{i\beta}|} \sum_{K_{ij} \in n_{i,j}} (R^T_{ik} \rightarrow R_{kj}) \quad (5)$$

k_i denotes the ith bioinformatics data member. n_i and n_j denote nodes that use ith and jth bioinformatics data members, respectively. $n_{i\beta}$ denotes n_i's β-constraint, i.e., $\{x \mid \mu_\beta \geq \beta\}$. $|n_\beta|$ is the number of members in $n_{i\beta}$. R is a m×n matrix. R_{ij} is $\mu_{nj}(k_i)$, i.e., the degree of $K_i \in n_j$. R^T is a transposed matrix of R, thus $R_{ij} = R^T_{ji}$.

3.4.2 External Relations

The external relation of bioinformatics data in Hadoop represents how much the bioinformatics data elements stored in a network node are used (referred to) by other nodes. From node *a*'s perspective, the external relation of node a with node b increases as the ratio of the bioinformatics data elements delivered from *a* to *b* increases. In terms of bioinformatics data reference, both *a* that owns a certain data element and *b* that refers to that element after receiving it from a are aware of this act of reference. On the other hand, when this data element is sent by *b* to other nodes, only *b* knows to whom the data is delivered, and *a* is unaware of this act of re-reference. The bioinformatics data elements associated with re-reference are not reflected as significantly as those associated with reference in calculating the external relation. Equation (6) is used to compute the external relation.

$$er(a,b) = \frac{mn \times tnme}{tmn} \sum_{j=1}^{tdme} \left(\frac{di}{tdme(1\,i)}\right) + \frac{rmn}{tmn \times unrm} \quad (6$$

Assume that a node sends false (unrequested) bioinformatics data, which increases the external relation score calculated using Equation (6). The internal relation score of this data is low as its relation of inclusion with other associated nodes is low. The proposed scheme prevents the network node relation from being partially determined by either the internal or external relation that has a predominant score.

3.5 Hierarchy of Bioinformatics Data

The following three steps are performed to hierarchically visualize bioinformatics data in Hadoop.

- Step 1: The internal and external relations of bioinformatics data are normalized using Equation (6).
- Step 2: Equation (7) is executed with the normalized internal and external relations.

$$ies(a \circledR\ b) = norm(ir(a,b)) + norm(er(a,b)) \quad (7)$$

In Equation (7), *ies()* denotes a function that calculates the sum of the internal and external relations. *norm()* denotes a normalization function. *ir(a,b)* denotes the internal relation of node *a* with node *b*, and *er(a,b)*denotes the external relation of *a* with *b*.

- Step 3: Equation (8) is used to represent network nodes into a hierarchy based on their summed relations. Using Equation (8), the nodes that have the sum of the internal and external relations greater than or equal to the average sum are selected, and they are arranged in a hierarchy. The hierarchy levels decrease as $bcut(n_i)$ in Equation (8) increases.

$$bcut(n_i) = \frac{\sum_{j=1}^{l} n_{ij}}{l} \quad (8)$$

Here, n_i denotes i^{th} node and l is the number of nodes. n_{ij} denotes the relation of i^{th} node with j^{th} node. The nodes with a higher relation score (i.e., the sum of internal and external relation scores) are placed in the upper layers of a hierarchy. For the nodes that are connected via multiple relationships, the relation scores of their parent and child nodes are taken into account and the ones with a higher score are placed in the upper layers of the hierarchy. In principle, there is only one connection (relationship) between two nodes but multiple connections might be allowed, if necessary. The proposed scheme can adjust the levels (layers) of a bioinformatics information hierarchy by adjusting the average sum score.

4. Experiments

This section evaluates the proposed visualization scheme with sample bioinformatics data sets. In the experiments, the performance (accuracy, overhead, and delay) of the proposed scheme was compared with that of the previous scheme.

4.1 Environment Settings

Table 2 shows the parameter settings of the experimental environment. The experiment was performed for 8 hours. The number of bioinformatics data elements varied from 1,000 to 30,000. The time interval for data generation was set to 0.01 ms. the similarity threshold settings were {1, 3, 5}.

4.2 Experiment Results

Figure 3 compares the proposed scheme with the previous scheme in terms of the accuracy of the hierarchically represented bioinformatics data. The average accuracy score of the proposed scheme is higher than that of the previous scheme by 11.3%. The proposed scheme processes the bioinformatics data in heterogeneous networks in parallel and correlates the data via the computed internal and external relations, without considering their types, functions, and characteristics. This contributes to improving the accuracy of data visualization.

Figure 4 shows the overhead for hierarchically visualizing the bioinformatics datasets in Hadoop. The communication overhead of the proposed scheme is lower than that of the previous scheme by 5.6%. This indicates that distributed storage and parallel processing of large data sets offered by Hadoop contributes to decreasing the machine overhead.

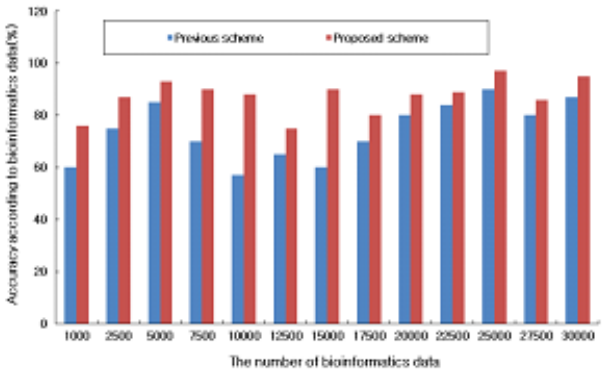

Figure 3. Data accuracy of the proposed vs. previous scheme.

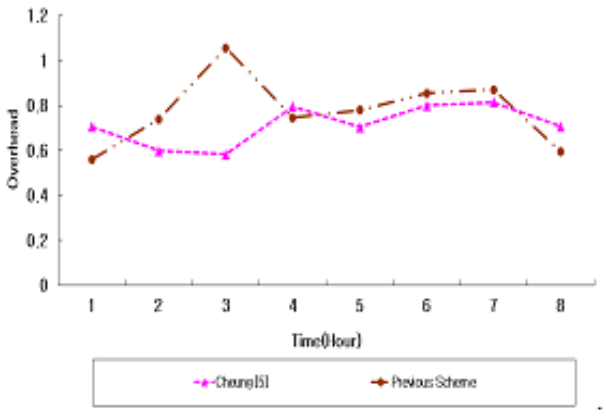

Figure 4. Overhead of the proposed vs. previous scheme.

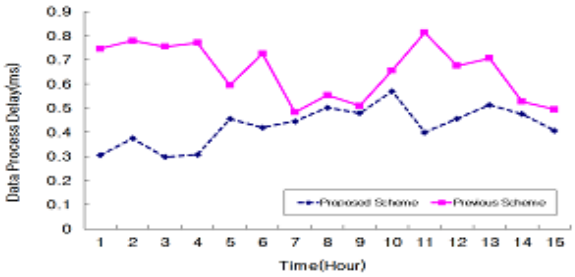

Figure 5. Delays in data processing of the proposed vs. previous scheme.

Figure 5 shows the delays in processing the bioinformatics data over time. The processing delay of the proposed scheme is lower than that of the previous scheme by 8.7%. In the proposed scheme, delays in processing heterogeneous data are reduced as it considers only the internal and external relations of the data, irrespective of their types, functions, and characteristics.

5. Conclusion

This paper proposes a visualization scheme for bioinformatics information in Hadoop to enhance the understanding of the information. The proposed scheme adopts the fuzzy relation theory to correlate bioinformatics data in heterogeneous networks. It computes the internal relations of bioinformatics data elements via the fuzzy relational product and the external relations via bioinformatics data exchanges among network nodes, and creates a hierarchical visual representation based on the computed relations. The proposed scheme does not take into account the properties of heterogeneous bioinformatics data, such as type, functionality, and characteristics, thus lowering computational costs. The experiment results show that the proposed scheme decreases processing delay by 8.7% and communication overhead by 5.6% in comparison with the previous scheme. In the future, the proposed scheme will be applied to a real-world bioinformatics application.

6. References

1. Kher S, Dickerson J, Rawat N. Biological pathway data integration trends, techniques, issues and challenges: A survey. 2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC); 2010. p. 177–82.

2. Xiaohua H. Data mining and its applications in bioinformatics: techniques and methods. 2011 IEEE International Conference on Granular Computing (GrC); 2011. p. 3.
3. Jordan C, Stanzione D, Ware D, Lu J, Noutsos C. Comprehensive data infrastructure for plant bioinformatics. 2010 IEEE International Conference on Cluster Computing Workshops and Posters (CLUSTER WORKSHOPS); 2010. p. 1–5.
4. Dixon S, Yu X. H. Bioinformatics data mining using artificial immune systems and neural networks. 2010 IEEE International Conference on Information and Automation (ICIA); 2010. p. 440–45.
5. Dai W, Cheng JL, Wang QW. A semantic integration system for heterogeneous bioinformatics data. 2012 2nd International Conference on Computer Science and Network Technology (ICCSNT); 2012. p. 1072–6.
6. Abdul-Rahman S, Bakar AA, Mohamed-Hussein ZA. Optimizing big data in bioinformatics with swarm algorithms. 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE); 2013. p. 1091–5.
7. Chong A, Gedeon TD, Koczy LT. Hierarchical fuzzy classifier for bioinformatics data. Proceedings Seventh International Symposium on Signal Processing and Its Applications; 2003. p. 45–8.
8. Ahmed E. Resource capability discovery and description management system for bioinformatics data and service integration - an experiment with gene regulatory networks. 11th International Conference on Computer and Information Technology; 2008. p. 56–61.
9. Jiang PY, Sun XX, Chen EZ, Kun S, Chiu RWK, Lo YMD, Hao S. Methy-Pipe: an integrated bioinformatics data analysis pipeline for whole genome methylome analysis. 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW); 2010 p. 585–90.
10. Staiano A, Ciaramella A, Raiconi G, Tagliaferri R, Amato R, Longo G, Miele G, Donalek C. Data visualization methodologies for data mining systems in bioinformatics. IEEE Proceedings of International Joint Conference on Neural Networks (IJCNN '05); 2005. p. 143–8.
11. Aloisio G, Cafaro M, Epicoco I, Fiore S, Mirto M. A semantic grid-based data access and integration service for bioinformatics. CCGrid 2005, IEEE International Symposium on Cluster Computing and the Grid 2005. 2005. 1:196–203.
12. Wegener D, Rossi S, Buffa F, Delorenzi M, Ruping S. Towards an environment for data mining based analysis processes in bioinformatics & personalized medicine. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW); 2011 p. 570–7.
13. Yang J, Parekh R, Honavar V, Dobbs D. Data-driven theory refinement algorithms for bioinformatics. International Joint Conference on Neural Networks (IJCNN'99); 1999. p. 4064–8.
14. Chen WJ, Brown DG. Optimistic bias in the assessment of high dimensional classifiers with a limited dataset. The 2011 International Joint Conference on Neural Networks (IJCNN); 2011. p. 2698–703.
15. Liu YL, Liu XM, Yang L. Analysis and design of heterogeneous bioinformatics database integration system based on middleware. 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME); 2010. p. 272–5.
16. Chitralegha M, Thangavel K. Protein sequence motif patterns using adaptive Fuzzy C-Means granular computing model. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME); 2013. p. 96–103.