The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 45–51

Integrated Data Mining Strategy for Effective Metabolomic Data Analysis

Younghoon Kim Inho Park Doheon Lee

Department of Bio and Brain Engineering Korea Advanced Institute of Science and Technology
Yuseoung gu, Daejeon, Korea

Abstract Disease diagnosis using molecular profiles has gained more attention during the last decades. Among the molecular diagnosis study, metabolomics has been a recently emerging field as promising tools for early detection of diseases. However, due to complexity and largeness of the metabolic profile data, data mining techniques have been essential to handle, process, and analyze the data, and also it is not obvious to apply the data mining techniques to such data, accordingly suggesting the need for suitable data mining strategies for the metabolomic data analysis. In this study, we propose required data mining procedures for effective metabolomics studies including description of current limit or future prospect and our approaches, consisting of preprocessing, dimension reduction, feature analysis and selection, classification, and automated data processing software.

1 Introduction

Disease diagnosis using molecular profiles has gained more attention during the last decades. Among the molecular diagnosis study, metabolomics has been a recently emerging field as promising tools for early detection of diseases. However, due to complexity and largeness of the metabolic profile data, data mining techniques have been essential to handle, process, and analyze the data, and also it is not obvious to apply the data mining techniques to such data, accordingly suggesting the need for suitable data mining strategies for the metabolomic data analysis. Here, we propose the integrated data mining strategies, as shown in Figure 1. The various types of metabolomic data are converted into standard data formats, treated by preprocessing procedures, and then analyzed with appropriate techniques including dimension reduction, classification, clustering, and feature analysis, which will be discussed in remaining parts of this paper, with the introduction about our metabolomic data analysis software, MetaAnalyzer.

2 Data Mining Procedures

2.1 Preprocessing Techniques

Among the whole data mining procedures, in fact, it is well-known that the preprocessing techniques are the most important and difficult part, and there are many issues here that we should deal with.

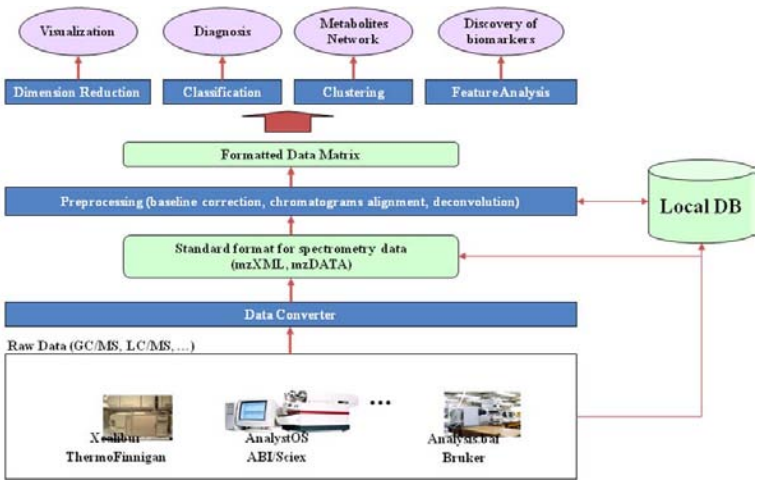

Figure 1: Overview of the data mining strategy for metabolomics data analysis.

2.1.1 Handling and processing a different kind of metabolomic data

There have been many kinds of metabolomic data such as GC-MS (gas chromatography - mass spectrometry), LC-MS (liquid chromatography), CE-MS (capillary electrophoresis), and ESI-LC-MS (electrospray ionization liquid chromatography). In addition, new types of mass spectrometers have been being developed recently including GC*GC-MS, tandem MS, and so on. Therefore, there is need for the processing technique to carefully handle all these kinds of data in consideration to nature of each data. Besides, there are lots of different companies that produce the spectrometers, and the data formats of each spectrometer may be different each other. Accordingly, as in the other fields, the need for a standard format on metabolomic data has increased, and the standards such as mzXML [1] and mzDATA [2] have been developed and widely used. MetaAnalyzer, a software platform for metabolomic data analysis that our research group has developed, supports mzXML as a default data format.

2.1.2 Normalization of data

The noise and background can occur when using electrospray for ionization of samples from chromatography, and thus there should be noise reduction and baseline reduction techniques. To deal with these problems, CODA [3] which stands for component detection algorithm has been developed. Besides, there can be systemic variation between samples. Therefore, we have to adjust it, and can properly treat it with the lowness-based normalization technique [4] which has been used for normalization of microarray data. MetaAnalyzer adopts CODA and lowness normalization as preprocessing methods for these issues.

2.1.3 Identification and quantification of metabolites

After removing noise and background described above, there should be peak alignment techniques for peak shift problems caused by variation of arrival time of compounds from multiple samples. For this issue, COW (correlation optimized warping), DTW (dynamic time warping), and PTW (parametric time warping) have been proposed [5].

As an alternative approach, the algorithm that performs the alignment by clustering retention time of each peak corresponding to each compound has been also proposed. [6, 7] Second, there can be overlapped chromatographic peaks in chromatography results, and for these peaks the algorithm to identify each peak is needed, which is called deconvolution algorithm. There is the popular algorithm named AMDIS [8] that NIST has developed for it. As an alternative way, in our approach, first selecting a particular ion among overlapped peaks is performed, and then it is possible to quantify the metabolite by calculating area of each selected ion chromatogram instead of using the deconvolution. We have developed the selected ion chromatogram method, and MetaAnalyzer uses this method and COW for peak alignment. Figure 2 is the snapshot for the main window of the MetaAnalyzer showing representation of GC/MS data.

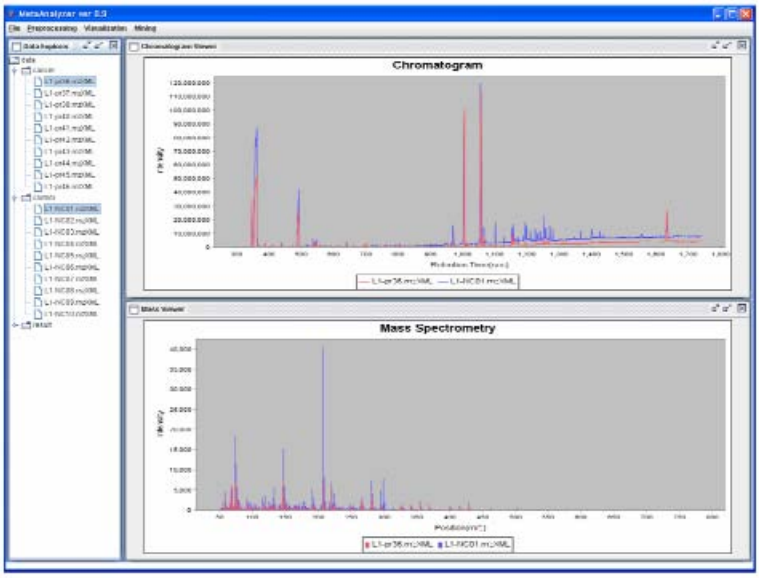

Figure 2: A main window of MetaAnalyzer software: exploring window on the left side and GC/MS data representation on the right side.

2.2 Dimension Reduction Techniques

Once we obtain metabolic profile data after proper preprocessing steps, both one of the easiest and the most powerful analysis tools is to see the data itself with

the naked eye. However, because of the high dimensionality of the data which is a inherited nature, in order to see the data directly, reduction of the dimension of the data into 2 or 3 dimensions is needed. For this purpose, there are two representative methods, PCA (principal component analysis) and PLS-DA (Partial least squares discriminant analysis), which are an unsupervised and supervised method respectively. MetaAnalyzer utilizes a PCA and PLS-DA as dimension reduction and visualization method of data. In Figure 3, we can see the example of the dimension reduction.

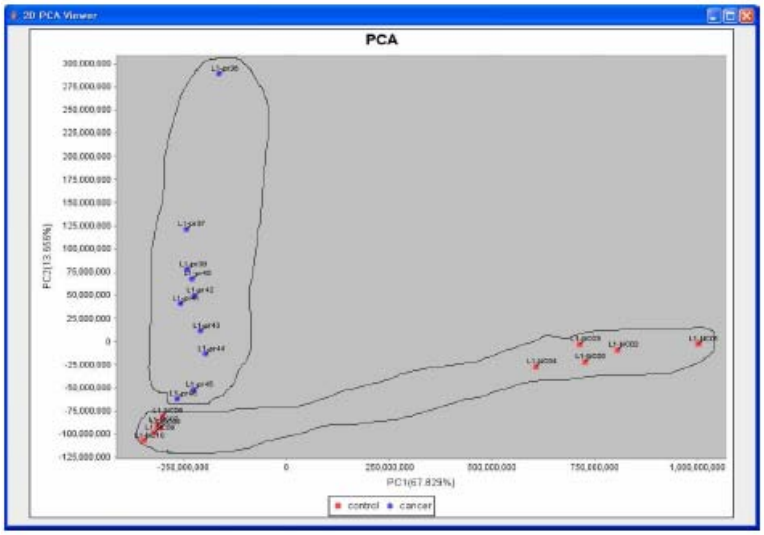

Figure 3: An example of results of Principle Component Analysis (PCA) for the data from patients with breast cancer: breast cancer patients on upper side and normal cases on lower side.

2.3 Feature Analysis and Selection Techniques

The main characteristic of metabolomic data is that there are large amounts of features, which are here metabolites, while the amount of samples is so small. This characteristic affects the results by statistical analysis or classification method applied to this data, resulting in unreliable ones, caused by at least more than thousands of metabolites and insufficient samples. Therefore, there is need for techniques of analysis about features and selection among them. Moreover, to avoid over-fitting to given data and keep general properties of classifiers that we have generated, also it is essential to use feature selection techniques. In addition, because by the feature selection techniques we are able to find a group of the most associated metabolites to the particular researches (e.g. diseases), the findings can be used as bio-markers and can be practically applied. There have been lots of algorithms on feature analysis and selection. Our group also has tried to develop a new method on it based on a genetic algorithm in careful consideration to nature of metabolomic data. MetaAnalyzer pro-

vides visualization tool for feature selection which describes different mass spectrum between two groups of samples, as shown in Figure 4.

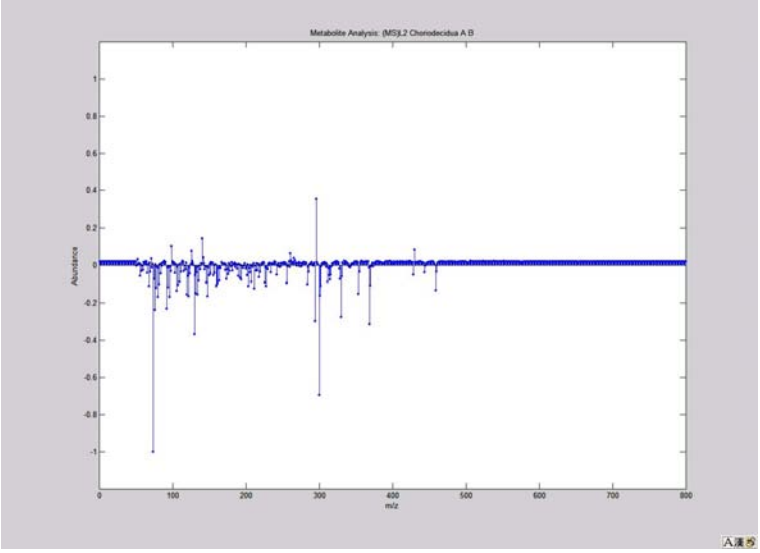

Figure 4: An analysis result of different mass spectrum pattern; the important features, which is m/z, will have high absolute values.

2.4 Classification Techniques

From given metabolic data, we can generate diagnosis models by classification techniques, and then using the generated models, we can diagnose patients by applying the data from them to the models. Moreover, since metabolic data can be easily obtained by simple acquisition of urine or serum from patients, it is possible to achieve simple and easy disease diagnosis techniques with low cost. There are a variety of classification algorithms, and in our consideration, Random Forest, which is regarded as extension of decision tree, can be suitable choice, because this algorithm provides additional information such as significant values on each feature which can be used in feature selection procedures as well as has a strong point in the case with large amount of features. Now MetaAnalyzer supports kNN and Random Forest. In Figure 5, for example, the classification accuracies for sample data by kNN classifier in MetaAnalyzer are shown.

2.5 Automated Data Processing Software

Since the whole procedures described above are not simple, the integrated, automated data processing software are really needed. To be useful tools, there are several requirements. First, the software needs to support as many kinds of data as possible, as explained in preprocessing section. For example, AMDIS software developed by NIST is an excellent tool, but there is limitation in using data of LC-MS or CE-MS.

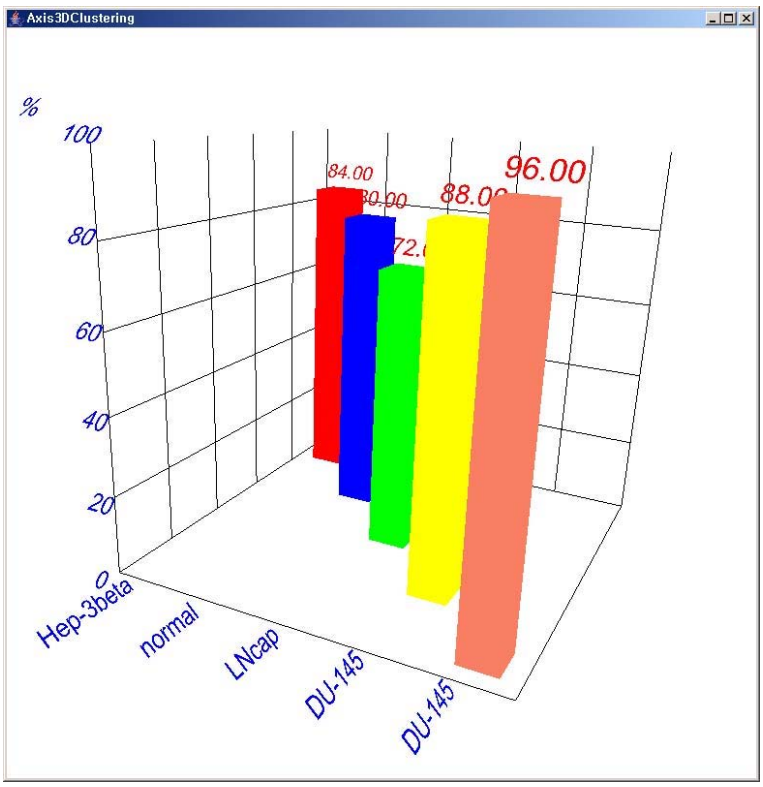

Figure 5: The representation for classification accuracy from the kNN classifier embedded in the MetaAnalyzer.

Second, platform-independent software and modular and flexible software, which allow for addition of new algorithms or functional modules as demands, are needed. For instance, MZmine is platform-independent software for LC-MS data, can extensible to additional new algorithms and functions and to other types of data such as GC-MS and CE-MS. Finally, the visualization techniques are also very important to analyze given data and then decide what kind of analysis procedures we have to take. Therefore, effective visualization techniques are needed including peak visualization parts and visualization on results from dimension reduction. To support the analysis procedures introduced above, MetaAnalyzer, a java-based software platform for metabolomic data analysis comprising all steps from preprocessing to classification for pattern recognition, has been developed by our research group.

3 Conclusion

A lot of metabolomic data have been generated and accumulated these days. Even though there is much information in the data, because we do not have good weapons or suitable analysis procedures to the data, we are not extracting all the

information in it enough. In this study, we have introduced suitable data mining procedures to characteristics of metabolomic data, containing our approach which is implemented to MetaAnalyzer.

Acknowledgement

This research was supported by a grant (2N3060) from the Biodiscovery Research Program funded by the Ministry of Science and Technology of the Korean Government. DL was supported in part by the Korea Ministry of Information and Communication under Grant C109006020001. We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics for providing research facilities.

References

[1] http://sashimi.sourceforge.net/software_glossolalia.html#MassWolf

[2] http://psidev.sourceforge.net/ms/#mzdata

[3] Willem Windig, et. al., A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry, Anal. Chem. Vol. 68, 1996

[4] Y. H. Yang et. al., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, NAR. Vol. 30, 2002

[5] A.M. van Nederkassel, et. al, A comparison of three algorithms for chromatograms alignment, Journal of chromatography A, Vol. 1118, 2006

[6] David P. De Souza et. al., Progressive Peak Clustering In GC-MS Metabolomic Experiments Applied to Leishmania Parasites, Bioinformatics. 2006

[7] Anthony L. Duran et. al., Metabolomics spectral formatting alignment and conversion tools (MSFACTs), Bioinformatics, Vol. 19, 2003

[8] S. E. Stein, An Integrated Method for Spectrum Extraction and Compound Identification from GC/MS Data, Journal of American Society of Mass Spectrometry, Vol. 10, 1999