Study on the Application of Data Mining in Bioinformatics

Mingyang Yuan

School of Science and Liberal Arts, New Jersey Institute of Technology, NJ Newark 07103, US

my244@njit.edu

Keywords: data mining, machine learning, bioinformatics

Abstract. In post-genome era, dealing with biology data from all of aspects has become a main work of bioinformatics. This study mainly analyzes the application of data mining in bioinformatics. Based on the summarize of the definition and main research contents of bioinformatics, the processing flow of bioinformatics data is illustrated. Then emphatically introduces data mining from the perspective of data preprocessing, dimension reduction and statistical machine learning in bioinformatics.

1. Introduction

With the successful implementation of "the human genome project", the amount of life science data has increased dramatically. While the traditional biological experiment method has difficult to meet the needs of dealing with the huge amounts of biological data[1]. In this circumstances, arises the bioinformatics which can solve the problem of molecular biology through calculation methods. The ultimate goal of bioinformatics is to be able to find the biological patterns and information hidden in the sea of biological data, and use the information to improve the understanding of important biological mechanism. Duo to the explosive growth of the available biological data, data mining and machine learning methods have become increasingly important for bioinformatics. The evolution of biological systems is determined by complex underlying mechanism, data mining approaches thus provide ideal tools to the exploration and analysis of them.

2. The Definition and Main Research Contents of Bioinformatics

2.1 The Definition of Bioinformatics

The name of bioinformatics first appeared in the late 1980 s, and has been widely developed as the development of human genome project and the high-throughput sequencing technology. Bioinformatics uses computer as a tool to storage, retrieval, and analysis the biological information, is one of the major cutting edge of bioscience and natural science[2]. The research of bioinformatics mainly focuses on genomics and Proteomics. The bioinformatics uses the analysis of genomic DNA sequence information as a source, simulates and forecasts the protein spatial structure after getting the information in protein coding regions, then make the drug design according to the function of specific protein.

2.2 The Main Research Contents of Bioinformatics

Based on the summarize of the research content of bioinformatics for more than ten years, the main research contents can be divides into the following parts[3, 4].

1) Sequence alignment. Sequence alignment is to compare the similarity or differences of two or more than two symbol sequence. Its theory basement is the theory of evolution. 2) Comparison and prediction of protein structure. Protein is a long chain of amino acids; the inner sequence of amino acid determines the protein's three-dimensional structure. 3) Gene identification and non-coding region analysis. Gene identification using biology experiments and computer to recognize the fragments which have biological characteristics. Non-coding region consists of introns, right now there is no general guidelines for the analysis of the non-coding region. 4) Molecular evolution and comparative genomics. In post genome era, there are more species which have complete genome data, provides the basis to describe the evolution phenomena and mechanism from the molecular level. Molecular evolution and comparative genomics study biological evolution through the analysis of the

similarities and differences of the same gene sequences from different species. 5) The evolution and logic of the genetic code. With the completion of a variety of biological genome sequencing, provides a new material for the study of the evolution and logic of the genetic code. 6) Drug design based on structure. Recently, a considerable amount of protein and nucleic acid, three-dimensional structures of sugars have been accurately determined which provides the basis for the drug molecular design based on the structure of the receptor.

3. The Processing Flow of Bioinformatics Data

In the post-genomic era, the data mining and processing of huge amounts of biological data has become an important task of bioinformatics[5]. Through preprocessing, data mining and analyzing the massive biological data, and establish mathematical model of quantitative or qualitative, researchers can further discuss the related properties among experiment data, biological processes and major diseases. Therefore, benefit mankind in the perspective of biological medicine. The integration of different types of biological data like genome sequencing text, gene expression level, yuan distribution and cell phenotype, and systematical mining of related mechanism and model of the disease is of great significance for prediction and treatment of disease.

As shown in figure1, the raw data of biology experiment turns into normalized data after being collected and standardized, then the corresponding mathematical model can be established to explore the potential mechanism linking, therefore the reliable support can be provided to the prediction and prognosis of some diseases. This is the normal process flow to solve different kinds of disease problems in bioinformatics. However, for certain diseases, still needs special calculation method in order to obtain more accurate analysis results.

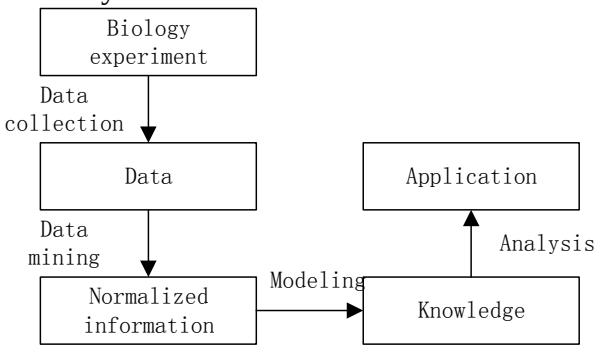

Fig. 1 Flowchart of bioinformatics data processing

4. Data Mining in Bioinformatics

4.1 The Definition of Data Mining

Data mining refers to the process that through the integrated use of a variety of algorithms, make a large amount of data from multiple sources for computer processing, in order to find the natural law behind data[6]. Data mining can be explained from the perspective of statistics, database and machine Learning. In the perspective of statistics, data mining can be explained as the exploration of useful and understandable statistical rules which can be found during the observation and analysis of reliable data. From the perspective of the database, Data mining is the discovery of useful knowledge from a huge amounts of data stored in a database or other information warehouse. From the perspective of machine learning, data mining is the extraction of implicit, unknown and potentially useful information from the massive data.

4.2 Biological Data Preprocessing

The main steps of biological data preprocessing including sample standardization, meta-analysis between samples and imbalanced data processing.

Biological data standardization mainly for the processing of different properties within the same sample, normalized each attributes for further analysis[7]. One of common methods for standardization is min-max normalization which is a linear transformation of the original data, the

result value distribution between [0, 1]. Another common used method is the z-score normalization, let each attribute of biological data minus its average then divided by its standard deviation. The data standardization is very necessary to process data within different classes and make various attributes comparable.

Meta–analysis mainly for standardization between different samples, make data from different experimental conditions comparable[8]. Meta-analysis is of great significance for the integration of data from different platform. Simple meta-analysis includes offset the mean or median by subtracting, for the biological gene expression data, can make the housekeeping genes which has relatively stable expression quantity as the reference value.

Imbalanced data processing applied to the circumstances in pattern classification field which has supervised learning biological data when the difference of sample size is too big. If handled improperly, can lead to a rather large sample category classification results. There are variety of processing methods, such as simulate to increase the number of small sample, delete the number of large sample category, etc. When calculating the classification effect, instead of a relatively simple absolute rate of prediction, it can be replaced by the synthesize of specificity and sensitivity.

4.3 Dimension Reduction of Biological Data

For the big amount of biological data which contains a large number of attributes, it is important to effectively filter useful properties[9]. Fully interpret the connected characteristics of diseases with fewer attributes is the main problem bioinformatics needs to solve. Dimension reduction methods can be divided into two categories: feature transformation and feature selection [10].

1) Feature transformation. Feature transformation refers to make mathematical space compression transformation from the original massive data in high dimensional space, make it mapped to a lower dimensional space and be able to reflect the main characteristics of the original data. Such methods include principal component analysis (PCA), principal component regression (PCR), partial least squares regression (PLS) and fisher linear differential analysis (FDA).

The PCA can make the original high dimensional attribute project to a new low dimensional attribute. Firstly, calculate the first principal component to minimize the residual, then calculate other principal components in the orthogonal direction. Each principal component is a linear combination of the original features and the first principal component can reflect the raw data's largest direction projection. The PCR is used to deal with the statistical learning problems which have outputs based on the principal component analysis. As to the PLS, the traditional linear least square method easy to produce the collinearity problem on the issue of high-dimensional data, while the partial least squares is usually used in the establishment of the regulation network or parameter estimation to deal with the small sample problems.

In the regression issues, the effect of partial least squares method is usually better than PCA and PCR. The PLS is also based on principal component analysis, the different between them is that the PLS also considers the sample label information which maximizes the ratio of the distance between and the distance among the class. For the massive biological data, the traditional dimension reduction methods need to be improved to fit data features.

2) Feature selection. Feature selection just choose a certain feature by calculating ‘yes’ or ‘no’. the methods of feature selection mainly can be divided into three sub-policies: the wrapper sub-policy, the filter sub-policy and the embedded sub-policy[11].

Wrapper sub-policy completely make each feature set as a "black box", the classification ability of each feature needs to combine the specific learning algorithm. The filter sub-policy filters the features in advance before the classification, the features’ properties completely decided by the nature of the training sample. The embedded sub-policy needs to combine with the specific problems and specific classifier. Feature selection strategy can make a score to each feature according to an index, then according to the score to sort those features and finally determines the number of selected features. This kind of strategy is usually very easy and simple, most of it only need linear time complexity. But it still has shortages, such as failing to consider interaction relations between the characteristics.

4.4 Statistical Machine Learning Methods

Statistical machine learning is the methods based on the statistical characteristics which is for high-throughput data. According to the application of output data, it can be roughly divided into two types: unsupervised learning and supervised learning.

1) The unsupervised learning. The unsupervised learning only analysis the features of input data, usually used to calculate the association or clustering among features. K-means clustering method is a kind of unsupervised learning methods commonly used to find cluster and the cluster center which is very easy and convenient to realize.

2) The supervised learning. The supervised learning is more commonly used in statistical machine learning. The supervised learning is used to learn the relationship between input and output characteristics using the corresponding mathematical models under the circumstances that the input and output data are both known. According to the output data type, it can be divided into two subclass: the regression (continuous output data) and the classification (discrete output data). For example, if the output data is different types of diseases, as well as control the normal label, the input data is biological experiment data, this can use the classification model in the field of machine learning to solve. The commonly used classification models are support vector machine (SVM) and K nearest neighborhood (KNN).

The SVM classifier is a classification algorithm based on the maximum hyperplane segmentation[12]. Classic support vector machine (SVM) is applied to two types of problems, makes the positive class (such as disease samples) and the negative class (such as normal control samples) of classification hyperplane has the maximum distance. For the minority types of problems requires a series of strategies to solve, such as “one-to-one "or" one-to-many "approach.

The KNN classifier is a nonparametric classification algorithm, uses the biological properties as the space distance, such as the Euler distance or Markov distance. The position sample label is decided by the first K samples which has the smallest distance. The KNN method belongs to passive learning which is easier to realize, especially suitable for the rapid calculation of high flux of massive data. Moreover, the KNN algorithm can also be improved, such as using kernel function to replace the traditional Euclidean distance.

Whether it is the SVM or the KNN classifier, the attribute selection is both needed in advance. The chosen of significant attributes combination can improve the effect of classification. Moreover, in order to prevent excessive fitting results in the process of learning, cross validation should be used to test the study results.

5. Conclusion

From the perspective of information science technology, the study of bioinformatics is a process from “data” to “discovery”. Data mining technology based on machine learning is playing an increasingly important role in the study of bioinformatics. As dealing with the massive biological data has become the significant work of bioinformatics. Through integrating multi-level data from the biological experiment and effectively application of suitable data mining methods, thus the regulation mechanism of typical disease can be studied in the angle of the whole system. Which is of great significance for life science.

References

[1]. Chou, K.C., *Structural bioinformatics and its impact to biomedical science.* Current Medicinal Chemistry, 2004. **11**(16): p. 2105-2134.

[2]. Fenstermacher, D., *Introduction to bioinformatics.* Journal of the American Society for Information Science and Technology, 2005. **56**(5): p. 440–446.

[3]. Palù, A.D., A. Dovier, and S. Will, *Introduction to the Special Issue on Bioinformatics and Constraints.* Constraints, 2008. **13**(1-2): p. 1-2.

[4]. Jones, D.T., M.J. Sternberg, and J.M. Thornton, *Introduction. Bioinformatics: from molecules to*

systems. Philosophical Transactions of the Royal Society of London, 2006. **361**(1467): p. 389-91.
[5]. Han, J., M. Kamber, and M. Kamber, *Data Mining: Concepts and Techniques, Morgan Kaufmann.* Machine Press, 2001 (in Chinese, 2006. **5**(4): p. 394-395.
[6]. Max Bramer BSc, P., CEng, FBCS, FIEE, FRSA, *Principles of data mining.* Drug Safety An International Journal of Medical Toxicology & Drug Experience, 2007. **30**(7): p. 621-2.
[7]. J, Q., *Microarray data normalization and transformation.* Nature Genetics, 2002. **32 suppl**: p. 496-501.
[8]. Tseng, G.C., D. Ghosh, and E. Feingold, *Comprehensive literature review and statistical considerations for microarray meta-analysis.* Nucleic Acids Research, 2012. **40**(9): p. 3785-99.
[9]. Jain, A.K., R.P.W. Duin, and J. Mao, *Statistical Pattern Recognition: A Review.* IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000. **22**(1): p. 4-37.
[10]. Tang, K.L., et al., *Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data.* Bmc Bioinformatics, 2010. **11**(1): p. 1-8.
[11]. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Applied Physics Letters, 2003. **3**(6): p. 1157-1182.
[12]. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2001. **16**(10): p. 906-14.