Abstract

Background/Objectives: Bioinformatic information in many diverse areas is being collected, managed, and stored in relation to the genome project. However, studies related to bioinformatics until now has very low efficiency due to improper management of the bioinformatic information. Methods/Statistical Analysis: In this paper, a data processing scheme applying the AHP algorithm is proposed for efficient management of bioinformatic information. The proposed scheme improves information accuracy by assigning property weights by hierarchically classifying the property information (type, property, priority, etc). Findings: Furthermore, the proposed scheme interconnects weighted bioinformatic information according to the weight to minimize the process time between the server and user, thereby improving the information accessibility. Application/Improvements: As a result of performance evaluation, the proposed scheme obtained improved results compared to the conventional schemes in terms of throughput and process time between the LSS (Location Service Server) and user.

Keywords: AHP, Algorithm, Big Data, Bioinformatics, Data Process

1. Introduction

Recently, boundaries between disciplines are being demolished and various technologies are being merged. Bioinformatics is an application area of informatics on which Biotechnology (BT) and Information Technology (IT) are converged\(^1\). To get solutions to biological problems, bioinformatics collect, manage, store, evaluate, and analyze biology data\(^3\).

Bioinformatics generated enormous amount of DNA sequence data because very sophisticated and complex technologies in relation to the Human Genome Project (HGP) and still explosively increasing biological data\(^1-6\).

Because bioinformatics produces large volumes of data at unpredictably high speed, it can be effectively utilized in various industries such as efficient integrated analysis technology, the integration, interconnection, search, and management technologies of bio data and bibliographic data distributed in various forms, the storage and management technologies of bio information that is increasing to petabytes level using storage devices such as SAN and NAS, ultrahigh-speed parallel computing technology for rapid analysis of bulky data, the development of new drugs and medical diagnosis which is a high value-added industry\(^7-9\).

However, the created bioinformatic information is not being utilized efficiently due to the rapid production speed and the lack of management and analysis technologies. Furthermore, the currency of studies related to bioinformatics is poor because the bioinformatic information has not been timely updated\(^10,11\).

Bioinformatics requires a cross-layer approach according to various types of information. Furthermore, cross-layer integration technologies need to be developed to construct a large-scale map of bioinformatic information. In order to actualize this, a new creative method is required\(^12,13\).

In this paper, an efficient data processing scheme applying the AHP algorithm in the bioinformatics environment is proposed. The proposed scheme applies the types,
properties, and priorities of bioinformatic information to the AHP algorithm so that bioinformatic information can be analyzed and processed in a short time. In particular, the proposed scheme improves the accuracy of bioinformatic information desired by users by assigning weights to the attributes, and hierarchically classified the attributes to minimize the process time between the server and user. Furthermore, the weighted bioinformatic information can be interconnected according to the weight, thereby improving information accessibility.

This paper is constructed as follows. Chapter 2 examines the definition of bioinformatics and AHP. Chapter 3 proposes an efficient data processing scheme applying the AHP algorithm for bioinformatic information. Chapter 4 compares the proposed scheme with conventional schemes. Finally, Chapter 5 presents conclusions.

2. Related Works

2.1 Bioinformatics

Bioinformatics refers to technology for integrated management, analysis, and processing of large volumes of distributed bio data and bibliographic information based on information and communication infrastructure and information technology efficient application services and for providing efficient application services to various areas of bio information (new drug development, medical diagnosis, and improvement of agricultural products, etc)\(^{14-16}\).

As shown in Figure 1, bioinformatics refers to information technology for collecting, managing, storing, evaluating, and analyzing biology data to get answers to biological problems. Bioinformatics involves the convergence of various disciplines and technologies such as basic biology and applied biology, medicine, pharmacology, mathematics, statistics, physics, chemistry, and engineering. The applications of bioinformatics cover the entire areas of life science\(^{17}\).

Figure 2 shows the components of bioinformatics. As shown in Figure 2, bioinformatics requires the process of acquiring data from various biological devices. The collected data is rearranged, processed, and interconnected to form databases. In particular, it is very difficult to compose databases by interconnecting data, and no proper solutions have been found until now. Therefore, there are many problems that need to be solved. Bioinformatics is now being applied to all the fields of molecular biology, as well as to genomics, transcriptomics, proteomics, metabolomics, pharmacogenomics, etc\(^{18,19}\).

2.2 AHP

AHP (Analytic Hierarchy Process) is a systematic approach to find the best alternative by hierarchically classifying various attributes of alternatives to problems that require organized assessment of mutually exclusive solutions when there are multiple goals of decision making or multiple evaluation criteria and by identifying the relative importance of each property\(^{12}\). In other words, the AHP method defines the elements to be included in the decision making process and hierarchically classifies the defined elements so that weights can be determined through 1:1 pairwise comparison. The application steps of AHP are hierarchically composed of the purpose of decision making, evaluation criteria for each stage, and alternatives\(^{20}\).

2.3 Previous Studies

To solve the haplotype assembly, a genetic algorithm for clustering fragments was proposed in\(^9\). Chromosomes are defined by binary strings (fragments) with the length of \(m\). When chrom, is 0 (or 1), the \(i\)th fragment is regarded as one of the first (or second) class members.

The particle swarm optimization method is similar to other many genetic algorithms\(^{13}\). The particle swarm optimization method was used for the composition of the MEC model haplotype as with some genetic algorithms.

The heuristic clustering method (K-means) based on the MEC model was published by\(^{22}\). Two pieces are

Figure 1. Definition of bioinformatics.

Figure 2. Components of bioinformatics.
selected at the primitive center in\(^{10}\). Other pieces are clustered by the hamming distance of the specified center and the distances between pieces.

### 3. Hierarchical Data Collection Method using AHP Algorithm

#### 3.1 Overview

Recently, biological data is explosively increasing in relation to the genome project. Bioinformatics is now used not only in biology, but also in all the fields of molecular biology including genomics, transcriptomics, proteomics, metabolomics, and pharmacogenomics.

To efficiently collect and analyze various pieces of bioinformatic information, a data collection scheme applying the hierarchical AHP algorithm is proposed in this paper. As shown in Figure 3, the proposed scheme consists of location service server, sensing device, and user terminal to apply the AHP algorithm to an actual network.

As shown in Figure 3, the proposed scheme largely consists of four steps. The bioinformatic information gathered from sensing devices is analyzed by the server and delivered to users for confirmation.

In particular, the proposed scheme analyzes the relationships of the decision making elements through the AHP algorithm and forms a hierarchical structure before verifying consistency by using the eigenvalues of the evaluation results by pairwise comparison.

The goal of the proposed scheme is to improve the accuracy of bioinformatic information by efficiently gathering and managing bioinformatic information through the AHP algorithm.

#### 3.2 Notations

The notations used in this paper are defined in Table 1.

#### 3.3 Hierarchical Bioinformatics Management Applying AHP Algorithm

##### 3.3.1 Creation of Attributes of Bioinformatics

To create bioinformatic information, the proposed scheme expresses pieces of bioinformatic information of various biology fields as a correlation matrix according to the number of attributes as shown in the following Equation (1):

$$ BI_{attr} = \begin{bmatrix} 0 & \cdots & \alpha_{ik} \\ \vdots & \ddots & \vdots \\ \alpha_{kj} & \cdots & 0 \end{bmatrix} $$

wherein \(\alpha_{ij}\) denotes the correlation between the attributes of bioinformatic information, \(attr_i\) and \(attr_j\); and \(\alpha_{ij}\) and \(\alpha_{ji}\) refer to the conditions where \(i = 0\) or higher and \(j\) is lower than 1 (0\(\leq i, j\leq 1\)). If \(\alpha_{ij}\) is 0, it means that the attributes of bioinformatic information, \(attr_i\) and \(attr_j\), have no correlation.

The information gathered from various areas is classified by AHP and the attributes of bioinformatics are represented by type, property, and priority. The number of attributes can be changed by the purpose and environment of extracting the bioinformatic information.

![Figure 3. Overall process of proposed scheme.](image)

**Table 1. Notation**

| Parameter | Notation |
|-----------|----------|
| BI \(_{attr_i}\) | Relation between attributes \(attr_i\) and \(attr_j\) \(i\)th attribute |
| Aw | Weight for the relative importance of the bioinformatics evaluation element |
| \(L_{max}\) | Largest eigenvalue |
| CI | Consistency index |
| CR | Consistency ratio |
| RI | Relational information ratio between elements |
3.3.2 Weighting of Bioinformatic Information

The proposed scheme expresses the weights of the relative importance of the evaluation elements using eigenvector as shown in the following Equations (2) and (3):

\[ Aw = nw \]  
\[ w' = (w_1, w_2, \ldots, w_n) \]

wherein \( Aw \) denotes the weight of the relative importance of bioinformatic evaluation element; \( A \) denotes the matrix for pairwise comparison, and \( nw \) denotes the evaluation elements of bioinformatic information.

Equations (2) and (3) cannot evaluate the accurate weight through pairwise comparison because the vector \( w \) is not accurately known. Therefore, vector \( w \) is determined by estimating through characteristic root decomposition as shown in the following Equation (4):

\[ A'W' = t_{\text{max}} W' \]  

Where \( t_{\text{max}} \) denotes the largest eigenvalue.

In Equation (4), \( t_{\text{max}} \) is always equal to or larger than \( n \). Therefore, as \( t_{\text{max}} \) gets loser to \( n \), the matrix \( A \) for pairwise comparison in the proposed scheme has no error and is consistent. This can be verified from the following Equations (5) and (6):

\[ CI = (t_{\text{max}} - n) / n - 1 \]
\[ CR = (CI / RI) \times 100 \]

Wherein \( CI \) denotes the consistence index, \( CR \) denotes the consistency rate, and \( RI \) denotes the correlation ratio between elements.

3.3.3 Bioinformatic Information Processing

Before the solution of bioinformatic information is received, the location service server (LSS) generates the solution information \( S_I \) of user \( U_i \) using the attributes of user \( U_i \) as shown in the following equation (7):

\[ S_I = (BI_{attr1}, BI_{attr2}, \ldots, BI_{attrn}) \]

wherein \( n \) denotes the number of solution attributes of user \( U_i \).

When the solution information \( S_I \) of user \( U_i \) is created as shown in equation (7), LSS generates a random key for user \( U_i \) to access the solution information \( S_I \) as shown in the following equation (8):

\[ \text{Generate } (r_1, r_2) \in Z^* \]

LSS delivers the attributes of user \( U_i \) to user \( U_j \) together with the random key \( (r_1, r_2) \) generated in equation (8) and the prior shared public key \( SK_i \).

User \( U_j \) updates the bioinformatic information together the state information. The state of bioinformatic information determines the update or no update of information depending on the value 0 and 1.

4. Evaluation

The proposed scheme is evaluated by security and performance. For performance evaluation, the throughput and delay time are compared with \( 10^{10}, 11 \).

4.1 Environment Setup

For performance evaluation of the proposed scheme, each user \( U_i \) sets the threshold of bioinformatic information to 1. It is assumed that LSS delivers bioinformatic information together with the random key \( (t_1, t_2) \) when user \( U_i \) requests it. Table 2 and Figure 5 show the parameters used in the simulation and the experiment environment.

4.2 Security Analysis

To prevent the replay attack of third parties when LSS delivers bioinformatics information to user \( U_i \), the proposed scheme creates random keys \( t_1 \) and \( t_2 \) so as to ensure security even if the information is tapped by third parties. Furthermore, the proposed scheme is safe from replay attack because it uses the connection information of one-way hash function to apply to the attributes of the solution information \( S_I \) of user \( U_i \) between LSS and user.

The proposed scheme sends the public key \( SK_i \) that has been agreed in advance between LSS and user.

Table 2. Parameter setup

| Parameter                        | Setting                  |
|----------------------------------|--------------------------|
| Number of bioinformatic property | nd={1,000, 5,000, 10,000}|
| Number of Property threshold     | p= {1, 2, 3, 4, 5}       |
| Data generation interval         | 0.01 ms                  |
| Initial self data set time       | 1 hours                  |

Figure 5. Simulation experiment environment.
Therefore, even if the information is leaked by third parties, it cannot be understood. Furthermore, the proposed scheme calculates the weight of AHP-based bioinformatics information to control unauthorized access of third parties between LSS and user. The proposed scheme performs mutual registration, authentication request, key exchange, transmission of device authentication information, and transmission of authentication result according to the weight information. Furthermore, it uses the random keys $t_1$ and $t_2$ as the connection information of the one-way hash function. Therefore, it can prevent attacks for multi-step service access authorization.

4.3 Performance Analysis

4.3.1 Throughput

Figure 6 compares the throughput of server generated during the transmission and reception of information by LSS in the bioinformatics environment. As a result of experiment, the throughput increased as the size of the exchanged bioinformatic information was greater. When the size of the exchanged bioinformatic information was a 64-byte packet, the throughput was small at 0.28 on average. However, when the bioinformatic information was processed in 1024-byte packets, the throughput increased to 0.4 on average. Furthermore, assuming that the server processes the same packet as shown in Figure 4, the throughput of the proposed scheme improved by 9.6% on average compared to the schemes in $^{10,11}$.

4.2.2 Delay Time

Figure 7 shows the delay time for the exchange of bioinformatic information between LSS and user. As a result of the experiment, the scheme in $^{10}$ indicated the highest delay time which was higher by 10.3% compared to the proposed scheme, and was lower by 6.1% compared to the scheme in $^{11}$. These results were obtained because the proposed scheme applied the AHP algorithm to the bioinformatic information and assigned weights.

5. Conclusions

With the recent increase of bioinformatic information at an unpredictably rapid rate, the need for managing bioinformatic information is rising. In this paper, a data collection scheme for efficiently gathering and managing information by applying the AHP algorithm to various pieces of bioinformatic information in the bioinformatics environment and assigning weights to the gathered information was proposed. Furthermore, the proposed scheme improved information accessibility by interconnecting the weighted bioinformatic information to minimize the process time between server and user. The performance evaluation results showed that compared to the conventional schemes, the proposed scheme improved throughput by 9.6% on average and decreased the delay time by 8.2% on average. In future studies, the system performance will be evaluated by applying the findings of this study to actual systems.

6. Acknowledgment

This work was supported by the Security Engineering Research Center, granted by the Ministry of Trade, Industry and Energy.

7. References

1. Disz T, Kubal M, Olson R, Overbeek R, Stevens R. Challenges in large scale distributed computing: Bioinformatics. Proceedings Challenges of Large Applications in Distributed Environments; Research Triangle Park, NC. 2005. p. 57–65.
2. Sumitomo J, Hogan JM, Newell F, Roe P. BioMashups: The new world of exploratory bioinformatics? Proceedings of IEEE 4th International Conference on eScience; Indiana, USA. 2008. p. 422–3.
3. Lengauer T. Algorithmic research problems in molecular bioinformatics. Proceedings of the 2nd Israel Symposium on the Theory and Computing Systems; Natanya, Israel. 1993. p. 177–92.
4. Alterovitz G, Ramoni MF. Bioinformatics and proteomics: An engineering problem solving-based approach. IEEE Transactions on Education. 2007 Feb; 50(1):49–54.
5. Saaty TL. How to make a decision: The analytic hierarchy process. European Journal of Operational Research. 1990 Sep; 48(1): 9–26.
6. Neelakanta P, Chatterjee S, Pappusetty D, Pavlovic M. Information-theoretic algorithms in bioinformatics and bio-/medical-imaging: A review. Proceedings of 2011 ICRTIT; Chennai, India. 2011. p. 183–8.
7. Lau KW, Siepen J. Bioinformatic approaches to improve the identification of peptides from proteomics experiments. Proceedings of the Institution of Engineering and Technology Seminar on Signal Processing for Genomics; Cambridge, UK. 2006. p. 23–45.
8. Jeong YS, Lee BK, Lee SH. An efficient device authentication protocol using bioinformatic. Processing of 2006 International Conference on Computational Intelligence and Security; Guangzhou, China. 2006. p. 855–8.
9. Wang RS, Wu LY, Li ZP, Zhang XS. Haplotype reconstruction from SNP fragments by minimum error correction. Bioinformatics. 2005 Feb; 21(10):2456–62.
10. Wang Y, Feng E, Ruisheng RW. A clustering algorithm based on two distance functions for MEC model. Computational Biology and Chemistry. 2007 Apr; 31(2):148–50.
11. Bustamam A, Burrage K, Hamilton NA. Fast parallel markov clustering in bioinformatics using massively parallel computing on GPU with CUDA and ELLPACK-R sparse format. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2012 May; 9(3):679–92.