Genetic Algorithms Successfully Select Potential Biomarkers in Mass Spectrometry-Based Metabolomics

Wei Zou, Vladimir V. Tolstikov
Metabolomics Core, Genome Center, University of California, Davis, CA

Introduction

Research on metabolic profiling samples of the high complexity, biological variance, and large compositional dynamic range poses many challenges for components separation, detection, and data analysis[1]. Complex and large datasets generated with techniques such as hydrophilic interaction chromatography (HILIC–LC–ESI–MS), reversed-phase liquid chromatography (RP–LC–ESI–MS), and gas chromatography (GC–TOF–MS) all coupled to mass spectrometry require modern computational tools and robust data mining technologies[2, 3].

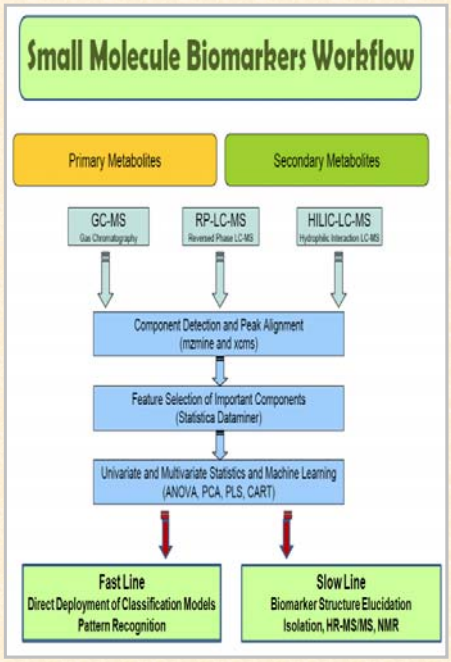
Biomarker discovery requires finding small subset of the most prominent metabolites that could be extended from training set to testing set, validated and further used for discrimination studies and/or diagnostics of the whole population of the particular organism. Feature selection is a technique commonly used in machine learning to select a subset of relevant feature for building robust learning models[4]. Univariate feature selection methods test one metabolite at a time for its ability to discriminate a dependent variable, such as genotype differences in the current case. Then top most significant metabolites are used to develop a statistical model. Multivariate methods take into consideration the synergy among metabolites. Based on different subsets of metabolites, many possible models are evaluated and the most predictive model is identified and selected[5, 6]. It was reported[7] that Markov chain Monte Carlo and Genetic Algorithms (GA) are two promising multivariate approaches in analysis of the LC–ESI–MS metabolomics datasets.

In the present study, three independent and complementary analytical techniques for metabolic profiling of six investigated genotypes of one-year-old *Pinus taeda L.* were applied. Unsupervised methods, such as principle component analysis (PCA) and clustering, and supervised methods, such as classification were used for data mining. GA was probed for selection of the smallest subsets of potentially discriminative classifiers.

References

1. Fiehn, O.; Kopka, J.; Trethewey, R. N.; Willmitzer, L. *Anal Chem* **2000**, *72*, 3573-3580.
2. Shulaev, V. *Brief Bioinform* **2006**, *7*, 128-139.
3. Jain, A. K.; Duin, R. P. W.; Mao, J. *Transactions on Pattern Analysis and Machine Intelligence* **2000**, *22*, 4-37.
4. Saeys, Y.; Inza, I.; Larranaga, P. *Bioinformatics* **2007**, *23*, 2507-2517.
5. Lee, J. W.; Lee, J. B.; Park, M.; Song, S. H. *Computational Statistics & Data Analysis* **2005**, *48*, 869-885.
6. Zhang, X.; Lu, X.; Shi, Q.; Xu, X.-q.; Leung, H.-c.; Harris, L.; Iglehart, J.; Miron, A.; Liu, J.; Wong, W. *BMC Bioinformatics* **2006**, *7*, 197.
7. Goodacre, R. *J Exp Bot* **2005**, *56*, 245-254.
8. Trevino, V.; Falciani, F. *Bioinformatics* **2006**, *22*, 1154-1156.

Methods



Kind T et al. Analytical Biochemistry (2007) 363:185-195.

Xconvert program included in Xcalibur was used to convert the LC-ESI-MS Xcalibur (*.raw) files to netCDF (*.cdf) format. Automatic peak finding, deconvolution, and alignment were performed using XCMS running on the open statistical platform R or MarkerView 1.1 (Applied Biosystems/MDS Sciex, Concord, Ontario, Canada). GC–TOF–MS data were pre-processed and annotated by BinBase[1].

Preliminary data exploration was accomplished using unsupervised methods such as PCA and clustering. PCA analysis used R package pcaMethods in Bioconductor project. Cluster analysis of the PCA scores was performed using partitioning methods such as K-means using the function kmeans() in R package stats, hierarchical agglomerative methods such as Ward's method using the function hclust() in R package stats, and multiscale bootstrap resampling using R package pvclust, and model-based clustering approach using R package mclust which assumes a variety of data models and applying maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters.

Genetic Algorithms (GA) are a class of algorithms based on the principle of biological evolution, suitable for finding approximate solutions to global optimization problems when there is a very large pool of possible solutions. GA procedure incorporates operators such as biological inheritance, mutation, selection, and recombination on chromosomes, initial sets of candidate solutions[8]. Starting from a randomly generated set of chromosomes and a criteria function for evaluating the fitness of an individual chromosome, GA procedure repeatedly selects the fittest candidate solutions in each generation and lets them reproduce and keep the population size constant. This process stops when the goal fitness is achieved. Goal fitness is defined as the average reachable fitness in a reasonable amount of generations. Feature selection using GA procedure and further classification were performed using R package GALGO[8].

All calculations were performed in an R integrated development environment (IDE) RKWard under Kubuntu 7.10. The commercial MarkerView 1.1 was used for validation running on Windows XP.

Results and Discussion

Unsupervised Analysis without Feature Selection: PCA score 3D plot by pcaMethod showed that group 1and 2 formed clusters and were well separated from the other groups, whereas the rest of the groups could not significantly differentiate from each other (Figure 1). pvclust and mclust are good choices for clustering if grouping information is not available.

Feature Selection Using GA: It was found, that 10 most frequent classifiers in HILIC, RP LC–ESI–MS, and GC–TOF–MS data occurred at least 100 times in 2000 models (Table 1). In parallel GA feature selection was applied to peak tables generated with MarkerView 1.1 (Table 2) for LC/MS data. One can find different candidates on the lists suggesting dependence on peak picking algorithms used.

Classification and Prediction after Feature Selection: The respective fittest GA model predicted with a specificity range of 0.85 to 0.98 and a sensitivity range of 0.36 to 1.0 for GC–TOF–MS data (Figure 2), a specificity range of 0.87 to 0.99 and a sensitivity range of 0.66 to 0.94 for HILIC–LC–ESI–MS data, and a specificity range of 0.88 to 1.0 and a sensitivity range of 0.74 to 0.96 for RP–LC–ESI–MS data. In parallel we applied MarkerView 1.1 for PCA analysis with GA selected classifiers (Figure 3).

Metabolic Networking among GA Selected Features: A major difference between GA and other machine learning approaches is its ability to determine relationship among feature components, providing valuable information about metabolite interactions, metabolic pathways, and clinical diagnosis. Therefore, a highly positive or negative correlated group of feature components was preferred for a model. The metabolite network of GC–TOF–MS data illustrated that many saccharides are interwoven heavily, indicating that carbohydrates metabolic pathways involved (Figure 4).

Figure 1. 3D PCA score plots of the GC–TOF–MS data. Groups were color-coded as: 1-black, 2-red, 3-green, 4-blue, 5-cyan, 6-pink.

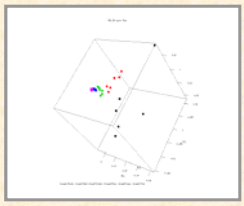


Figure 2. Heatmap plot of GC–TOF–MS data using the fittest GA model.

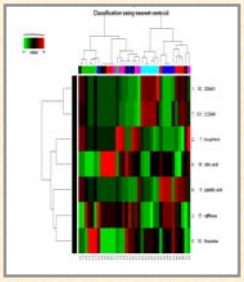


Figure 3. PCA score plots generated with the only GA selected classifiers of the GC–TOF–MS data.

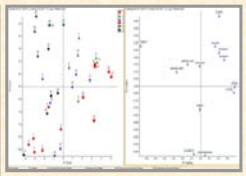


Figure 4. Network interactions among feature components of GC–TOF–MS data. The line thickness represents the dependency strength. Top components were color-coded: black, red, green, blue, cyan, pink, yellow, and gray.

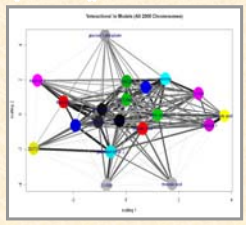


Table 1. Top 10 classifiers (XCMS peak list)

Rank	ions	m/z (Dalton)	RT (sec) frequency
HILIC-LC-ESI-MS			
1	negative	549/220	794
2	positive	1048/372	690
3	positive	405/205	621
4	positive	875/96	478
5	negative	1003/385	440
6	negative	747/487	347
7	negative	743/162	307
8	positive	995/325	307
9	positive	709/457	269
10	negative	832/923	252
RP-LC-ESI-MS			
1	positive	1341/1375	716
2	negative	508/609	457
3	positive	1237/1460	397
4	positive	324/533	370
5	negative	1005/965	358
6	positive	304/549	303
7	negative	741/711	247
8	positive	305/547	196
9	negative	661/965	156
10	negative	153/139	148

Table 2. Top 10 classifiers (MarkerView peak list)

Rank	ions	m/z (Dalton)	RT (min) frequency
HILIC-LC-ESI-MS			
1	negative	827/14.9	151
2	negative	291/9.7	140
3	negative	947/19.5	101
4	positive	828/14.9	95
5	positive	649/12.2	86
6	negative	355/14	82
7	positive	188/9.5	80
8	negative	847/13.5	78
9	negative	289/11.5	76
10	negative	264/5.5	72
RP-LC-ESI-MS			
1	positive	1261/24.5	232
2	positive	257/13.4	225
3	positive	1303/22.7	166
4	negative	744/7.8	156
5	positive	738/6.5	147
6	negative	315/12.9	135
7	positive	313/22.3	99
8	negative	785/11.3	93
9	negative	319/16.1	85
10	negative	530/24.7	71

Conclusion

The present study demonstrated that combination of the comprehensive metabolic profiling utilizing three complementary analytical methods for MS data acquisition and GA technique for feature selection presenting intriguing avenue for finding and exploration small subsets of strong classifiers. Data pre-processing is extremely important for further analysis. Parameters optimization is shown to be essential for avoiding over fitting tendency in multivariate approach. Preliminary results of this study are promising. Developed methods will be further validated and applied for large-scale metabolomic studies involving different organisms. Generation of small subsets of classifiers with high discriminatory ability is particular attractive for diagnostic test developments.