International Journal of Hybrid Information Technology
Vol.8, No.6 (2015), pp.135-144
http://dx.doi.org/10.14257/ijhit.2015.8.6.14

A Fuzzy C-mean based Data Mining Algorithm Used in the Bioinformation

Yang Zaihua

College of technology, Xi'an International University, 710077, Xi'an Shaanxi China
yangzaihua@xaiu.edu.cn

Abstract

Data mining technology is a powerful tool to solve the problem. It is widely used to identify potentially useful information. Fuzzy principle based fuzzy C- means is a modification of commonly used C- means clustering technique. In the paper, a modified fuzzy C-means algorithm is used in the gene sequence. The modification is through taking a pseudo F statistics into the method. In the simulation, we use nodes instead of gene to verify the validity. According to the simulation, we get the optimal cluster number, the structure of the classification of the nodes. In order to test the performance of the algorithm, it has been used to process large amount of data, and results show that it has higher processing speed and stable performance. The algorithm can be used in the gene description.

Keywords: *bioinformation, fuzzy C-means, data mining, gene*

1. Introduction

The number of nucleic acid, protein and other biological data accumulated by human is increasing rapidly with hitherto unknown speed. Capacity of the PDB and GenBank and other databases is expanding at geometrically speed. How to handle and analyze vast amounts of biological data is a huge challenge for biological information works.

Bioinformatics [1-3] is generally considered to be part of the computational molecular biology and computer-related part [4-7]. Broadly speaking, bioinformatics is the use of various computers and related techniques to extract biological data, storage, processing and analysis. Perspective on genetic analysis [8], bioinformatics mainly refers to computer processing and analysis of nucleic acid [9-10] and protein sequence data [11-12], three-dimensional structure data of proteins.

The birth and development of bioinformatics make the rapid growth of nucleic acid [13], protein structure [14-15] and function of data [16], various disease-related data [17] and biological literature data. How to make better use of such huge quantities of data is a great challenge for computer works and biologist. Data mining technology [18-19] is a powerful tool to solve the problem.

Data mining is non-trivial process to identify an effective, novel, potentially useful and ultimately understandable patterns from the database. Data mining process usually includes the following steps: data cleaning, data integration, data selection, data transformation, mining, pattern evaluation, and knowledge representation. Data mining is product of database technology, machine learning and statics and other multidisciplinary. Therefore, data mining inherited the theory and method above. The object of data mining includes: file, relational databases, heterogeneous and legacy database systems, data warehouses, multimedia databases, spatial databases, time databases, sequence data, Web, text, etc. Data mining tasks generally include a description and prediction. The obtained models are mainly on association analysis, classification, regression, clustering.

ISSN: 1738-9968 IJHIT

In this paper, based on the idea of the data mining, cluster analysis with a modified fuzzy C-means algorithm is used to describe the gene sequence. The main contribution is the modification of a processing gene method, and the remainder of the paper is shown as the following: Data mining is introduced in the Section 2. Clustering algorithm is summarized in Section 3. New method is proposed in Section 4. The simulation results and analysis is shown in Section 5 and the conclusion is described in Section 6.

2. Data Mining

2.1 Main Methods

(1) The decision tree, rules and method. Decision tree is a kind of simple method to represent knowledge. It gradually classifies cases into different categories, which are like a flow chart with a tree structure. In which each node of the tree structure is corresponding to a non-class attributes, each edge is corresponding to each possible value of this attribute, and each leaf node of the tree structure represents a category.

(2) Classification and clustering analysis method. The classification analysis is an important task in data mining. Its input set is a group of revelation collection and several tags. First, gives a mark for each record, which means that enlightenment is classified by tags, and then checks the calibration record and describes the characteristic of these records. Clustering analysis is a descriptive work. It searches and identifies a limited collection or kinds of cluster so as to describe the data. Briefly, it is to identify a set of clustering rules and divides the data into several classes.

(3) Neural network method. Neural network method is a kind of nonlinear prediction model on the physiological structure of the neural network. It carries out pattern recognition through the study. It is a set of connected input and output unit (neurons), and each connection between units is associated with a weight. In the network learning phase, the network implements the corresponding input samples and its corresponding category according to adjusting the weight. When the neural network training is completed, data is imported to the inlet of neural network which have been trained, and classification results can be directly obtained from the outlet. Now neural network based data mining tools is more and more popular, and it can be divided into the forward neural network and self-organizing neural network, etc.

(4) Genetic algorithm. Genetic algorithm is the optimization technique which is based on the theory of evolution. It combines with the design method such as genetic variation and natural selection.

(5) The method of association rules. The correlation analysis of data mining is to find interesting association, correlation, or causal structure between item sets and the frequent model of item sets from a large amount of data. Correlation could be divided into simple associations, temporal correlation and causal relationship. The purpose of correlation analysis is to find out the event correlation which hidden in the data and difficult to be found and even conflict with human consciousness.

2.2 The Basic Process of Data Mining

The realization of data mining mainly has four stages: determine business objects, data preparation, data mining, and analysis of results.

(1) Identify business objects: define business objects and understand the purpose of data mining is an important step in data mining. The final result is unpredictable, but the explored problem should be predictable, the data mining without the desired effect will reduce the efficiency.

(2) Data preparation phase can be divided into three sub-phases: data integration, data selection, data pre-processing.

(3) Process mining stages as follows: hypothesis generation → selection tools → knowledge discovery→confirm the discovery of knowledge.

(4) Result analysis: according to the purpose of policy makers to analyze information, present the most valuable information to the user through visualization with an understandable and observable way.

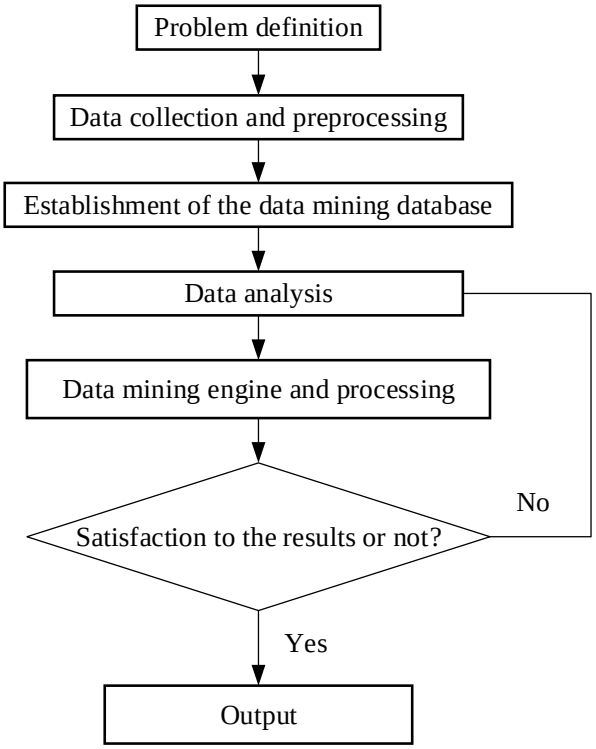

Figure 1. Flowchart of Data Mining

3. Clustering Algorithm

In the gene cluster analysis, there is C- means clustering, hierarchical clustering, self-organizing map (SOM) clustering and principal component analysis (PCA) and so on. All of them are common clustering algorithms [19-22].

Fuzzy C- means [23-24], which is based on fuzzy principle, is a modification of commonly used C- means clustering technique and is a kind of clustering algorithm based on fuzzy partition. It was proposed in 1974 and then popularized. The main idea is that N vectors are divided into C fuzzy clusters in the FCM algorithm. Then, cluster center of each cluster will be calculated to minimize the objective function.

Compared to the C- means algorithm, fuzzy weight index [25-26] is introduced in the objective function. It avoids the defect of membership degree with just two values of 0 and 1. This modification can reflect the actual relationship between data points and cluster center. In the process of cluster, gene sequence alignment is usually used to establish the similarity matrix between gene sequences as the cluster analysis data. But in cases of numerous gene sequence data, this method will greatly increase the computational complexity of establishing process of the similar matrix.

4. Modified Clustering Algorithm

Fuzzy C- means clustering method with the basis of pseudo F statistic and fuzzy theory has been introduced. The algorithm is used in the clustering analysis of gene sequences.

Optimal cluster number will be given according to the pseudo F statistic to achieve the best clustering result.

4.1 Discriminant Function

Pseudo F Statistics are a statistic from the ANO2VA, and it is widely used in F test. Here, we define a pseudo F statistics ratio for the sample with P dimension. It can be described as follows:

$$PFS = \frac{t_r(S_B^P)(n-c)}{t_r(S_w^P)(c-1)} \tag{1}$$

Where, $t_r(S_B^P)$ is the trace of S_B^P; S_B^P and S_w^P are inside and between classed scatter matrix of samples with P dimensional variables; c is the class number; n is the number of genes.

$$S_B^P = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} V_i V_i^T \tag{2}$$

$$S_w^P = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} (N_k - V_i)(X_k - V_i)^T \tag{3}$$

Where, $V_i, (i = 1,2,...,c)$ is the center of class i; $N_k, (k = 1,2,...,n)$ is the value of sample i, and $V_i, (i = 1,2,...,c)$ $N_k, (k = 1,2,...,n)$ are both vectors with P dimensions.

u_{ik} Can be defined as follows:

$$u_{ik} = \begin{cases} 1 & N_k \in Gi \\ \\ 0 & N_k \notin Gi \end{cases} \tag{4}$$

With the definition of PFS and formulas described above, characteristics of PFS can be described as follows:

(1) If c is fixed, this means the classification number is fixed. If obtaining the maximum PFS value, $tr(SPW)$ should be a minimum. For the k-means clustering method, if making the inside class scatter matrix trace $tr(SPW)$ be minimum, corresponding classification result will be satisfactory;

(2) Considering the condition of PFS value varies with the C value. With the increase of C value, $tr(SPW)$ is always tending to decline and $tr(SPB)$ is always tending to rise. However, PFS value is not always risen. In fact with the increasing of C, ratio of $(n-C)/(C-1)$ is falling. It can be predicted that PFS will reach its maximum value at a certain value of C. This is the optimal searched classification number.

4.2 Clustering Algorithm

The objective function of FCM algorithm can be defined as follows:

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^m \left\| n_k - V_i \right\|^2 \tag{5}$$

Where, $U = \{u_{ik}\}_{c\times n}$ is a membership matrix; $V = \{v_1, v_2, ..., v_c\}_{c\times n}$ is the center coordinates of C clusters. $\|n_k - v_i\|$ is the Euclidean distance between sample data point N_k and the cluster center V_i; m is a fuzzy weight index, and it is used to strengthen the degree of contrast of membership of N_k to each class.

FCM clustering problem is to get optimal value of U and V to make the objective function $J_m(U,V)$ be a minimum value. With the optimization of the objective function, necessary conditions of minimum value of the function have been given:

$$u_{ik} = \frac{D_{ik}^{\frac{1}{1-m}}}{\sum_{j=1}^{c}\left(D_{jk}^{\frac{1}{1-m}}\right)} \tag{6}$$

Where, $D_{ik} = \|n_k - v_i\|^2$ is the distance between sample data point n_k and the cluster center V_i

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m n_k}{\sum_{k=1}^{n}(u_k)^m}, \quad i = 1,2,..,C \tag{7}$$

Construct the Similarity coefficient r_{ij}:

$$r_{ij} = 1 - 0.01875\frac{\sum_{k=0}^{n_{ij}} dis(n_{ik}, n_{jk})}{N_{ij}} \tag{8}$$

Where, $dis(n_{ik}, n_{jk})$ is the Euclidean distance.

In the basis of original fuzzy clustering algorithm, pseudo F statistic has been taken into the algorithm. The modified fuzzy clustering algorithm can be briefly described as follows:

Input: the fuzzy weight index m, threshold ε and a maximum iteration number t; output: the best cluster number and clustering results and PFS value.

(1) Initializing the cluster center V, and set $k = 0$, cluster number $C = 4$;

(2) Initializing fuzzy clustering matrix $U^{(0)}$;

(3) When it is in the iteration of k, center of matrix $V^{(k)}$ should be calculated by the formula described above;

(4) Calculate the objective function $J_m^{(k)}$. If

$$-\varepsilon < J_m^{(k)} - J_m^{(k-1)} < \varepsilon \text{ Or } k \geq t \tag{8}$$

The iteration will be terminated, and output the corresponding class center, membership matrix and the value of the objective function. Then go to step (6), otherwise go to step (5)

(5) Set $k = k + 1$, and use formula to update the membership matrix, and go to step (3);

(6) According to the corresponding class center and clustering results, scatter matrix between and within classes and will be calculated. Calculate the value of PFS when current number of clusters is C. If $PFS(C) > PFS(C-1)$ and $C \leq X$, then $C = C + 1$. Where, X is the total number of samples. Then, go to step (3), otherwise output the results.

5. Analysis of Simulation Results

Table 1 shows the detail information of the data set. The data set includes 20 nodes, and each node is composed of 4 sequences.

Table 1. The Characteristic Matrix of the Data Set

	Characteristic 1	Characteristic 2	Characteristic 3	Characteristic 4
Node 1	23.152	36.089	29.356	29.848
Node 2	22.214	36.376	30.701	33.425
Node 3	23.418	38.615	28.912	29.366
Node 4	22.899	37.072	30.747	30.379
Node 5	22.119	36.204	28.456	31.025
Node 6	22.338	36.662	25.361	30.891
Node 7	21.912	36.708	29.321	31.316
Node 8	21.742	36.998	26.581	30.482
Node 9	21.788	36.093	29.273	29.723
Node 10	23.276	36.286	25.810	29.332
Node 11	22.234	36.903	30.954	33.249
Node 12	22.431	36.337	29.371	32.988
Node 13	21.858	37.012	28.782	31.721
Node 14	23.438	37.209	26.819	29.817
Node 15	22.367	35.902	27.927	30.489
Node 16	22.470	36.491	26.358	33.670
Node 17	21.872	36.879	32.206	33.234
Node 18	22.899	37.002	31.528	33.569
Node 19	21.434	36.785	30.329	31.988
Node 20	22.633	36.912	26.612	33.515

Table 2. The Optimal Clustering Results

cluster	Node
Ⅰ	node 5; node 6, node 7, node 8, node 9;
Ⅱ	node 11, node 12, node 13;
Ⅲ	node 16; node 17, node 18,
Ⅳ	node 19, node 20;
Ⅴ	node 1, node 2, node 3, node 4;
Ⅵ	node 10, node 14, node 15.

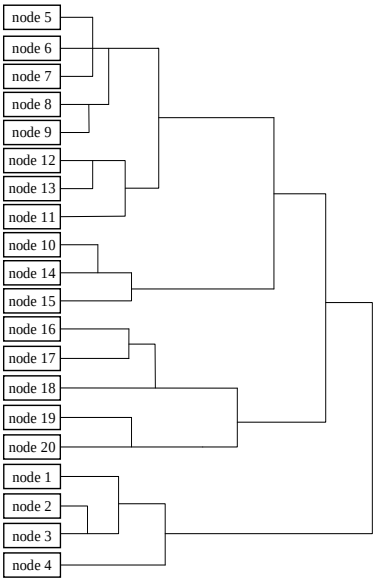

Figure 3. Results of Hierarchical Clustering

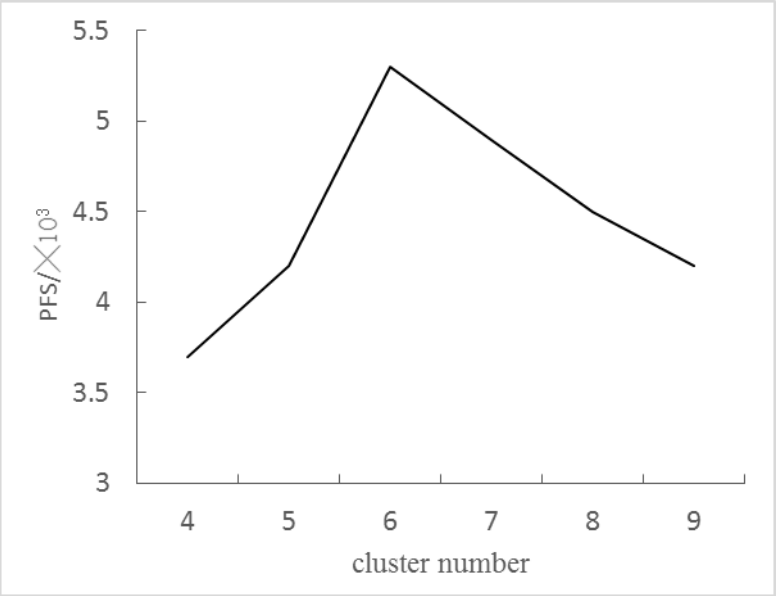

Figure 4. PFS Curve with the

Figure 4 shows the PFS changing with number of clusters. From the Figure 4 we can get the result of maximum PFS value appears at 6 clusters. This means that the optimal clustering results appear when clustering number is 6.

As shown in Figure 2, it is the optimal structure of the optimal clustering results. Groups with short evolutionary distance are classified as a cluster. And it can be used to prove the effectivity of the algorithm proposed in the paper.

Besides, we test the calculation time when the algorithm is used to process large amount of data. The algorithm is compared with the SAM and MSTClust methods.

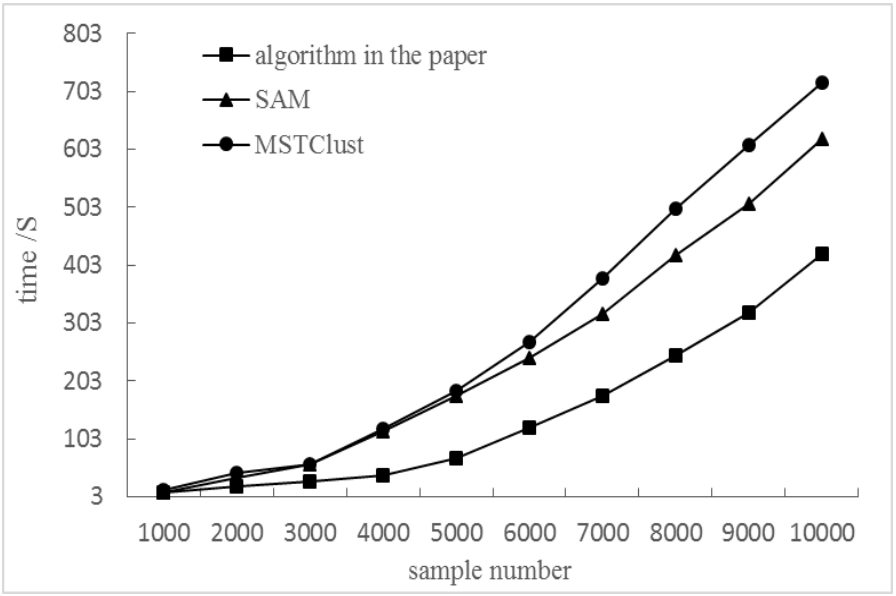

Figure 5. Comparison of Execution Time

Figure 5 shows the comparison of the execution time of the algorithm proposed in the paper, SAM and MSTClust algorithms. We can see that the algorithm proposed in the

paper has a faster calculation speed. When it is used in the calculation with large amount of data, it can improve efficiency and save time.

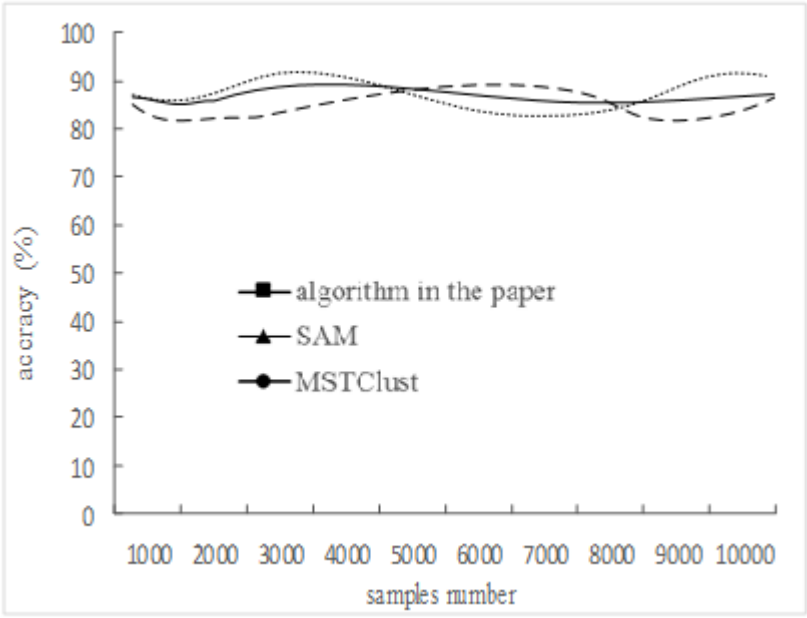

Figure 6. Comparison of Clustering Stability

Figure 6 shows the stability of the algorithm of the algorithm proposed in the paper, SAM and MSTClust algorithms. We can see that the algorithm proposed in the paper has less fluctuation, which means the more stable in the calculation. When it is used in the calculation with small amount of data to large amount of data, it can keep the stability and provide the accuracy.

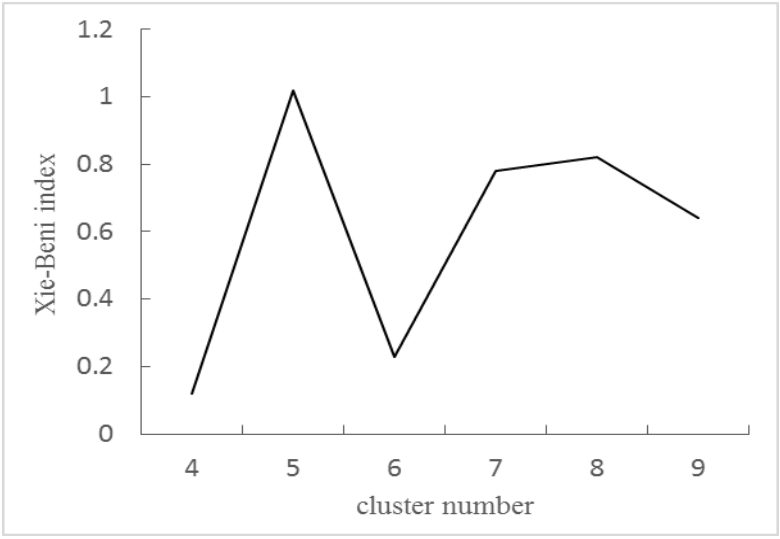

Figure 6. Comparison of Clustering Stability

Figure 6 shows the variation of Xie-Beni indicator with the variation of cluster number. It can be seen that when the cluster number is 6, the indicator is quite small. This means the distance between classes is long. It can be seen the effectively of the algorithm.

6. Conclusions

The number of nucleic acid, protein and other biological data is accumulated by human increases at a faster speed. How to classify the vast amounts of biological data is a huge challenge for biological information works. Data mining technology is a powerful tool to solve the problem. It is widely used in the scientific research and other fields to identify potentially useful information.

Fuzzy principle based fuzzy C- means is a modification of commonly used C- means clustering technique. The main idea is dividing fuzzy clusters and calculating the cluster center to minimize the objective function.

In the paper, a modified fuzzy C-means algorithm is used in the gene sequence. The modification is through taking a seudo F statistics. This method is used to classify the cluster for gene sequence. In the paper, we use nodes instead of gene to verify the effectivity. According to the simulation, we get the optimal cluster number, the structure of the classification of the nodes. Then, we test the algorithm in dealing with large amount of data and the results shows that it has higher processing speed and stable performance. The results can be used to prove the validity and the algorithm can be used in the gene description.

References

[1] L. Chen, S. Weixing, S. Sheng and A. Zhilong, "Gene expression patterns combined with bioinformatics analysis identify genes associated with cholangiocarcinoma", Computational Biology and Chemistry, vol. 47, **(2013)**, pp. 192-197.
[2] T. Andrew, C. Rory, B. Peter, Y. Li, T. Douglas and E. Scott, "5th Workshop on Workflows in Support of Large-Scale Science", WORKS, **(2010)**.
[3] G. Kreshna, J. C. Sacchettini and T. R. Joerger, "Database approaches and data representation in structural bioinformatics", Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE, **(2007)**, pp. 425-432.
[4] P. Manjunath, M. Jeganathan, S. Thangarasu, S. Janakiraman and B. Elumalai, "Computational systems biology of oxLDL induced macrophage foam cell formation-A multilayer regulatory network analysis of atherogenic process", Journal of Chemical and Pharmaceutical Research, vol. 5, no. 10, **(2013)**, pp. 39-44.
[5] E. Alexander, "Some experiences with solving semidefinite programming relaxations of binary quadratic optimization models in computational biology", Asia-Pacific Journal of Operational Research, no. 2, **(2014)**.
[6] S. Ernesto, D. Natalia, M. Jefferson and S. Dimas, "CENCALC: A computational tool for conformational entropy calculations from molecular simulations", Journal of Computational Chemistry, vol. 34, no. 23, **(2013)**, pp. 2041-2054.
[7] A. Anna, C. Simon, C. Giosuè, D. Simona, P. Lucia, A. Stefano, O. Francesco and C. Gabriele, "Molecular interaction fields in drug discovery: Recent advances and future perspectives", Wiley Interdisciplinary Reviews: Computational Molecular Science, vol. 3, **(2013)**, pp. 594-613.
[8] J. I. Lucas-Lledó, V.-S. David, C. Aguado and M. Cáceres, "Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm", BMC Bioinformatics, vol. 15, no. 1, **(2014)**.
[9] R. P. Singh, O. Byung-Keun and C. Jeong-Woo, "Application of peptide nucleic acid towards development of nanobiosensor arrays", Bioelectrochemistry, vol. 79, no. 2, **(2010)**, pp. 153-161.
[10] W. Ronghu and T. B. McMahon, "Investigation of proton transport tautomerism in clusters of protonated nucleic acid bases (cytosine, uracil, thymine, and adenine) and ammonia by high-pressure mass spectrometry and ab initio calculations", Journal of the American Chemical Society, vol. 129, no. 3, **(2007)**, pp. 569-580.
[11] B. C. H. Chang and S. K. Halgamuge, "Approximate symbolic pattern matching for protein sequence data", International Journal of Approximate Reasoning, vol. 32, no. 2-3, **(2013)**, pp. 171-186.
[12] Z. Qingda, J. Qingshan, L. Sheng, X. Xiaobiao and L. Lida, "An efficient algorithm for protein sequence pattern mining", ICCSE 2010 - 5th International Conference on Computer Science and Education, Final Program and Book of Abstracts, **(2010)**, pp. 1876-1881.
[13] B. Chalothorn, D. Boonsong, R. Nisanath, S. Chaturong, P. Nattawee, S. Khatcharin and V. Tirayut, "Pyrene-labeled pyrrolidinyl peptide nucleic acid as a hybridization- responsive DNA probe: Comparison between internal and terminal labeling", RSC Advances, vol. 4, no. 17, **(2014)**, pp. 8817-8827.

[14] W. Shigang, W. Fengjuan, J. Shufeng and Z. Hongjun, "Research on particle swarm optimization for protein structure prediction", Advanced Materials Research, **(2013)**, pp. 605-607, 2497-2500.
[15] T. W. De Lima, F. R. Antonio, G. P. H. Ribeiro, D. A. C. Botazzo and D. S. I. Nunes, "Evolutionary approach to protein structure prediction with hydrophobic interactions", Proceedings of GECCO: Genetic and Evolutionary Computation Conference, vol. 425, **(2007)**.
[16] J. Xiaoyu, N. Naoki, S. Martin, K. Simon, G. David and E. D. Kolaczyk, "Combining hierarchical inference in ontologies with heterogeneous data sources improves gene function prediction", Proceedings - IEEE International Conference on Bioinformatics and Biomedicine, BIBM, **(2008)**, pp. 411-416.
[17] J. Qinghua, W. Guohua and W. Yadong, "An approach for prioritizing disease-related microRNAs based on genomic data integration", Proceedings - 2010 3rd International Conference on Biomedical Engineering and Informatics, BMEI, vol. 6, **(2010)**, pp. 2270-2274.
[18] X. Zhang, "Plain discussion of data mining technology research", 2011 IEEE 3rd International Conference on Communication Software and Networks, ICCSN, **(2011)**, pp. 296-298.
[19] J. Hongfen, L. Yijun, Y. Feiyue, X. Haixu, Z. Mingfang and G. Junfeng, "Study of clustering algorithm based on fuzzy C-means and immunological partheno genetic", Journal of Software, vol. 8, no. 1, **(2013)**, pp. 134-141.
[20] C. Jianmei, L. Hu, S. Yuqing, S. Shunlin, X. Jing, X. Conghua and N. Weiwei, "Possibility fuzzy clustering algorithm based on the uncertainty membership", Jisuanji Yanjiu yu Fazhan/Computer Research and Development, vol. 45, no. 9, **(2008)**, pp. 1486-1492.
[21] W. Guowei, Y. Lin and Y. Kai. Proceedings of IEEE ICIA 2006 - 2006 IEEE International Conference on Information Acquisition, **(2006)**, pp. 1443-1447.
[22] J. Jin-Tsong, C. Chen-Chia and C. W. Tao, "Interval competitive agglomeration clustering algorithm", Expert Systems with Applications, vol. 37, no. 9, **(2010)**, pp. 6567-6578.
[23] C. Chen-Chia, J. Jin-Tsong and L. Chih-Wen, "Fuzzy C-means clustering algorithm with unknown number of clusters for symbolic interval data", Proceedings of the SICE Annual Conference, **(2008)**, pp. 358-363.
[24] M. Wenping, J. Licheng, G. Maoguo and L. Congling, "Image change detection based on an improved rough fuzzy c-means clustering algorithm", International Journal of Machine Learning and Cybernetics, vol. 5, no. 3, **(2014)**, pp. 369-377.
[25] W. Feiyue, X. Zhisheng and D. Longjun, "Comprehensive evaluation model-based fuzzy math of tailings dam stability", Applied Mechanics and Materials, vol. 44-47, **(2011)**, pp. 3408-3412.
[26] B. Hua, "A fuzzy AHP based evaluation method for vendor-selection", Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology, ICMIT, **(2008)**, pp. 1077-1081.

