NAIST-IS-DD0661033

Doctoral Dissertation

Integrative analysis of transcriptomics and metabolomics in *Escherichia coli*

Hiroki Takahashi

February 5, 2009

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of SCIENCE

Thesis Committee:

Professor Shigehiko Kanaya	(Supervisor)
Professor Naotake Ogasawara	(Co-supervisor)
Professor Kotaro Minato	(Co-supervisor)

Integrative analysis of transcriptomics and metabolomics in *Escherichia coli* *

Hiroki Takahashi

Abstract

In the era of post-genomics, a systematic and comprehensive understanding of the complex events of the organisms is a great concern in biology. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS) is the best MS technology for obtaining exact mass measurements owing to its great resolution and accuracy, and several outstanding FT-ICR/MS-based metabolomics approaches have been reported. In the present study, I proposed a procedure for metabolite annotation on direct-infusion FT-ICR/MS by taking into consideration the classification of metabolite-derived ions using correlation analyses. Integrated analysis based on information of isotope relations, fragmentation patterns by MS/MS analysis, co-occurring metabolites, and database searches (KNApSAcK and KEGG) can make it possible to annotate ions as metabolites and estimate cellular conditions based on metabolite composition. A total of 220 detected ions were classified into 174 metabolite derivative groups and 72 ions were assigned to candidate metabolites in the present work. Metabolic profiling has been able to distinguish between the growth stages with the aid of PCA. The constructed model using PLS regression for OD_{600} values as a function of metabolic profiles is very useful for identifying to what degree the ions contribute to the growth stages. Ten phospholipids which largely influence the constructed model are highly abundant in the cells. This approach can reveal that global modification of those phospholipids occurs as *E. coli* enters the stationary phase. Thus, the integrated approach involving correlation analyses, metabolic profiling, and database searching is efficient for high-throughput metabolomics. Furthermore, I performed the transition point analysis by applying the statistical method, Linear Dynamical System (LDS) to transcriptomics and metabolomics data, respectively and detected a time lag between transcriptional and metabolite levels. Finally, the integrative analysis of transcriptomics and metabolomics was performed based on gene-to-metabolite correlation analysis by taking into consideration a time lag.

Keywords: metabolomics, transcriptomics, FT-ICR/MS, cDNA microarray, LDS, *Escherichia coli*

.*Doctoral Dissertation, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0661033, February 5, 2009.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Preliminaries

Since the completion and publication of the *Haemophilus influenzae* genome sequence in 1995 [Fleischmann et al. 1995], several high-throughput experimental technologies for post-genomics analyses have been dramatically advanced. Biological research was transformed from a relatively data poor discipline into one that now is data rich [Joyce and Palsson 2006]. Recent technologies allow us to analyze large number of genes or proteins simultaneously, whereas a few years ago each gene or protein was studied as a single entity. In a systems biology approach, a cell is considered as a system which receives dynamically changing environmental cues and transduces these signals into the observed behavior, i.e. change of phenotype or change of physiological response [De Keersmaecker et al. 2006]. The discipline of systems biology is expected to provide a better understanding of cell biology by enabling the study of the function and behavior of molecular interactions in complex networks [Galperin and Ellison 2006]. In this chapter, I review representative high-throughput experimental technologies, i.e. metabolomics and transcriptomics.

1.2 Metabolomics

Information flow in biological systems follows the sequence DNA to RNA to protein. The phenotype of the organism is the product of its genotype within its environment. Amongst products along that flow, metabolites are the end products of cellular regulatory processes, and their levels can be regarded as the ultimate response of biological systems to genetic or environmental changes [Fiehn 2002]. In parallel to the terms ‘transcriptome’ and ‘proteome’, the set of metabolites synthesized by a biological system constitute its ‘metabolome’. An approach by which all the metabolites are identified and quantified, is called ***‘metabolomics’***, in analogy to ‘transcriptomics’ and ‘proteomics’. Metabolomics stands out from any other organic compound analysis in scale and in chemical diversity, i.e. all metabolites are aimed to be described, both primary and secondary metabolites, present in an organism or biological system. Compared to the linear 4-letter codes for genes or the linear 20-letter codes for proteins, metabolites have a much greater variability in the order of atoms and subgroups. The most striking feature of metabolomics lays in its integrative capacity, as part of the omics disciplines, which has resulted in a shift from mainly pure (organic) chemistry-based characterization into a biochemical context.

The methods in metabolomics to analyze the metabolite contents that are extracted from isolated cells or tissues are still being refined, and typically rely on mass spectrometry, NMR spectroscopy and vibrational spectroscopy. Modern techniques must capture hundreds of distinct chemical species, according to the highly diverse set of molecules and the large dynamic range. Despite of these challenges and consequent limitations, metabolomics is quickly becoming a popular tool for studying the cellular states of many systems due to several reasons, e.g. metabolomics can offer insights into metabolism that complement information obtained from transcriptomics and proteomics [Fridman and Pichersky 2005] and shed light on a large set of overlooked metabolic phenotypes, i.e.

silent mutants [Allen et al. 2003].

Mass spectrometry is a spectrometric method that allows the detection of mass-to-charge (*m/z*) species pointing to the molecular mass of the detected metabolites. As a developing technology in metabolomics applications, there are various configurations of mass spectrometers, in terms of ion acceleration and mass detection, ion production interfaces and ion fragmentation capabilities. Most MS applications in metabolomics make use of a separation method before mass detection, typically liquid chromatography (LC), gas chromatography (GC) or capillary electrophoresis (CE). Such separation step introduces an extra dimension for identification (retention time) to the data, and reduces the complexity of the data analysis by avoiding ion suppression at the source. GC-MS is the most popular technology in metabolomics research. Fiehn et al. (2000) used *Arabidopsis thaliana* leaf extracts and automatically quantified 326 distinct compounds, assigning a chemical structure to half of them. The GC-MS approach was also used for studying metabolism in potato tuber tissues derived from either transgenic plants or plants exposed to different environmental conditions [Roessner et al. 2001a; Roessner et al. 2001b]. On the other hand, since metabolites of different empirical formulas have different masses, very high mass resolution (> 100,000) is required to resolve them. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS) is the only MS system capable of routinely achieving this level of resolution with a sufficiently fast data acquisition rate. With FT-ICR/MS, separation of metabolites is achieved solely by ultra-high mass resolution, eliminating the need for time consuming chromatography and derivatization. First application of direct-infusion FT-ICR/MS for metabolomics was performed by Aharoni et al. (2002), in which they analyzed the ripening shift in strawberry fruit during four consecutive steps of development. Recently, several researches were conducted by the use of direct-infusion FT-ICR/MS [Oikawa et al. 2006; Nakamura et al. 2007; Hounsome et al. 2009]. In this dissertation, I developed bioinformatics method on the platform of direct-infusion FT-ICR/MS for metabolomics analysis, and characterized growth stage specific metabolites based on time series samples of *Escherichia coli*.

1.3 Transcriptomics

The field of ***'transcriptomics'*** provides information about the presence and the relative abundance of RNA transcripts, thereby indicating the active components within the cell. Since the mid-late 1990s in advance of proteomics or metabolomics, countless genome-wide studies have examined the dynamics of gene expression in many model systems and environments. Microarrays and serial analysis of gene expression (SAGE) represent the most well-used approaches and have been applied to many model systems. Transcriptomics can be used to identify genes that are potentially involved in particular modules. For example, by sporulating yeast cells and recording the transcriptomics data, transcripts that are upregulated or downregulated could be identified and the corresponding genes postulated to function in the sporulation module [Chu et al. 1998; Priming et al. 2000]. Even though transcriptomics studies provide crucial information regarding the expression state, or primary genomics readout of the cell, it must be recognized that various levels of post-transcriptional control might rival its importance and are not captured [Meta et al. 2005].

Early transcriptomics experiments used small sample numbers and model organisms with relatively small genomes [DeRisi et al. 1997]. Future experiments, however, will deal with hundreds of samples and with organisms that have larger, more complex genomes. A preview of what is to come can be seen in the work reported by Hughes et al (2000), in which 300 samples, half of which were done in duplicate, and 63 negative controls were used to characterize undefined ORFs and potential drug targets in yeast. Bioinformatics is moving towards methods that try to incorporate as much available knowledge as possible.

In this dissertation, I focused on cDNA microarray, in which two mRNA samples to be compared are reverse transcribed into cDNA, labeled using two different fluorophores (usually a red fluorescent dye, Cy5, and a green fluorescent dye, Cy3) and then hybridized simultaneously to the glass slide. Intensity values generated from hybridization to individual DNA spots are indicative of gene expression levels, and comparisons in gene expression levels between the two samples are derived from the resulting intensity ratios.

1.4 Dissertation outline

This chapter introduces ‘omics’ approaches. Section 1.2 and 1.3 describe metabolomics and transcriptomics, respectively. Section 1.4 provides an outline of this dissertation. The following four chapters of this dissertation address ‘omics’ approach in time series of *E. coli*. Chapter 2 introduces metabolomics informatics. Section 2.1 describes where bioinformatics is in biology. Section 2.2 describes about bioinformatics for metabolomics. Section 2.3 explains about bioinformatics tools I used in this dissertation. Chapter 3 considers the non-targeted metabolomics analysis based on time series experiments of *E. coli*. Section 3.1 explains where metabolomics research is in recent biological research. Section 3.2 describes Materials and methods I used in this dissertation. In Section 3.3, I discuss about the results of metabolomics analysis. Chapter 4 considers the integrative analysis of transcriptomics and metabolomics data. Section 4.1 introduces integrative analyses of several omics data set. Section 4.2 describes Materials and methods I used in this dissertation. In Section 4.3, I discussed about the results of integrative analysis. Finally, Chapter 5 is concluding remarks of this dissertation.

Chapter 2
Metabolomics informatics

2.1 Introduction

Bioinformatics is playing a more and more significant role in the study of modern biology. Bioinformatics is currently a popular term for the application of computational and analytical methods to biological problems [Yu et al. 2002]. The rise of bioinformatics has been largely due to the diverse range of large scale data that require sophisticated methods for analysis. The availability of different types of high-throughput experimental data in the late 1990s has expanded the role of bioinformatics and facilitated the analysis of higher order functions involving various cellular processes [Kanehisa and Bork 2003]. In this section, I review recent advance in bioinformatics for metabolomics and describe the developed bioinformatics platform for Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS).

2.2 Bioinformatics for metabolomics

2.2.1 Identification of metabolites using MS

Metabolite assignments using MS are usually obtained by combining accurate molecular mass, isotopic distribution, fragmentation patterns and other mass spectrometric information available, e.g. databases (DB). The calculation of the chemical formulas that fit a certain accurate mass is generally one of the first steps to obtain a set of alternatives that can lead metabolite annotation for the detected ions. Using an instrument that can provide very high mass accuracies, the range of possibilities of molecular formulas is limited and can lead to the correct molecular formula especially for lower *m/z* values.

One of the most powerful methods for narrowing down the number of molecular formulas is to make use of the isotopic pattern. Compounds that were synthesized by natural precursors comprise monoisotopic and isotope masses according to the natural average abundance of stable isotope abundances [De Laeter et al. 2003]. Table 2.1 shows information of isotope weight and abundance for some elements (C, H, O, N, P, and S).

Kind and Fiehn (2006 and 2007) performed the isotope filter for removing candidate molecular formulas, i.e. using M+1 and M+2 pattern for the formulas as one of the constraints. This is an efficient strategy that can remove more than 95% of false positives. On the other hand, assuming that the intensity of the second isotopic signal corresponds to the ^{13}C signal for small organic molecules, the number of carbons that the molecule contains can be unraveled by natural abundance of ^{13}C (1.07%). The number of carbons in the molecular formula is estimated using the following equation:

$$n = (^{13}\text{C isotopic ion intensity}/^{12}\text{C isotopic ion intensity}) \times (0.9893/0.0107) \quad (2.1)$$

, where *n* represents the number of carbons in the molecule. In view of rigorous atomic mass, mass differences between isotopes of atoms are not identical, i.e. mass differences between ^{1}H and ^{2}H, ^{12}C and ^{13}C, ^{14}N and ^{15}N are 1.0063u, 1.0033u, and 0.9970u, respectively (as shown in 'Isotope difference' column in Table 2.1). If accurate mass values can be detected, the isotope difference for individual atoms can theoretically provide information for estimating molecular formula. In addition, measurements using MS involve multivalent ions derived from identical molecule. Figure 2.1 shows an illustrative example of a fictitious molecule, whose molecular weight is 1,002. For example, in the negative ion mode, monovalent, divalent, and trivalent ions are detected as the ion with *m/z* = 1,001, 500, and 333, respectively, assuming that atomic weight of H^+ is equal to 1. Search for compounds with molecular weights around 500 and 300 based on *m/z* values may lead to false assignments because these *m/z* values are originated from identical compound as *m/z* = 1,001. Therefore, it is necessary to distinguish multivalent ions from obtained MS data and then search the candidates by using several DBs.

The fragmentation pattern of a mass signal can provide structural information about the fragmented ion. From the fragments obtained the structure of the original molecule can be deduced. An *O*-glycosylated flavonoid will, for example, fragment on the glycosidic linkage and only afterwards in the aglycone backbone, if sufficient energy is provided. Isolating one ion and performing tandem MS to the successively obtained fragments can be highly informative for tracking functional groups and connectivity of fragments for structure elucidation of the metabolites. In addition, obtaining accurate mass fragments is also another advantage when there is little knowledge about the possible atomic arrangements of the molecular ion.

The most straight-forward approach for obtaining confirmation of the identity of metabolites in a biological sample is to test commercially available standard compounds on the same analytical system. This approach, however, implies the commercial availability of such standard compounds, i.e. we can get a limited set of compounds as standard. When standard compounds are available, these are useful not only for confirmation of the identity of compounds but also for undergoing quantitative analyses.

Table 2.1: Atomic weights and isotopic compositions of C, H, O, N, P, and S [De Laeter et al. 2003]

Isotope	Atomic mass/u	Mole fraction (%)	Isotope difference
^{1}H	1.0078250319	99.9885	0.993585146
^{2}H	2.0014101779	0.0115	
^{12}C	12	98.93	1.003354838
^{13}C	13.003354838	1.07	
^{14}N	14.0030740074	99.632	0.997034966
^{15}N	15.000108973	0.368	
^{16}O	15.9949146223	99.757	2.004245778
^{17}O	16.99913150	0.038	
^{18}O	17.9991604	0.205	
^{31}P	30.97376149	100	-
^{32}S	31.97207073	94.93	1.99579614
^{33}S	32.97145854	0.76	
^{34}S	33.96786687	4.29	
^{36}S	35.96708088	0.02	

‘Isotope difference’ column corresponds to atomic mass differences between first and second abundant isotopes on earth of individual atoms, i.e. ^{1}H and ^{2}H, ^{12}C and ^{13}C, ^{14}N and ^{15}N, ^{16}O and ^{18}O, ^{32}S and ^{34}S.

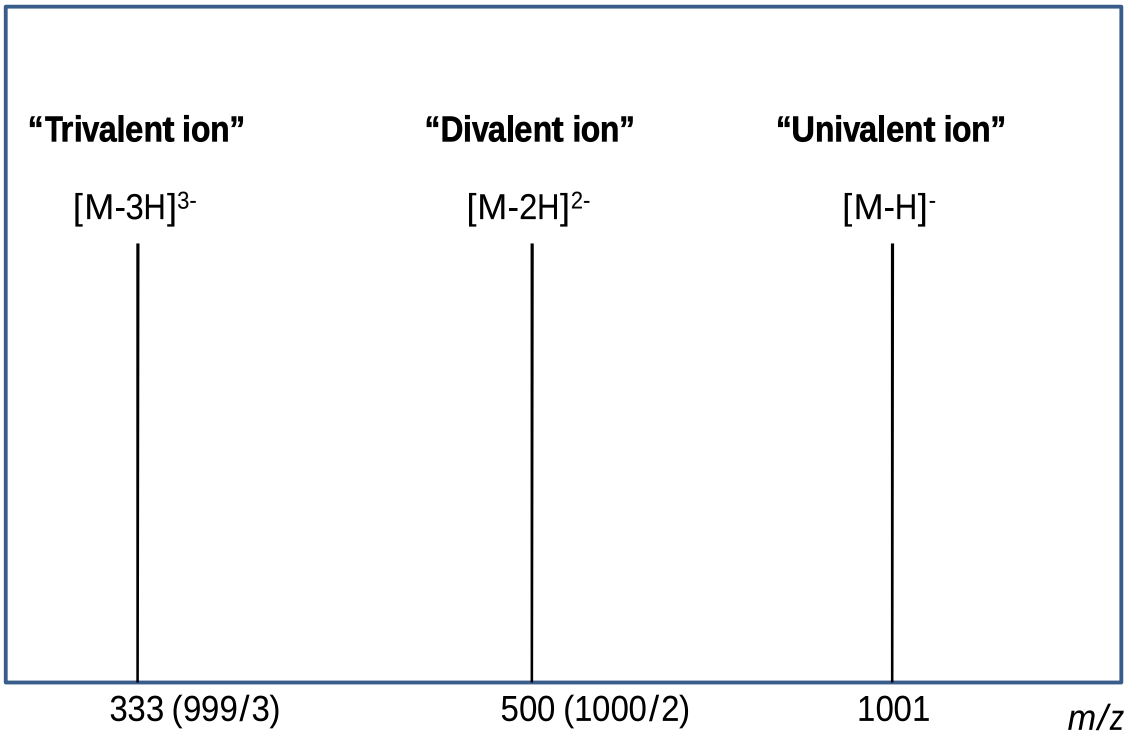


Figure 2.1. Illustrative example of multivalent ions. This is an example for the fictitious molecule, whose molecular weight is 1,002, assuming that the atomic weight of H^+ is equal to 1. Divalent, and trivalent ions are detected the ions with *m/z* = 500, and 333, respectively. Parenthesis indicates that molecular weight is divided by the valence of the ion.

2.2.2 Databases

In metabolite annotation, databases play crucial role for searching the candidate metabolites for each obtained *m/z*. There are several available databases with good number of accumulated compounds. In fact, the bridge between experimental data and the available chemical databases is still weak. Table 2.2 shows several MS databases. Some identification tools such as elemental composition calculation or molecular mass calculation exist among the different instrumental software, but these tools seldom allow spectral matching facilities linked to public databases, like in proteomics applications. More specialized databases might be useful for metabolite annotation, though construction of public metabolite databases has been started by the laboratories within the community.

One of the largest initiatives for the identification of metabolites is the Human Metabolome Project where MS and NMR data are combined with information on molecules [Wishart et al. 2007]. The detailed description of the methods of sample preparation and analysis, conditions of the analytical experiment, chemical information about the metabolites (name, IUPAC name, chemical descriptors such as Chemical Abstracts Service (CAS) registry numbers and InChi and/or structural information, links to chemical databases), experimental spectra and biological source are some of the features included in the metabolite databases. A large portion of compounds accumulated on PubChem [Wheeler et al. 2006] and SciFinder are not natural compounds but artificially synthesized compounds. KNApSAcK and KEGG have accumulated only natural compounds, associated with information of source organisms for each compound.

Table 2.2: Number of metabolite records present in databases

DB	Source	No. Records
KNApSAcK	Nara Institute of Science and Technology (NAIST)	23,287
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Kyoto University / Tokyo University	14,000
Human Metabolome Database (HMDB)	Genome Algebra and Genome Canada	2,300
Golm Metabolome Datbase (GMDB)	Max Planck Institute of Molecular Plant Physiology	
SciFinder	Chemical Abstracts Service (CAS)	30,500,000
PubChem	National Institutes of Health (NIH)	10,100,000

2.3 Bioinformatics tools developed in this dissertation

Different analytical instruments have been applied to metabolome analysis. Analytical tool is necessary for each analytical platform, as there are several analytical methods depending on the microarray platforms. This is also true in the case of the direct-infusion electrospray ionization FT-ICR/MS analysis. In this dissertation, I used our four bioinformatics tools, i.e. DrDMASS+, KNApSAcK, DPClus for metabolomics analysis, and TREBAX for transcriptomics analysis.

2.3.1 FT-ICR/MS analysis tool, DrDMASS+

I have developed the bioinformatics platform for FT-ICR/MS analysis tool, called DrDMASS+. Figure 2.2 shows the main window of DrDMASS+ and the data processing scheme of DrDMASS+ is shown in Figure 2.3. Appendix E describes the detail of this procedure. I briefly describe DrDMASS+ software in this section. The first requirement for the success of metabolomics is the ability to mine the generated data and to perform reliable and comparative analysis. To attain this, a bioinformatics scheme consisting of four stages has been developed: (i) peak correction, (ii) multivariate data processing, (iii) unsupervised learning such as principal component analysis (PCA) and batch-learning SOM (BL-SOM), and (iv) supervised learning such as partial least squares (PLS) regression.

(i) **Peak correction.** Though FT-ICR/MS affords extremely high resolution *m/z* values, analytical data fluctuations are generally associated with the *m/z* values at the three or four decimal place level. So, initially, appropriate *m/z* values must be estimated from the observed *m/z* values. The experimental *m/z* values for the internal mass calibrants (IMCs) were fixed to their theoretical values, and the *m/z* error calibration data were reflected in the *m/z* compensation for all other ion species in each spectral scan (Fig. 2.4).

(ii) **Multivariate data processing.** After compensating *m/z* values, ion peak matching among ten independent scans was done for repeated identifiable *m/z* values. The threshold levels of ion appearance frequencies are freely adjustable. The intensity values of repeatedly observed ions were converted into percentage values of total ion intensity. Thus, metabolomics data from a single biological sample consisted of averaged *m/z* values with intensity information from ten spectral scans.

(iii) **Unsupervised learning.** I implemented two unsupervised learning methods, PCA and BL-SOM. PCA is a multivariate method to project a distribution of data points in a multidimensional space into a space of fewer dimensions, and BL-SOM is a method to classify such data points into groups (grids) accommodating similar decrease/increase patterns [Kanaya et al. 2001; Abe et al. 2003].

(iv) **Supervised learning.** PLS is a method for linearly relating a data matrix $\mathrm{X}\,(M \times N)$ to a vector $\mathbf{y}\,(M \times 1)$ where M and N represent the number of samples and parameters, respectively. The PLS model is represented by Equations (2.2) and (2.3):

$$\mathbf{X} = \sum_{k=1}^{L} \mathbf{t_k}\ \mathbf{p_k^T} + \mathbf{E} \tag{2.2}$$

$$\mathbf{y} = \sum_{k=1}^{L} \mathbf{t_k}\ q_k + \mathbf{e}\,. \tag{2.3}$$

Here, $\mathbf{p_k}$ and q_k are called the loading vector of $\mathbf{X}$, and the coefficient of $\mathbf{y}$ for the *k*th component, respectively. L is the number of components and $\mathbf{t_k}$ is a score vector for the *k*th component. $\mathbf{E}(M \times N)$ and $\mathbf{e}(M \times 1)$ represent the residual matrix and vector, respectively. The number of PLS components, L, is determined to maximize a predicted correlation coefficient (R_{pred}) by leave-one-out cross-validation for each component according to Equation (2.4):

$$R_{pred} = 1 - \frac{\sum \left(y_{obs} - y_{pred}\right)^2}{\sum \left(y_{obs} - \bar{y}_{obs}\right)^2}. \tag{2.4}$$

Here, y_{obs} is an experimental y value, y_{pred} is a predicted y value, and $\bar{y}_{obs}$ is the mean of y_{obs}. The PLS equations (Equations (2.2) and (2.3)) can also be transformed into a linear form represented by Equation (2.5) [Boulesteix and Strimmer 2007; Takahashi et al. 2008]:

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{f}. \tag{2.5}$$

Here, $\mathbf{b}$ is a regression coefficient vector and its elements are represented by b_j (*j* = 1, 2, …, N).

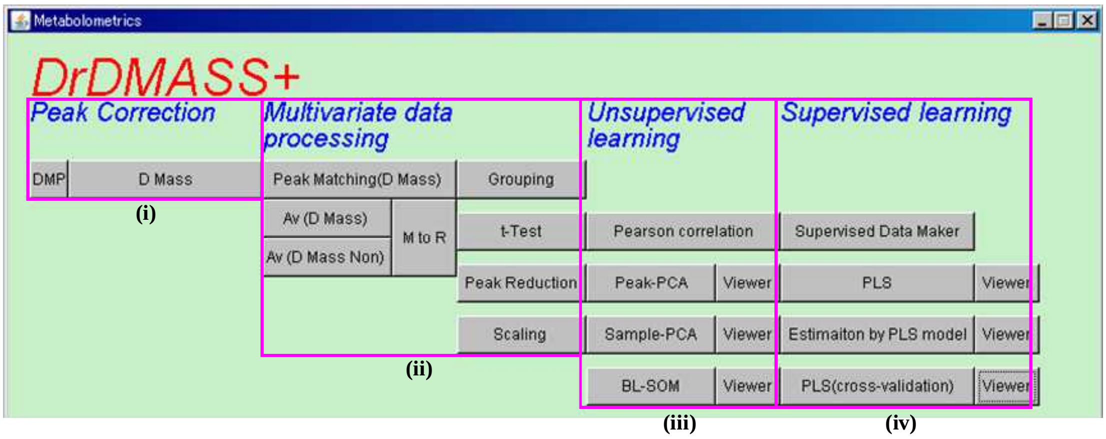


Figure 2.2. Operation window of DrDMASS+ software. (i)-(iv) correspond to each step of DrDMASS+ scheme.

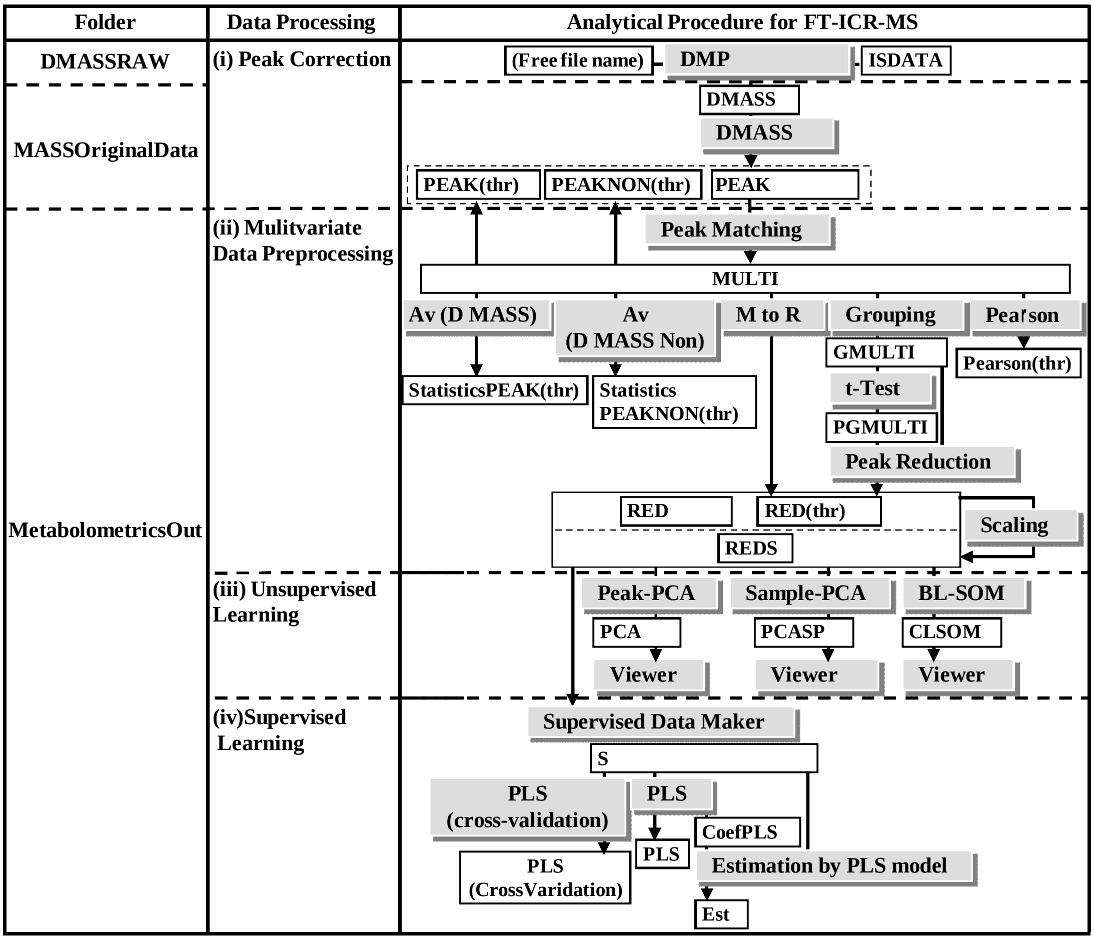


Figure 2.3. Flow diagram of data preprocessing in DrDMASS+ software. Silver boxes correspond to individual processes, and white boxes correspond to prefix in input/output file names. Appendix E describes the detail of the instruction.

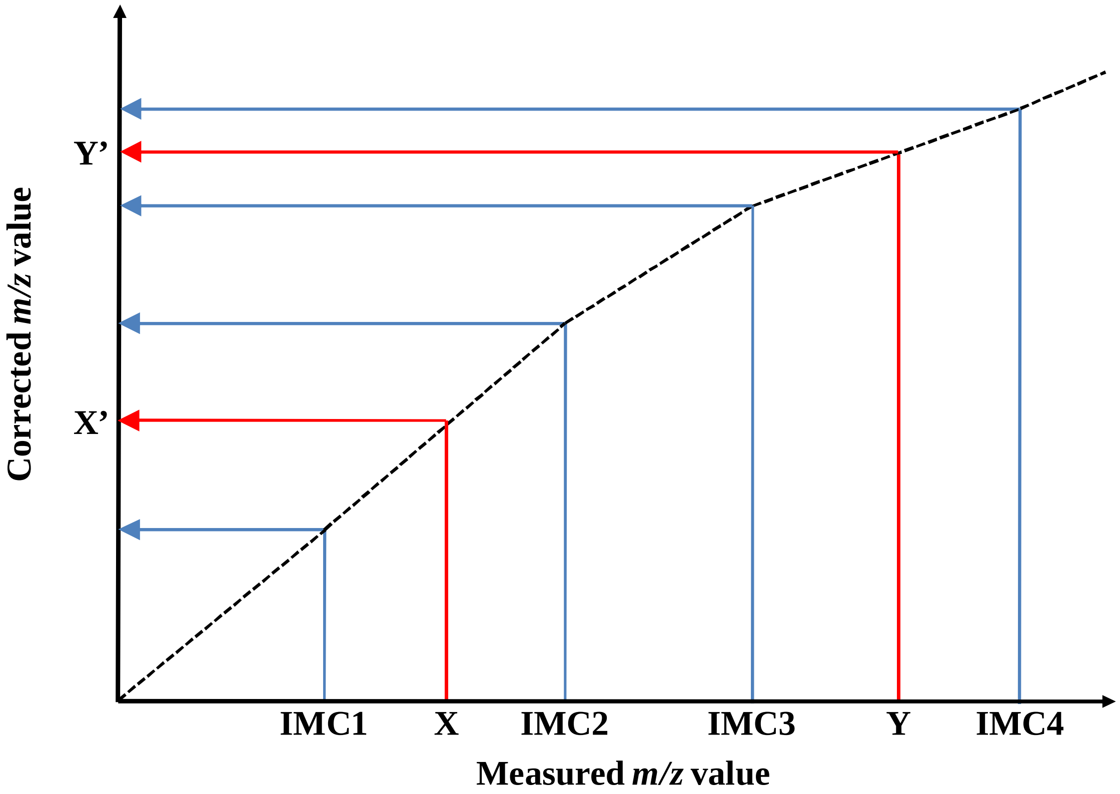


Figure 2.4. Graphical illustration of peak correction. Abscissa and ordinate axes correspond to measured *m/z* value and corrected *m/z* value, respectively. For example, measured values X and Y on abscissa axis which result in X’ and Y’, are corrected by using IMCs (1-4).

2.3.2 Species-metabolite relationship database, KNApSAcK

KNApSAcK is the species-metabolite relationship database which allows high-throughput prediction of metabolite identities from FT-ICR/MS. Information on metabolites in the database can be searched by metabolite name, organism, molecular weight, molecular formula, and mass spectral data taking into consideration the ionization modes ($[M+NH_4]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+H]^+$, and $[M-H]^-$). A total of 46,093 species-metabolite pairs encompassing 23,287 metabolites have so far been compiled (as of 5th September 2008) [Shinbo et al. 2006]. The database enables a high-throughput search using FT-ICR/MS analysis data for metabolite-species relationships together with detailed metabolite information including molecular weight, molecular formula, chemical structures, CAS numbers, biological functions, and references of academic papers.

2.3.3 Graph clustering software, DPClus

In order to elucidate molecular networks within the cell, it is important to extract correlated relations between genes, proteins, metabolites, etc. DPClus is a graph clustering software that can extract densely connected clusters using an algorithm that is based on density and periphery tracking of clusters [Altaf-Ul-Amin et al. 2006]. While using DPClus, it is necessary to provide a value of minimum density for the generated clusters (d), a minimum value for cluster property for the nature of periphery tracking (cp_{in}), and a minimum number of objects in a cluster.

2.3.4 Microarray analysis tool, TREBAX

TREBAX is a microarray analysis tool mainly for cDNA microarray data [Kobayashi et al 2007]. Gene expression levels are evaluated by measuring the fluorescence intensity for each spot, and there is usually some experimental variation that occurs in every microarray experiment. It is, therefore, important to minimize experimental variation, and although several methods of microarray normalization have been developed [Quackenbush 2002; Yang et al. 2002], there are usually some false-positive data arising when analyzing gene expression data collected via microarrays. Normalization of the logarithmic ratio of expression intensity between target (R_i) and control (G_i) experiments was carried out based on MA plots [Dudoit et al. 2002], which can show the intensity-dependent ratio of raw microarray data. The MA plot uses M_i ($\log_{10}(R_i/G_i)$) as the y-axis and A_i ($\log_{10}\sqrt{R_iG_i}$) as the x-axis. By plotting values of A_i on the abscissa and M_i on the ordinate of a coordinate system, it is possible to evaluate the bias error with respect to the average logarithmic intensities. The normalized log ratio M''_i was estimated as the difference between M_i and baseline M'_i. Here, using the relation between M_i and A_i ($M_i = f(M_i) + \varepsilon_i$, where ε_i is the difference between M_i and $f(A_i)$ for the *i*th gene for the MA plot), the baseline for the *i*th gene was estimated by $M'_i = f(A_i)$. With this methodology, it is assumed that there was no large error due to expression intensity in the majority of the spots. Figure 2.5 shows one example of the MA plots before and after normalization using TREBAX.

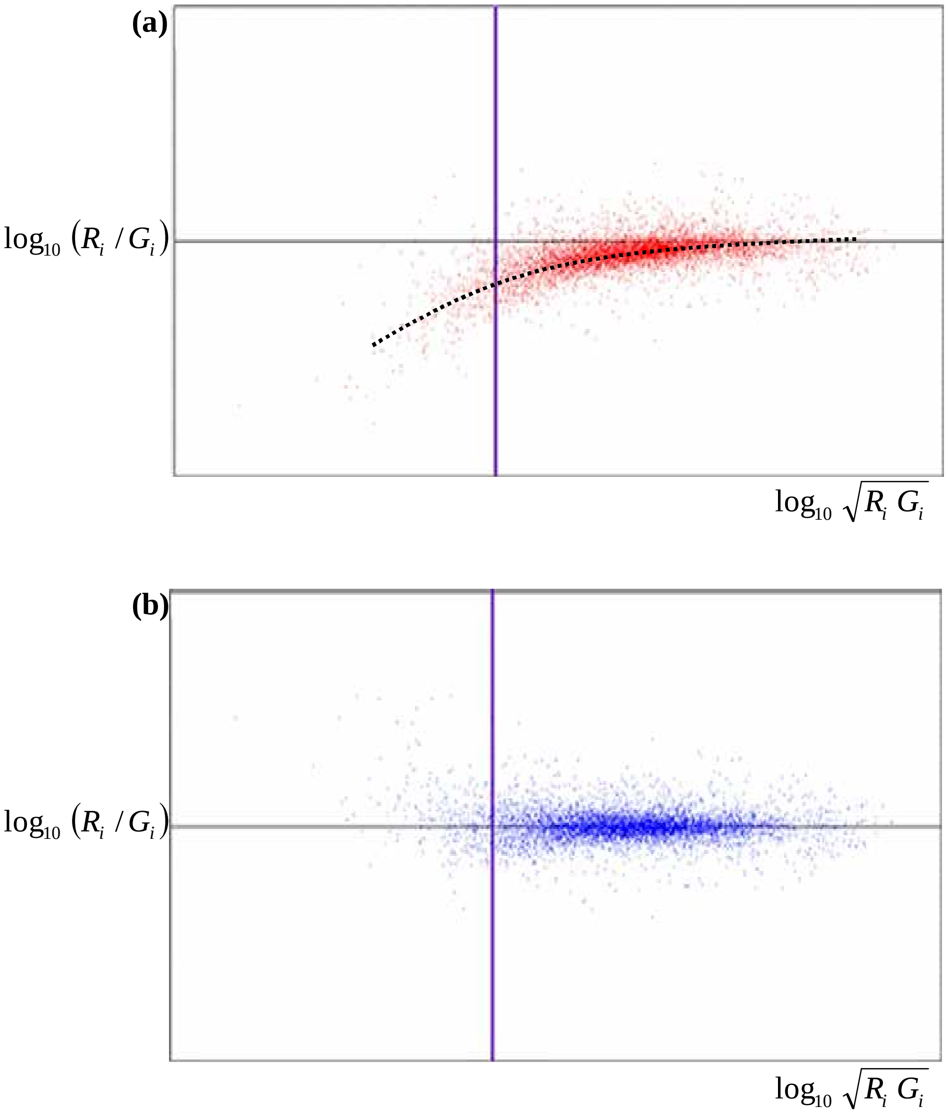


Figure 2.5. MA plot. **a** MA plot before and **b** after normalization using TREBAX.

Chapter 3
Nontargeted metabolomics in *Escherichia coli*

3.1 Introduction

Comprehensive metabolomics is clearly distinct from conventional metabolism studies in that it addresses whole cellular activities rather than just focusing on enzymes, reactions, or metabolites. Over the past decade, methods that offer both high resolution and sensitivity for the measurement of a vast number of metabolites have been established and two major approaches, targeted and non-targeted metabolomics studies, have been developed [Fiehn 2002; Villas-Boas et al. 2005]. Targeted metabolomics plays a crucial role in understanding the primary effects of genetic alternations based on restricted information of a class of metabolites, and analytical procedures often need to include processes for identification and quantification of selected metabolites. Only recent advances in mass spectrometry have allowed non-targeted metabolomics, which is intended for unbiased analyses such as mapping metabolite profiles in the whole cellular processes in a given organism.

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS) is the best MS technology for obtaining exact mass measurements owing to its great resolution and accuracy [Marshall et al. 2002; Aharnoni et al. 2002], and several outstanding FT-ICR/MS-based metabolomics strategies have been reported [Hirai et al. 2004; Hirai et al. 2005; Tohge et al. 2005; Oikawa et al. 2006; Nakamura et al. 2007; Suzuki et al. 2008]. Development of a general scheme for FT-ICR/MS-based metabolomics, with the aid of its potential for the high resolution measuring power together with ion signal intensity information, should thus make a significant contribution to metabolomics studies. To attain this purpose and to understand the cell system based on the components of metabolites, I apply chemometrics and bioinformatics approaches to FT-ICR/MS data. Among a variety of metabolomics strategies, FT-ICR/MS offers a unique opportunity in non-targeted metabolomics studies owing to its extreme accuracy (below 1 ppm) in the mass measurement. Thus, chemical formulas and molecular identities of metabolites can be predicted with the aid of high precision mass spectrometry data, and can also be easily linked to reported metabolites.

Metabolomics research currently confronts a problem associated with high-throughput data acquisition technologies including chromatography-coupled MS and FT-ICR/MS which have facilitated simultaneous detection and quantification of a large number of metabolite-derived peaks without metabolite assignment [Hall 2006]; a very similar situation has arisen in genomics research in that technologies for determination of nucleotide sequence in the whole genome has progressed without annotations of gene functions [Stein 2001]. Progress in annotation of metabolites in metabolomics can bridge the gap between the data and their biological interpretation. The problem with annotation of metabolites is that there is only a piece of information about peaks corresponding to precise molecular weight for metabolite-derived ions in MS, but when quantities of ions in a time series experiment are measured, metabolite-derived ions such as isotope ions and multivalent ions could be categorized by correlations between ions originated from identical metabolites, which can lead to more precise annotation of ions as described in Section 2.2.1. Thus, correlation analysis of ions may be a powerful approach to annotation of metabolites in metabolomics.

In this dissertation, I propose a procedure for metabolite annotation using the data obtained from FT-ICR/MS by taking into consideration classification of metabolite-derived ions. Here, I perform the non-targeted comprehensive analysis of metabolomics for the time series measurements in *E. coli*, and discuss a metabolic profiling scheme on the basis of FT-ICR/MS analyses furnished with a bioinformatics scheme including data preprocessing, classification of ions originated from identical metabolites, and supervised and unsupervised learning algorithms for metabolomics.

3.2 Materials and methods

3.2.1 Strains and growth conditions

The strain used in this dissertation was *E. coli* K-12 W3110. An aliquot (8 ml) of an overnight liquid culture of W3110 in LB medium at 37℃ was inoculated into 2 l LB (pH 7.4) medium in a 3 l jar-fermenter. Cells were grown continuously at 37℃ for ～12 h, adjusting the agitation speed to 300 r.p.m. with fixed 2 l min^{-1} air flow rate. Growth was monitored by measuring the optical density at 600 nm (OD$_{600}$).

3.2.2 Sample preparation

A culture medium was passed through a 0.45-μm-pore-size filter (Durapore Membrane, Millipore). Residual *E. coli* cells on the filter were washed with Milli-Q water and then plunged into 2 ml methanol [Soga et al. 2003]. After sonication for 1 min, the methanol solution was kept at 4℃ for ～20 h. The solution was then filtered through disposable membrane filter units (DISMIC-13JP, ADVANTEC), evaporated, and stored at -80℃ until use. Upon FT-ICR/MS analysis, the extracts were dissolved in 50% (v/v) acetonitrile/water. A set of 2,4-dichlorophenoxy acetic acid ($[M-H]^-$ = 218.96212), ampicillin ($[M-H]^-$ = 348.10235), 3-[(3-cholamidopropyl) dimethylammonio] propanesulfonic acid ($[M-H]^-$ = 613.38920), and tetra-*N*-acetylchitotetraose ($[M-H]^-$ = 829.32078) was used as the internal mass calibrants (IMCs) in the negative ion mode analysis.

3.2.3 FT-ICR/MS conditions

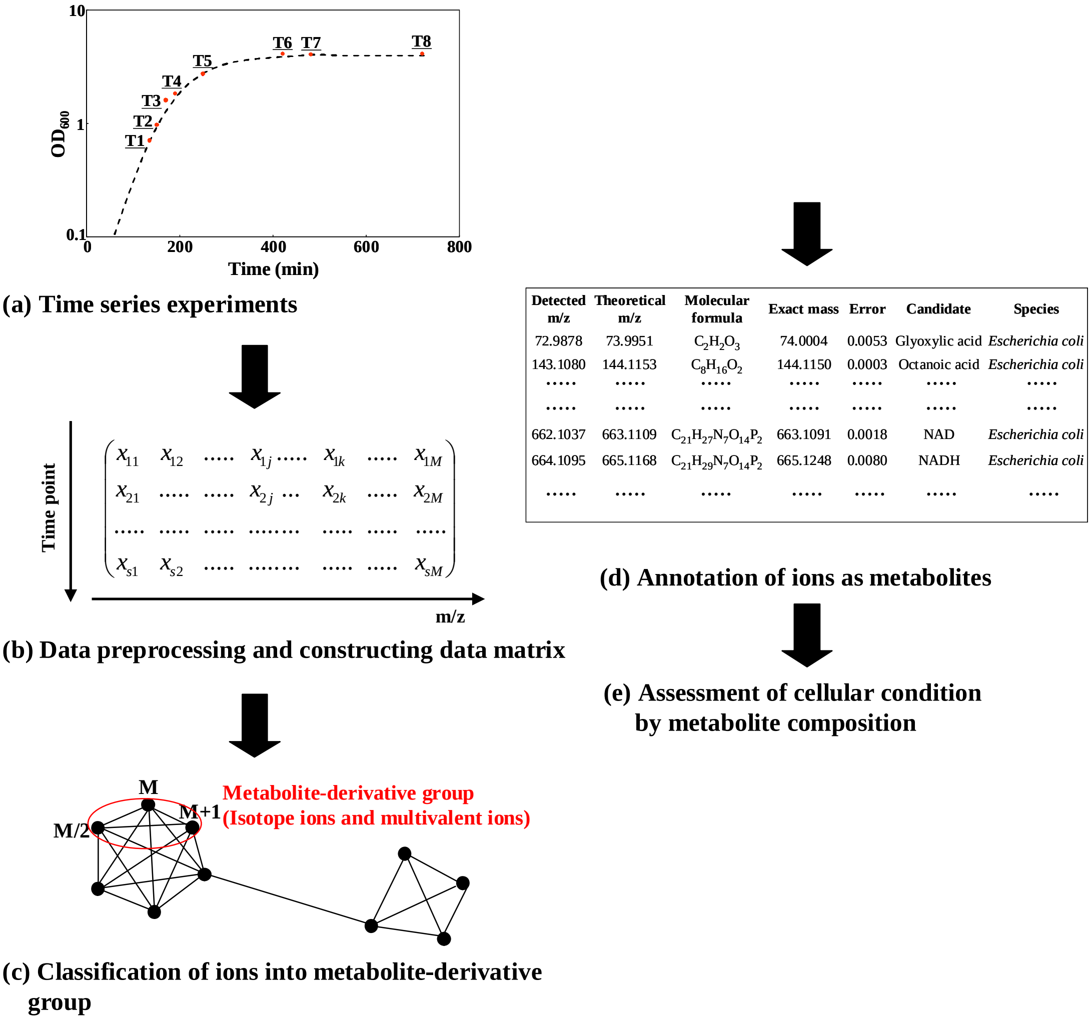
Mass analysis was done in the negative ion mode using an IonSpec Explorer FT-ICR/MS (IonSpec) equipped with a 7-tesla actively shielded superconducting magnet. Ions were generated from an ESI source with a fused silica needle of 0.005-inch i.d. Samples were infused using a Harvard syringe pump model 22 at a flow rate of 0.5 to 1.0 μl min^{-1} through a 100 μl Hamilton syringe. All the experimental events were controlled using Omega8 software (IonSpec). Briefly, the potentials on the electrospray emitters were set to -3.0 kV for the negative electrosprays. The base pressure in the source region was approximately 5×10^{-5} torr (1 torr = 133.3 Pa). For the negative electrosprays, sample solutions were prepared in 50% (v/v) acetonitrile/water with 0.1% (v/v) of ammonium hydroxide. Ionized metabolites were accumulated for a period of 2,500-5,000 ms in a hexapole ion trap/guide and transferred through a radiofrequency-only quadrupole into the FT-ICR cell in the superconducting magnetic field, where they were again trapped. The direct current potential in the negative ion mode analyses was 2 V during the ion accumulation and -2 V for the ion transfer into the FT-ICR cell. These ions trapped in the hexapole were extracted for transfer into the FT-ICR cell. In the negative ion modes, the potential on the extraction plate was -12 V during the ion trapping and were reversed to 2 V for the extraction. The base pressure in the analyzer region was set to approximately 4×10^{-10} torr. ESI-MS spectra were acquired over the *m/z* range 55-1,000 from 1,024,000 independent data points. MS/MS analyses were done using the sustained off-resonance irradiation SORI-CID methods [Gauthier et al. 1991; Laskin and Futrell 2005]. SORI R_f was set at 0.5-1.5 V, and the N_2 collision gas was used with a 400-ms pulse.

3.3 Results and discussion

3.3.1 Data processing of FT-ICR/MS: from data acquisition to assessment of cellular conditions according to metabolite composition

The concept of FT-ICR/MS data processing from data acquisition of a time series experiment to describe cellular conditions from exponential to stationary growth phase by metabolite consists of five steps (Fig. 3.1). Time series experiments are a popular method for studying a wide range of biological systems. In bacteria, there are a few reported papers which comprehensively analyzed bacteria intrametabolites [Brauer et al. 2006]. However, to my knowledge, there is no paper about bacteria which addresses total intrametabolite profiling. In order to elucidate intrametabolite profiling in a whole cell, I performed the time series experiment in *E. coli* (Fig. 3.1a). Samples were collected at 135, 150, 170, 190, 250, 420, 480, and 720 min postinoculation (which correspond to T1, T2, T3, T4, T5, T6, T7, and T8, respectively), and metabolites were extracted, and measured by FT-ICR/MS. FT-ICR/MS raw data were processed for differential metabolomics according to the peak correction and peak matching of the DrDMASS+ program as described in Section 2.3.1. I selected *m/z* values whose appearance frequencies were higher than 50% among ten scans. Thus, differential metabolomics was studied in terms of corrected *m/z* values with average signal intensities of reproducible ions from ten independent spectral data. The observed *m/z* values for ions in individual measurements in the time series experiments were calibrated with those of internal standards [Oikawa et al. 2006; Takahashi et al. 2008]. Peak matchings were carried out to make a matrix consisting of intensities for *m/z* values and time points (Fig. 3.1b) utilizing a metabolomics platform, based on FT-ICR/MS incorporating the metabolite profiling tool DrDMASS+. After the processing step, 220 independent ions were detected in the negative ion mode analysis. Thus, time series data matrix consists of intensities of 220 independent ions corresponding to metabolites for eight measurement points.

There are many ions originated from identical metabolites, i.e. isotope ions and multivalent ions. If detected ions are classified into identical metabolite-derived ion groups, I can use further information for annotating chemical structures in metabolites, because isotope pattern allows us to estimate the number of carbons in molecular formulas for metabolites, and the actual number of metabolites included in samples can also be estimated. When different detected ions are derived from identical metabolite, those should be theoretically correlated with each other in time series data. So, correlations between ions can lead to the estimation of molecular formula by using isotope information. This step was carried out by DPClus software (Fig. 3.1c) developed by Altaf-Ul-Amin et al. (2006). DPClus software can make it possible to visualize the correlation network, and give us complete subgraphs when the density for the generated clusters is equal to 1. All nodes within a complete graph are connected with each other. Thus it is expected that multiple ions derived from identical metabolite can be detected within a complete graph, if the density is set to be 1 in DPClus. After classification of ions into specific metabolite derivative groups, I performed annotation of ions as metabolites using public natural compound databases, KNApSAcK [Shinbo et al. 2006] and KEGG [Bairoch 2000; Goto et al. 2002; Kanehisa et al. 2006] (Fig. 3.1d), and cellular conditions were characterized by the composition of metabolites using two approaches, supervised and unsupervised learning. Cellular condition could be assessed by the metabolite composition using principal component analysis (PCA), and the relationship between cell densities and the metabolite composition, reflecting transition from exponential to stationary phases, could be understood by using partial least squares (PLS) regression (Fig. 3.1e). Marker metabolites significant in exponential and stationary growth were determined using PLS regression.


10
1
0.1
OD_{600}
0
200
400
600
800
Time (min)
T1
T2
T3
T4
T5
T6
T7
T8
Time point
m/z
x_{11} x_{12} ..... x_{1j} ..... x_{1k} ..... x_{1M}
x_{21} .... x_{2j} ... x_{2k} ..... x_{2M}
....
x_{s1} x_{s2} .... x_{sM}
M
M/2
M+1
Metabolite-derivative group
(Isotope ions and multivalent ions)

(a) Time series experiments

(b) Data preprocessing and constructing data matrix

(c) Classification of ions into metabolite-derivative group

Detected m/z	Theoretical m/z	Molecular formula	Exact mass	Error	Candidate	Species
72.9878	73.9951	$C_2H_2O_3$	74.0004	0.0053	Glyoxylic acid	*Escherichia coli*
143.1080	144.1153	$C_8H_{16}O_2$	144.1150	0.0003	Octanoic acid	*Escherichia coli*
.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....
662.1037	663.1109	$C_{21}H_{27}N_7O_{14}P_2$	663.1091	0.0018	NAD	*Escherichia coli*
664.1095	665.1168	$C_{21}H_{29}N_7O_{14}P_2$	665.1248	0.0080	NADH	*Escherichia coli*
.....	.....	.....	.....	.....	.....	.....

(d) Annotation of ions as metabolites

(e) Assessment of cellular condition by metabolite composition

Figure 3.1. Data processing scheme consisting of five steps. **a** Time series experiments in *E. coli*. The growth curve shows eight time points (135, 150, 170, 190, 250, 420, 480, and 720 min postinoculation corresponding to T1, T2, T3, T4, T5, T6, T7, and T8, respectively), at which samples were taken, and metabolites were extracted, and measured by FT-ICR/MS. **b** Data structure after data preprocessing by DrDMASS+ including peak correction and peak matching. M and s show the number of detected ions and samples, respectively. **c** Classification of ions into metabolite derivative groups by DPClus based on the correlations between detected ions. **d** Annotation of ions by searching metabolite databases (KNApSAcK and KEGG). **e** Assessment of cellular conditions according to metabolite composition by using multivariate analyses.

3.3.2 Classification of ions into metabolite derivative groups

The difference of *m/z* value between isotope ions originated from carbon atom (1.0033 u) is a clue for determining whether or not the ions are originated from identical metabolites. Furthermore, ions, originated from identical metabolites, occurring in different ion valence are also detected. Isotope intensity pattern of a metabolite in an MS chart can serve as a powerful additional constraint for removing wrong elemental composition candidates [Kind and Fiehn 2007]. When intensities of ions are correlated to each other in a time series experiment, those ions would be expected to be originated from an identical metabolite. Tautenhahn et al. (2007) successfully combined highly correlated pairs of mass signals in LC-MS to chemical relation hypothesis groups. Thus, taking into consideration the differences of *m/z* values for ions and correlation of time series profiles of ions, isotope ions can be classified into metabolite derivative groups, which lead to estimation of molecular formula of metabolites. To attain this, I visualized all correlations in a time series experiment between ions. Pairwise ion-ion correlations were calculated by Pearson's correlation coefficient (r) [Fisher 1958]. I extracted a set of 742 unique binary relations involving 148 ions by the threshold $r \geq 0.9$ ($p \leq 2.3\times10^{-3}$, n = 8) and visualized this by using the graph-clustering method called DPClus [Altaf-Ul-Amin et al. 2006]. Out of total 220 detected ions, 72 ions do not show significant correlation with other ions. Figure 3.2 shows the configuration of the 742 relations including 148 ions assigned to 11 isolated clusters (ID = 1 to 11). Two largest isolated subgraphs consisting of 43 and 28 ions, respectively, can be characterized by six clusters (ID = 1-1 to 1-6) and three clusters (ID = 2-1 to 2-3), of size $>$ 2, which are all complete graphs where an edge connects every pair of distinct vertices within the same cluster. Ions assigned to multiple complete subgraphs are depicted by blue nodes. Relations between ions and cluster IDs are listed in Appendix A.

I assume that ions which belong to the same cluster and have appropriate *m/z* difference of ^{13}C and certain valences have originated from identical metabolites. Initially, to determine isotopic ion pairs, I searched ion pairs under conditions that the ion pairs have not only correlation with each other but also appropriate *m/z* difference for certain *k*-valence, i.e. $\mathrm{M}^{-} + \mathrm{H}^{+} = 2\mathrm{M}^{2-} + 2\mathrm{H}^{2+} = \ldots = k\mathrm{M}^{k-} + k\mathrm{H}^{+}$. Furthermore, to determine ion pairs originated from identical metabolites, the search was extended to ions other than isotope ions. Thus, 19 metabolite derivative groups consisting of multiple ions including isotope and multivalent ions were identified (Fig. 3.2, surrounded by red broken lines). In total, 148 ions were classified into 102 metabolite derivative groups which include isotope ions and multivalent ions.

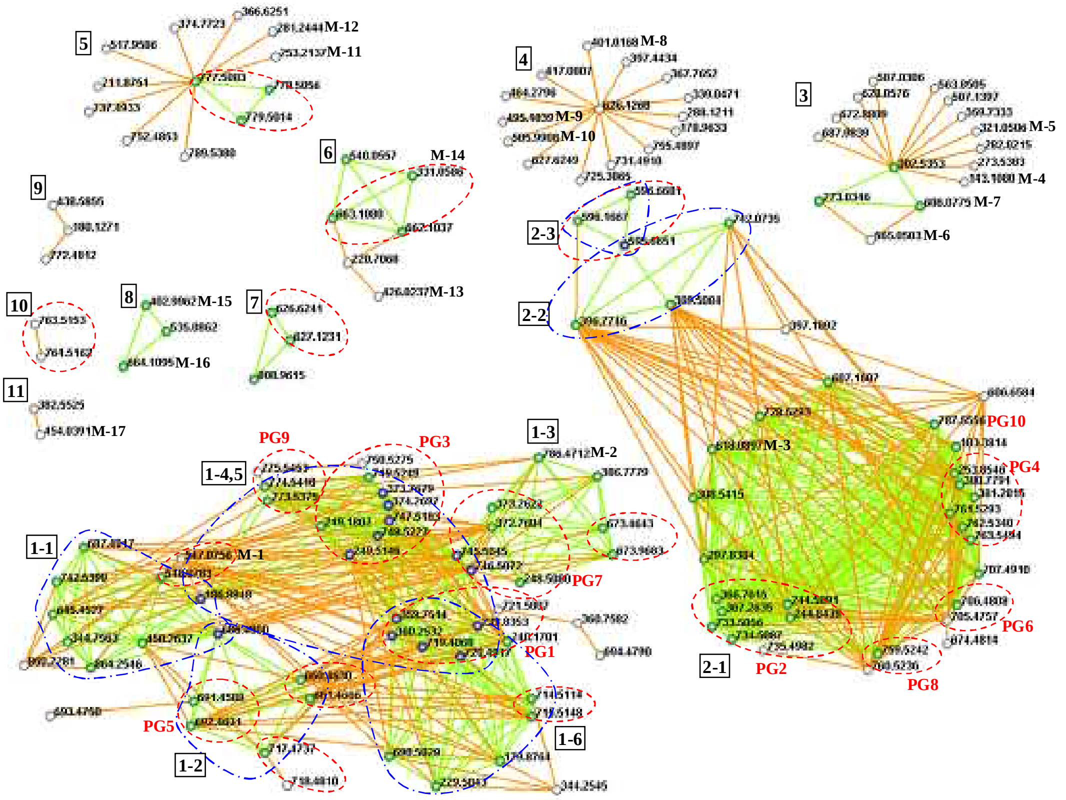


Figure 3.2. Correlation analyses based on the graph clustering. A graph sharing correlation between ions and densely connected clusters is shown. Each boxed black number (1-11) corresponds to a cluster ID detected by the graph clustering. Each node corresponds to an ion with *m/z* value indicated. The colors of nodes represent the ions within a cluster (*green*), the common ions among clusters (*blue*), and the other ions (*silver*). The intracluster edges are *green* and intercluster edges are *orange*. The thick blue broken circles show the clusters 1-1, 1-2, 1-3, 1-4,5, 1-6, 2-1, 2-2, and 2-3. The red dotted circles show isotope ions. PG1-PG10 (phosphatidylglycerol) are shown in *red*. M-1 to M-17 near the nodes are the identities of ions which have candidates according to the KNApSAcK search: M-1, dTDP-L-rhamnose; M-2, BE 32030B; M-3, ADP-L-glycero-beta-D-manno-heptopyranose; M-4, octanoic acid; M-5, dTMP; M-6, UDP-D-glucose, UDP-D-galactose; M-7, UDP-*N*-acetyl-D-mannosamine, UDP-*N*-acetyl-D-glucosamine; M-8, dTDP; M-9, kinamycin A, kinamycin C;, M-10, ATP, dGTP; M-11, omega-cycloheptanenonanic acid; M-12, oleic acid,

cis-11-octadecanoic acid, omega-cycloheptylundecanoic acid; M-13, adenosine 3’, 5’-bisphosphate, ADP, dGDP; M-14, NAD; M-15, UDP; M-16, NADH; M-17, antibiotic MI 178-34F18A2, antibiotic MI 178-34F18C2

3.3.3 Annotation of ions

The concept of metabolite annotation comprised of mass spectral annotation and biological metadata annotation including description of actual experimental conditions that help unravel the biological role of metabolites by their changes in levels in response to genetic and environmental perturbation [Fiehn et al. 2005; Scholz and Fiehn 2007]. In this dissertation, I use the term ‘metabolite annotation’ to describe a procedure of providing chemical characterization to individual metabolite-derived ions; thus, the annotation procedure can be classified as a mass spectral annotation, which is important for interpretation of cellular conditions according to metabolite compositions. There are two distinct ways to provide metabolite annotation: an exhaustive computation of all chemically possible isomeric structures or a query of databases for known natural compounds. In this dissertation, I annotated ions based on the latter method together with additional evidence of chemical information such as MS/MS fragmentations. Three publicly available databases concerning natural products are PubChem [Wheeler et al. 2006], KEGG [Kanehisa et al. 2008], and KNApSAcK [Shinbo et al. 2006]. The PubChem database is comprised of records for over 19.6 million compounds with over 11 million unique structures including small molecules, particularly diagnostic and therapeutic agents. In this dissertation, detected ions are natural compounds and it is better to search the databases that mainly contain natural products. In KEGG, the metabolic pathways are constructed by interspecies gene relations such as orthologs and paralogs, so metabolite-species relationships can be obtained via information of enzymes. The KEGG database focuses on metabolites related to known metabolic pathways and includes around 13,000 metabolites. On the other hand, the relationships between metabolites and their biological origins have been addressed systematically in the KNApSAcK database, which has accumulated 46,093 records (species-metabolite pairs) encompassing 23,287 metabolites (as of 5th September 2008). The total number of secondary metabolites for which molecular structures have been elucidated is estimated to be 50,000 [De Luca and St Pierre 2000]. So, around 47% of metabolites have been compiled in the database and this is considered to be enough for searching candidates including species information. As the first stage, I searched metabolites in two databases (KEGG and KNApSAcK) by molecular weights estimated from *m/z* values for ions.

Isotope patterns allow us to estimate the number of carbons in molecular formulas for metabolites because natural compounds on earth reflect the natural abundance of stable elemental isotopes, such as ^{13}C (which is found at approximately 1.07%) [De Laeter et al. 2003]. The abundance of isotope ions is dependent on the actual elemental composition and can therefore serve as a powerful filter in calculating unique elemental compositions from mass spectral data [Kind and Fiehn 2007]. In view of rigorous atomic mass, mass differences between isotopes of atoms are not identical, e.g. mass differences between ^{1}H and ^{2}H, ^{12}C and ^{13}C, ^{14}N and ^{15}N are 1.0063 u, 1.0033 u, and 0.9970 u, respectively. Several software methods calculate isotope patterns of compounds based on the assumption that mass differences of atomic isotopes for different atoms can be considered to be identical [Boecker et al. 2006]. Because of the extent of high resolution in FT-ICR/MS, the isotope differences cannot be neglected, i.e. it could be possible to separately detect each isotope ion containing ^{2}H, ^{13}C, ^{15}N and so on. But intensities of isotope compounds with isotope atoms other than ^{13}C would be too small to be detected, because the probability of ions containing ^{2}H, ^{15}N and so on is much lower compared with ions containing ^{13}C. So assuming that an isotope ion M+1 is derived from only ^{13}C, a relative ratio of M (^{12}C) and M+1 (^{13}C) separated by the difference (1.0033 u) of *m/z* values for two peaks can allow us to estimate how many carbon atoms a compound should contain without prior information about the structure. In addition to this, MS/MS fragmentation patterns provide structural information of metabolites, so I performed MS/MS analysis for the five peaks corresponding to *m/z* = (A) 662.1037, (B) 719.4868, (C) 733.5056, (D) 747.5183, and (E) 761.5293.

In ion A, the intensity of *m/z* = 662.1037 is highly correlated with those of *m/z* value 663.1080 in cluster 6, so those would be isotope ions, i.e. *m/z* = 662.1037 (M) and *m/z* = 663.1080 (M+1) because of the difference 1.0043. The number of carbon atoms estimated by the intensity ratio of 662.1037 to 663.1080 was in the range of 19 and 21 at the 99% confidence interval of the *t* test (Table 3.1). I got 845 possible molecular formulas consisting of six types of atoms (C, H, O, N, P, and S) in the range of ±0.01 for an ion with *m/z* = 662.1037. After reducing candidates that do not have the estimated number of carbon atoms, I could get 92 possible candidates, i.e. about 89% candidate molecular formulas were excluded. The candidate metabolite for ion A according to the KNApSAcK search (no hits in KEGG database) is nicotinamide adenine dinucleotide (NAD) ($C_{21}H_{27}N_7O_{14}P_2$), and ions obtained from MS/MS analysis (*m/z* = 540.0782, 328.0532) for ion A are consistent with the fragmentation pattern of NAD (Fig. 3.3), i.e. fragmentation ions with *m/z* = 540.0782 and 328.0532 could be assigned to ($[C_{15}H_{20}N_5O_{13}P_2]^-$) [theoretical *m/z* = 540.0533] and ($[C_{10}H_{11}N_5O_6P]^-$) [theoretical *m/z* = 328.0447], respectively. Thus, I annotated the ions corresponding to *m/z* = 662.1037 and 663.1080 in cluster 6 as NAD and also *m/z* = 331.0586 in cluster 6 as a doubly charged ion ($[M-2H]^{2-}$)

of NAD.

Next, I annotated four selected monoisotope ions *m/z* = (B) 719.4868, (C) 733.5056, (D) 747.5183, and (E) 761.5293. Though the candidate metabolites could not be obtained by the database search, fragmentation ions for those were obtained by MS/MS analyses in Figure 3.4a-d. In the MS/MS spectrum corresponding to the ion with *m/z* = (B) 719.4868 (Fig. 3.4a), two peaks for fragment ions (i.e. *m/z* = 253.2181 and 255.2337) could be assigned to an unsaturated fatty acid ($C_{16}H_{30}O_2$) [theoretical *m/z* = 253.2167 ($[R_2O]^-$)] and a saturated fatty acid ($C_{16}H_{32}O_2$) [theoretical *m/z* = 255.2324 ($[R_1O]^-$)], indicating that the ion with *m/z* = 719.4868 is a phosphatidylglycerol (PG). All ions (B-E) possess some common identifiable peaks (i.e. *m/z* = 255.2337, 391.2260, 465.2628, and 483.2735 in Fig. 3.4a), suggesting that they are similar types of molecules, i.e. four ions B-E, referred to as PG1 to PG4, respectively, would be different types of PGs summarized in Figure 3.5. The numbers of carbon atoms estimated at the 99% confidence interval of the *t* test were also true for all four ions, suggesting that identification of isotope ions based on the graph clustering and estimating the number of carbon atoms by the confidence interval of the *t* test could be also reliable to reduce the number of candidate molecular formulas. I also checked the effect of other constraints for reducing candidates, e.g. using element ratio constraints (H/C 0.2-3.1, O/C 0-1.2, N/C 0-1.3, P/C 0-0.3, and S/C 0-0.8) [Kind and Fiehn 2007], but there was no impact after reducing by the *t* test (element ratio column in Table 3.1), suggesting that if the isotope pattern data for a metabolite in a time series can be given, the relative ratio of isotope ions (M and M+1) can efficiently narrow down candidate molecular formulas even without other constraints. Though incorporating chromatographic separation systems into the FT-ICR/MS system is helpful to estimate the relative ratio of isotope ions and also to predict the candidate molecular formula of unknown ions in a single measurement, time series data set can also ensure the possibility of candidate molecular formulas from a statistical perspective, i.e. the confidence interval of the *t* test.

It has been reported that PGs are composed of various molecular species [Ishinaga et al. 1979]. In this dissertation, another six metabolite derivative groups can be annotated as PGs by following three ‘rules’ in fatty acid metabolism (Fig. 3.6):

(1) Cyclopropane fatty acid (CFA) formation occurs as one of the modifications of phospholipids [Chang and Cronan 1999; Grogan and Cronan 1997]. A mass difference of 14.0157 corresponding to CFA was obtained in five pairs of PGs (PG1 and PG2, PG3 and PG4, PG5 (*m/z* = 691.4588) and PG6 (*m/z* = 705.4757), PG7 (*m/z* = 745.5045) and PG8 (*m/z* = 759.5242), and PG9 (*m/z* = 773.5375) and PG10 (*m/z* = 787.5556));

(2) An elongation process occurs in fatty acids [Magnuson et al. 1993], i.e. a mass difference of 28.0313 u corresponds to one cycle of two-carbon addition in fatty acid biosynthesis, which was obtained in six pairs of PGs (PG5 and PG1, PG1 and PG3, PG7 and PG9, PG6 and PG2, PG2 and PG4, and PG8 and PG10);

(3) A desaturation process, i.e. a mass difference of 2.0157 was obtained in two pairs of PGs (PG3 and PG7, and PG4 and PG8). So, annotation of PG5 to PG10 could be validated by enzyme reactions in lipid metabolism.

I searched the other 169 metabolite derivative ions using KNApSAcK (threshold was set to be ± 0.01), and obtained 163 candidate metabolites from the search of the entire metabolite inventory in the database. Based on the species-metabolite relationships and MS/MS analyses above, I was finally able to assign 33% of 220 detected ions to candidate metabolites. If the search was restricted to only bacteria-metabolite relationships of the KNApSAcK database, then I found 26 ions are related to 38 metabolites (Table 3.2). Out of these, there is only one whose candidates have different molecular formulas. The other 25 ions correspond to unique elemental compositions, suggesting that the information of species-metabolite relationships is efficient to extract reliable lists of candidate metabolites. In Table 3.2, there are several candidates for the molecules, which have never been found in *E. coli*. There are three isomeric candidate molecules for *m/z* = 281.2444. This ion is most likely to oleic acid, because other two candidates have never been reported in *E. coli* based on the species-metabolite relationships of KNApSAcK. In addition, the candidates for *m/z* = 454.0391 are antibiotic MI 178-34F18A2 and antibiotic MI 178-34F18C2, that would never be found in *E. coli*. Information about these molecular formulas, i.e. $C_{20}H_{19}Cl_2NO_7$, could be useful clue for estimating the molecular formula of this ion, although these are not likely to be true candidates for *E. coli*. Overall, not only structure information, but also species-metabolite information could help us reach the exact annotation of ions. Also, taking into consideration the possibility of detecting pieces of compounds is necessary for annotation scheme, since *E. coli* has peptide antibiotics such as microcin B17 [Roy et al. 1999]. In this dissertation, the percentage of ions annotated to candidate metabolites is much higher than that in the case of a plant reported by Nakamura et al. (2007) (10% of peaks in *Arabidopsis thaliana*).

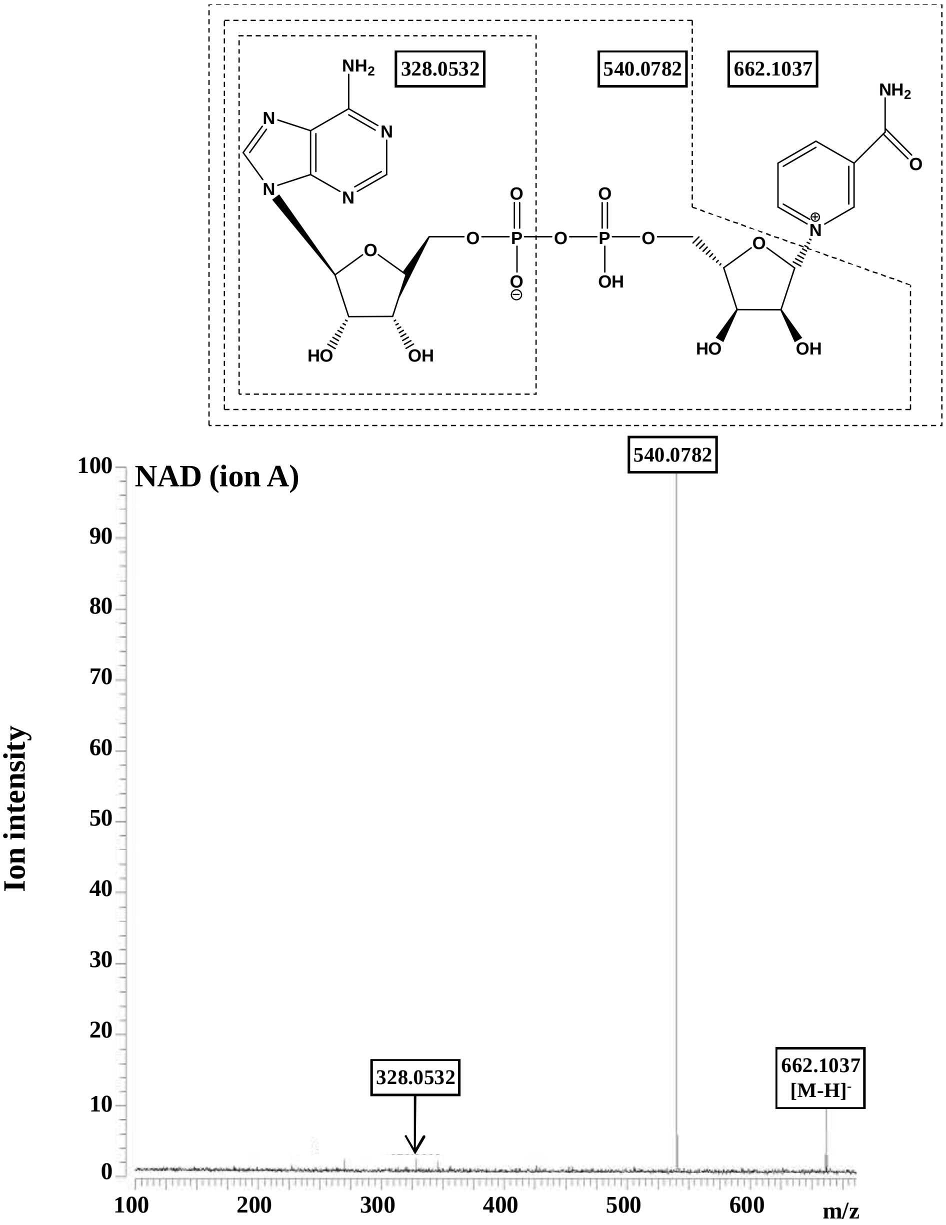


Figure 3.3. MS/MS analysis of nicotinamide adenine dinucleotide (NAD) ion with *m/z* = 662.1037 in the negative ion mode analysis. [M-H]$^-$ corresponds to the detected ion.

Table 3.1: Summary of reduction of candidates using the isotope pattern in ions in MS/MS analyses

Ion ID	Cluster ID	Monoisotope (M)	Isotope (M+1)	Difference	Number of candidates±0.01	Estimated carbon number	Number of estimated candidates	Element ratio	Candidate	Actual number of carbon atoms
Ion A	6	662.1037	663.1080	1.0044	874	19-21	92	90	NAD	21
Ion B	1	719.4868	720.4917	1.0048	146	36-40	33	33	PG1	38
Ion C	2	733.5056	734.5087	1.0032	167	38-44	34	34	PG2	39
Ion D	1	747.5183	748.5227	1.0044	175	39-40	12	12	PG3	40
Ion E	2	761.5293	762.5340	1.0047	219	28-60	102	102	PG4	41

M in ‘monoisotope’ column corresponds to $[M-H]^-$ in the negative ion mode analysis.

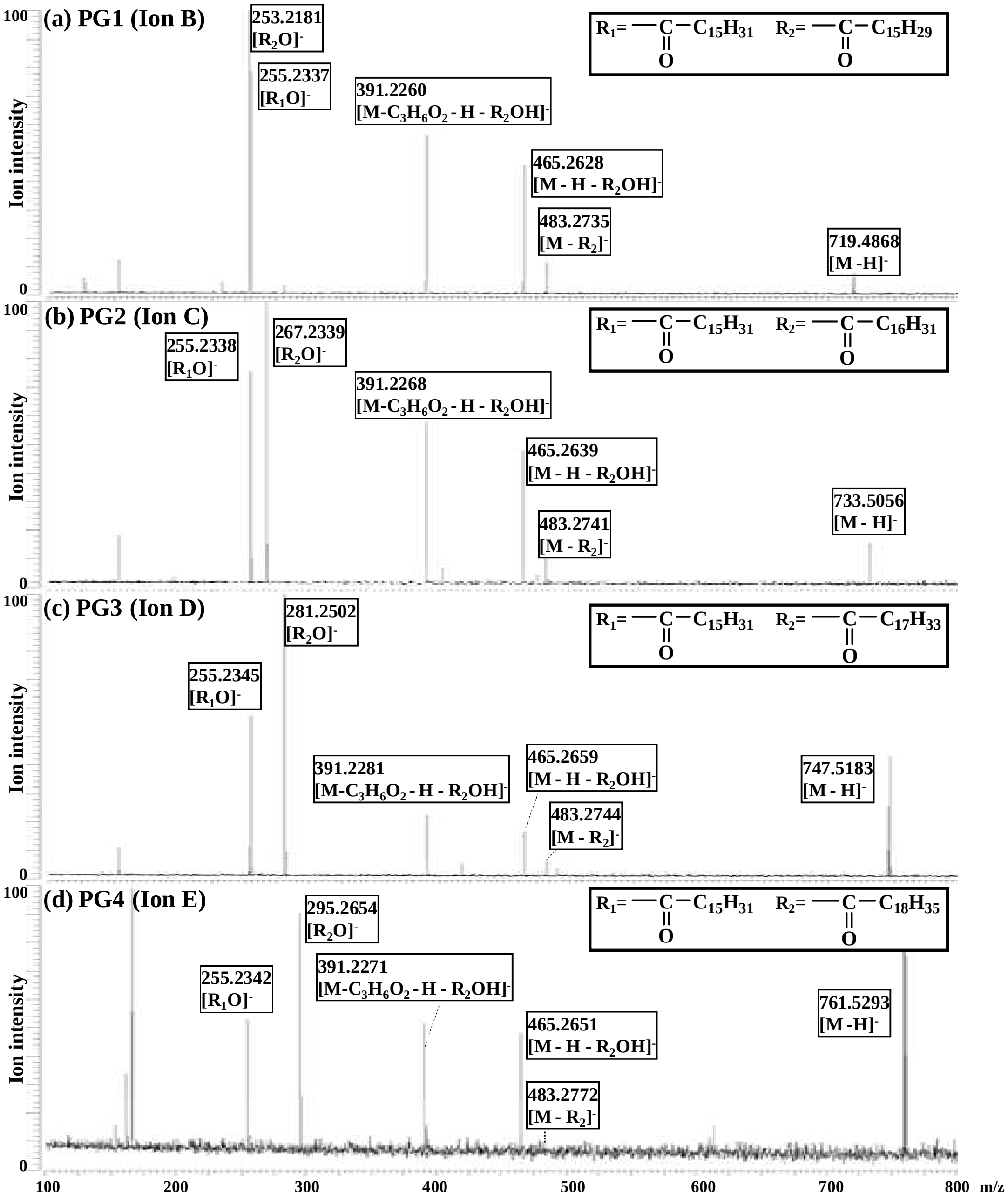


Figure 3.4. MS/MS analyses of the four ions in the negative ion mode analysis. $[M\text{-}H]^-$ corresponds to the detected ion. **a-d** Fragmentation patterns of phosphatidylglycerols 1-4 (PG1-PG4) ions with *m/z* = 719.4868, *m/z* = 733.5056, *m/z* = 747.5183, and *m/z* = 761.5293. R_1 and R_2 correspond to fatty acids.

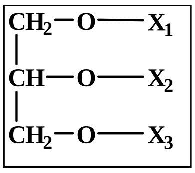
ID	Combination of three substructures (X_1, X_2, X_3)			
PG1	$-C(=O)-C_{15}H_{29}$	$-C(=O)-C_{15}H_{31}$	$-P(=O)(OH)-O-CH_2-CHOH-CH_2OH$	
PG2	$-C(=O)-C_{16}H_{31}$			
PG3	$-C(=O)-C_{17}H_{33}$			
PG4	$-C(=O)-C_{18}H_{35}$			

Figure 3.5. Molecular structures of PG1-PG4 determined by MS/MS analyses. Chemical structures in *left*, *middle*, and *right* columns correspond to substructure X_1, X_2, and X_3 of phosphatidylglycerols, respectively.

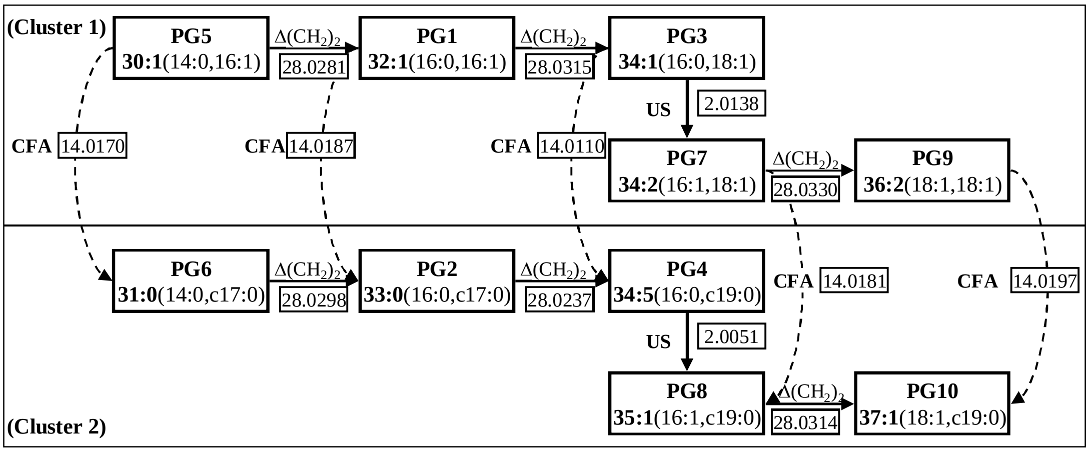


Figure 3.6. Relation of mass differences among PG1 to PG10. PG *xx:y* head groups, *xx* total number of carbons in the fatty acid chains, *y* number of double bonds; *c* cyclopropane; *CFA* cyclopropane fatty acid formation; *US* unsaturation. Theoretical $\Delta(CH_2)_2$, CFA and US are 28.0313, 14.0157, and 2.0157, respectively.

Table 3.2: Summary of candidates for ions based on KNApSAcK search using bacteria-metabolite relationships

Detected m/z	Theoretical m/z	Molecular formula	Exact mass	Error	Candidate	Species
72.9878	73.9951	$C_2H_2O_3$	74.0004	0.0053	Glyoxylic acid	*Escherichia coli*
143.1080	144.1153	$C_8H_{16}O_2$	144.1150	0.0003	Octanoic acid	*Escherichia coli*
253.2137	254.2210	$C_{16}H_{30}O_2$	254.2246	0.0036	omega-Cycloheptanenonanoic acid	*Alicyclobacillus acidocaldarius*
253.2185	254.2258	$C_{16}H_{30}O_2$	254.2246	0.0012	omega-Cycloheptanenonanoic acid	*Alicyclobacillus acidocaldarius*
281.2444	282.2516	$C_{18}H_{34}O_2$	282.2559	0.0042	Oleic acid	*Escherichia coli*
		$C_{18}H_{34}O_2$	282.2559	0.0042	cis-11-Octadecanoic acid	*Lactobacillus plantarum*
		$C_{18}H_{34}O_2$	282.2559	0.0042	omega-Cycloheptylundecanoic acid	*Alicyclobacillus acidocaldarius*
297.2410	298.2482	$C_{18}H_{34}O_3$	298.2508	0.0026	alpha-Cycloheptaneundecanoic acid	*Alicyclobacillus acidocaldarius*
297.2467	298.2540	$C_{18}H_{34}O_3$	298.2508	0.0032	alpha-Cycloheptaneundecanoic acid	*Alicyclobacillus acidocaldarius*
297.2516	298.2589	$C_{18}H_{34}O_3$	298.2508	0.0081	alpha-Cycloheptaneundecanoic acid	*Alicyclobacillus acidocaldarius*
321.0506	322.0579	$C_{10}H_{15}N_2O_8P$	322.0566	0.0013	dTMP	*Escherichia coli K12*
346.0570	347.0643	$C_{10}H_{14}N_5O_7P$	347.0631	0.0012	AMP	*Escherichia coli*
		$C_{10}H_{14}N_5O_7P$	347.0631	0.0012	3'-AMP	*Escherichia coli*
		$C_{10}H_{14}N_5O_7P$	347.0631	0.0012	dGMP	*Escherichia coli*
401.0168	402.0241	$C_{10}H_{16}N_2O_{11}P_2$	402.0229	0.0012	dTDP	*Escherichia coli*
402.9962	404.0035	$C_9H_{14}N_2O_{12}P_2$	404.0022	0.0013	UDP	*Escherichia coli*
426.0237	427.0310	$C_{10}H_{15}N_5O_{10}P_2$	427.0294	0.0016	Adenosine 3',5'-bisphosphate	*Escherichia coli*
		$C_{10}H_{15}N_5O_{10}P_2$	427.0294	0.0016	ADP	*Escherichia coli*
		$C_{10}H_{15}N_5O_{10}P_2$	427.0294	0.0016	dGDP	*Escherichia coli*
454.0391	455.0464	$C_{20}H_{19}Cl_2NO_7$	455.0539	0.0075	Antibiotic MI 178-34F18A2	*Actinomadura spiralis MI178-34F18*
		$C_{20}H_{19}Cl_2NO_7$	455.0539	0.0075	Antibiotic MI 178-34F18C2	*Actinomadura spiralis MI178-34F18*
458.1112	459.1185	$C_{15}H_{22}N_7O_8P$	459.1267	0.0083	Phosmidosine B	*Streptomyces sp. strain RK-16*
495.1039	496.1112	$C_{24}H_{20}N_2O_{10}$	496.1118	0.0006	Kinamycin A	*Streptomyces murayamaensis sp. nov.*
		$C_{24}H_{20}N_2O_{10}$	496.1118	0.0006	Kinamycin C	*Streptomyces murayamaensis sp. nov.*
505.9908	506.9981	$C_{10}H_{16}N_5O_{13}P_3$	506.9957	0.0023	ATP, dGTP	*Escherichia coli*
547.0756	548.0829	$C_{16}H_{26}N_2O_{15}P_2$	548.0808	0.0020	dTDP-L-rhamnose	*Escherichia coli*
565.0503	566.0576	$C_{15}H_{24}N_2O_{17}P_2$	566.0550	0.0025	UDP-D-glucose	*Escherichia coli*
		$C_{15}H_{24}N_2O_{17}P_2$	566.0550	0.0025	UDP-D-galactose	*Escherichia coli*
606.0775	607.0848	$C_{17}H_{27}N_3O_{17}P_2$	607.0816	0.0032	UDP-N-acetyl-D-mannosamine	*Escherichia coli*
		$C_{17}H_{27}N_3O_{17}P_2$	607.0816	0.0032	UDP-N-acetyl-D-glucosamine	*Escherichia coli*
618.0897	619.0970	$C_{17}H_{27}N_5O_{16}P_2$	619.0928	0.0042	ADP-L-glycero-beta-D-manno-heptopyranose	*Escherichia coli*
662.1037	663.1109	$C_{21}H_{27}N_7O_{14}P_2$	663.1091	0.0018	NAD	*Escherichia coli*
664.1095	665.1168	$C_{21}H_{29}N_7O_{14}P_2$	665.1248	0.0080	NADH	*Escherichia coli*
741.4729	742.4801	$C_{32}H_{62}N_{12}O_8$	742.4814	0.0012	Argimicin A	*Sphingomonas sp.*
786.4712	787.4785	$C_{41}H_{65}N_5O_{10}$	787.4731	0.0054	BE 32030B	*Nocardia sp. A32030*
853.3166	854.3239	$C_{41}H_{46}N_{10}O_9S$	854.3170	0.0069	Argyrin G	*Archangium gephyra Ar 8082*
		$C_{45}H_{56}Cl_2N_2O_{10}$	854.3312	0.0073	Decatromicin B	*Actinomadura sp. MK73-NF4*
		$C_{39}H_{50}N_8O_{12}S$	854.3269	0.0030	Napsamycin C	*Streptomyces sp. HIL Y-82,11372*

The column of 'Detected *m/z*' corresponds to the $[M\text{-}H]^-$ ion in the negative ion mode analysis

3.3.4 Cellular conditions assessed according to metabolite composition

In this section, I describe growth stage specificity for annotated metabolites. Figure 3.7 shows (a) the growth curve, (b) the number of ions detected at each time point, and (c) expression profiles of metabolites in cluster 1-5. The number of ions detected in each cluster decreases toward T6 and thereafter increases toward T8, suggesting that after the exponential phase, composition of metabolites in *E. coli* would be largely changed at T6.

Ions in cluster 5 and 3 correspond to ion accumulation in T2 and T3 at the exponential phase (Fig. 3.7c), respectively, suggesting that these metabolites would be necessary only at certain cell states. A candidate for the ion with *m/z* = 281.2444 in cluster 5 obtained by KNApSAcK searching is oleic acid (M-12 in Fig. 3.2; error of *m/z* = 0.0042) which is a precursor of phospholipids and has one double bond, suggesting that biosynthesis of fatty acid with double bond might occur in the exponential but not stationary phase, and other ions in cluster 5 would be intermediate compounds in a pathway related to fatty acid biosynthesis.

Candidates for the ion with *m/z* = 565.0503 (M-6) in cluster 3 are UDP-D-glucose and UDP-D-galactose. Candidates for the ion with *m/z* = 606.0775 (M-7) are UDP-*N*-acetyl-D-mannosamine and UDP-*N*-acetyl-D-glucosamine, which are precursors of lipopolysaccharides (LPS) [Vimr et al. 2004], suggesting that LPS biosynthesis would occur only in the exponential phase and relate to abundances of UDP-D-glucose and UDP-D-galactose, and other ions in cluster 3 would be compounds related to LPS biosynthesis. A candidate for the ion with *m/z* = 143.1080 in cluster 3 is octanoic acid

(M-4), which is the direct precursor of a vitamin and lipoic acid, and is also an exponential phase specific metabolite. *E. coli* contains a pool of octanoic acid which can act as a substrate for lipoate ligase during lipoate starvation of a lipoic acid auxotroph [Ali et al. 1990]. The accumulation of octanoic acid at stage T3 would be needed in the exponential phase to prepare biosynthesis of vitamins. Ions in cluster 4 correspond to ion accumulation in T7 at the stationary phase (Fig. 3.7c), suggesting that ions in cluster 4 would be compounds related to the stationary phase.

According to profiles in Figure 3.7c, clusters 1 and 2 are exponential and stationary phase specific, respectively. It is well known that phospholipid production decreases dramatically at the stringent response [Merlie and Pizer 1973; Polakis et al. 1973], and the bulk of CFA synthesis occurs as cultures enter the stationary phase of growth [Magnuson et al. 1993]. Those facts are consistent with the structures of PG2, PG4, PG6, PG8, and PG10 in cluster 2 being CFA forms of PG1, PG3, PG5, PG7, and PG9 in cluster 1, respectively. In addition to this, CFA synthesis occurs in a broad range of phosphatidylglycerols after T5. Thus, cellular conditions of *E. coli* could be explained in terms of the composition of metabolites.

Unsupervised learning such as PCA and BL-SOM makes it possible to examine metabolic phenotyping of seedlings treated with different herbicidal chemical classes for pathway-specific inhibitions [Oikawa et al. 2006] and accurate classification of genes based on time series expression profiles which led to the prediction of gene functions [Hirai et al. 2005; Hirai et al. 2004; Yano et al. 2006]. Figure 3.8 shows the PCA projection of measurement points in time series data. The proportions, that is, percent variances to total variance, are 94.3% and 2.4% for the first and second principal components (PC1 and PC2), respectively. So the first two principal components, which can explain 96.7% of total variance, are enough to examine the differences in eight time points. The distribution of eight time points in the first two PCs as shown in Figure 3.8 implies that time points are clearly classified into two groups, an early group consisting of T1, T2, T3, T4, and T5, and a late group consisting of T6, T7, and T8, suggesting that the different growth stages could be represented by the metabolomics data. The former and latter roughly correspond to exponential and stationary phases in the growth curve of *E. coli*. This result shows that the metabolite profile in *E. coli* seems to be totally shifted from T5 to T6, which is also consistent with the transient point in the number of detected ions in Figure 3.7b.

To directly relate composition of metabolites to cellular conditions, I applied partial least squares (PLS) regression to the metabolite profiling data. PLS regression provides a quantitative model to estimate the cellular conditions based on the composition of metabolites. So in this dissertation, I focused on the PLS model to estimate cellular conditions from exponential to stationary phase based on intensities of *m/z* values in FT-ICR/MS and examined quantitative differences of metabolites using the PLS model. Growth of bacteria can be generally monitored by measuring the optical density at 600 nm (OD_{600}). A linear model for estimating the OD_{600} values according to the metabolite quantities in individual time points provides the useful information associated with quantitative differences of the metabolites between exponential and stationary phases. To attain this, I conducted PLS regression, which is applicable when the number of independent variables is very large compared with the number of samples. Using PLS regression, the OD_{600} value can be directly estimated from the corresponding intensity vector of *m/z* values, as follows:

$$\mathrm{OD}_{600} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_j x_j + \cdots a_m x_m \qquad (3.1)$$

, where x_j and a_j represent the intensity and coefficient for *j*th ion.

When the ion has a positive PLS regression coefficient, its ion’s level should increase from exponential to stationary phase because the optical density is saturated in the highest level of the growth curve. As shown in Figure 3.9, I got the best linear model in PLS regression with one component (R_{pred} = 0.94) as described in Equation (2.4). The Pearson’s correlation coefficient between the observed and predicted OD_{600} values is *r* = 0.97, suggesting that constructed model would work well, and is informative to clarify the relation between a growth stage and metabolite profile.

Next, I plotted the regression coefficients of each ion determined by using the proposed model in order to elucidate which metabolite is important for estimating the OD_{600} values (Fig. 3.10). The ions with negative and positive coefficients contribute to the constructed model, negatively and positively, and are dominant in exponential and stationary phase, respectively. Four ions (PG1, *m/z* = 719.4868; PG2, *m/z* = 733.5056; PG3, *m/z* = 747.5183; PG4, *m/z* = 761.5293) which were analyzed by MS/MS analysis as described above had the highest coefficients. Other annotated six ions (PG5, *m/z* = 691.4588; PG6, *m/z* = 705.4747; PG7, *m/z* = 745.5045; PG8, *m/z* = 759.5242; PG9, *m/z* = 773.5375; PG10, *m/z* = 787.5556) also had higher coefficients, suggesting that PLS analysis could extract stage-specific metabolites efficiently. Thus, the observed behavior of metabolites is highly reflected in the regression coefficients of the PLS model and the interpretation of the coefficients is fairly consistent with the transition of metabolites from exponential to stationary phase.

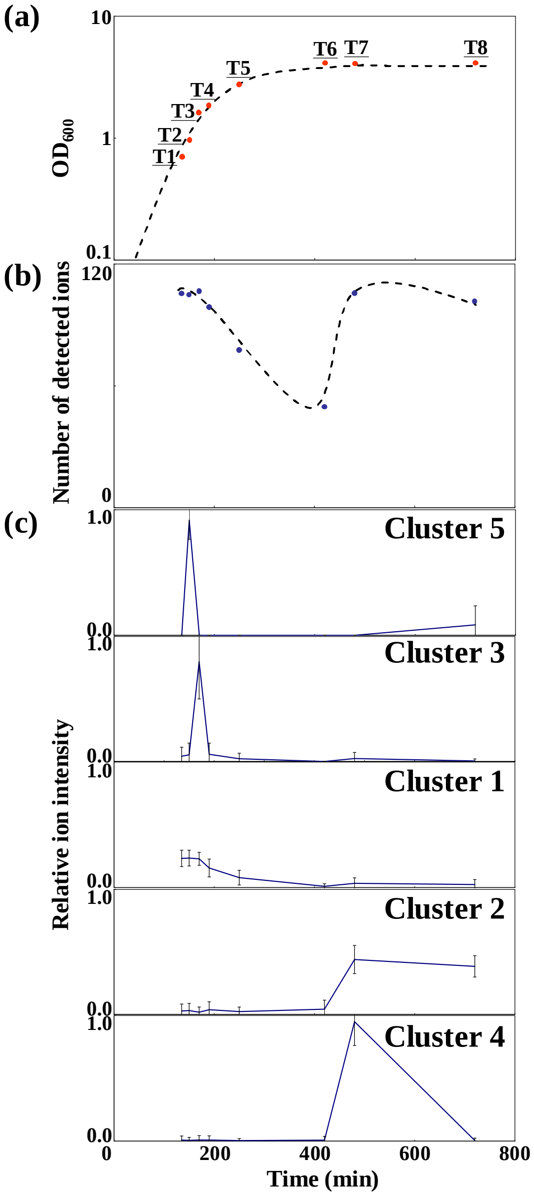


Figure 3.7. The time series profiles. **a** Growth curve. **b** Time series change of total number of detected ions at each time point. **c** Average expression profiles of ions in cluster 1-5. Error bars show standard deviation at each time point.

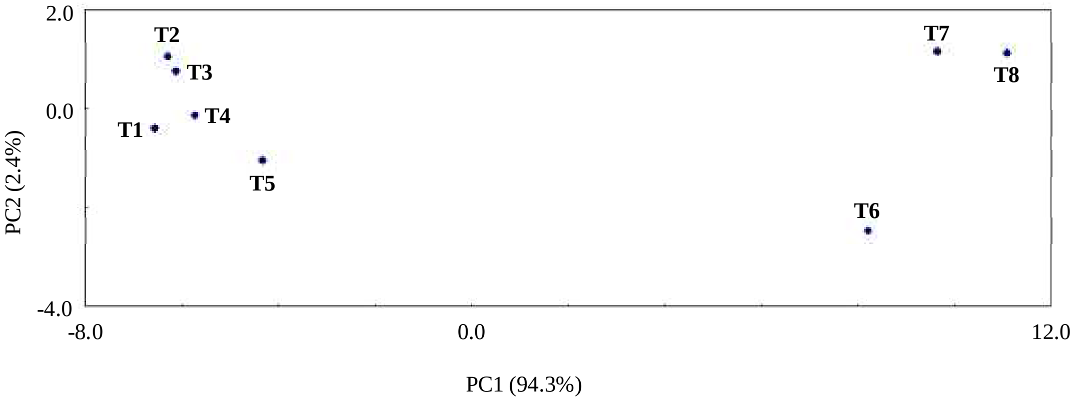


Figure 3.8. PCA analysis. Plot of eight time points are shown by using the first two PCs.

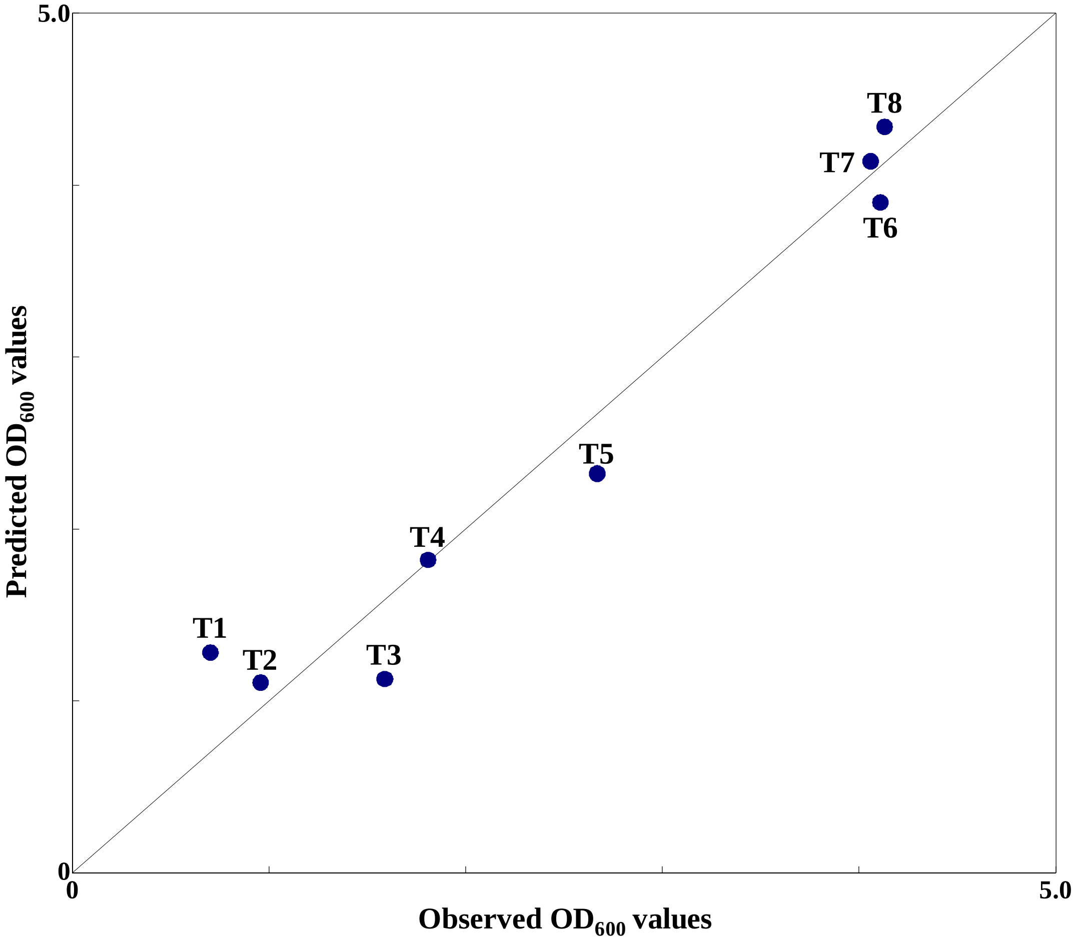


Figure 3.9. Predicted OD_{600} values by PLS analysis. Observed and predicted optical densities are based on the PLS model with the first component.

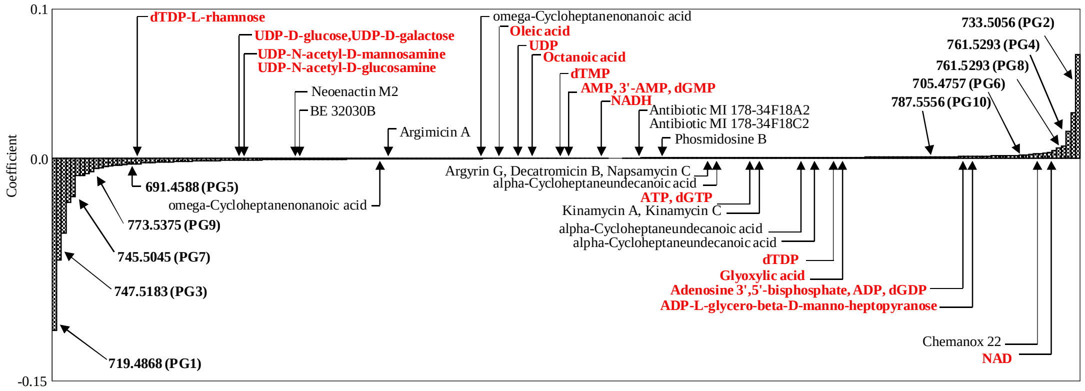


Figure 3.10. Intensity of regression coefficients. The metabolites written in *red* are reported metabolites in *E. coli*. The metabolites written in *black* are reported metabolites in other bacteria species.

Chapter 4
Integrative analysis of transcriptomics and metabolomics

4.1 Introduction

In the post-genomics era, a systematic and comprehensive understanding of the complex events in the organisms, e.g. the majority of gene products function not alone, but with other gene products, is a great concern in biology. Now, ‘omics’ approaches, i.e. genomics, transcriptomics, proteomics, and metabolomics, are required in order to understand organisms as a system. New advanced methods, strategies, and technologies for ‘omics’ studies should be mainly directed to the elucidation of regulation, and gene regulatory networks responsible for specific phenotypes at the different levels: genomics, transcriptomics, proteomics, and metabolomics in an integrative or systems biology perspectives [Castrillo and Oliver 2004]. Metabolomics, hence, offers insights into metabolism that complements information obtained from proteomics and transcriptomics [Fridman and Pichersky 2005] and has a potential to elucidate gene functions and networks, especially when integrated with transcriptomics. A promising approach is pair-wise gene-to-metabolite correlation analysis, which can reveal unexpected correlations and shed light on candidate genes for regulating the metabolite content. The systematic integration of transcriptomics, proteomics, and metabolomics facilitates the unbiased, information-based reconstruction of underlying biochemical networks [Hirai et al. 2004; Hirai et al. 2005; Urbanczyk-Wochniak et al. 2003; Fiehn et al. 2001]. Pir et al. (2006) integrated metabolomics with transcriptomics by using PLS modeling, and metabolite data were modeled as a function of the transcriptome to determine their congruence.

While clustering techniques have been applied to identify co-expressed genes in time series microarray analysis, several papers propose methods that could detect time lagged relationships of gene expression profiles [Balasubramaniyan et al. 2005; Ji and Tan 2005; Redestig et al. 2007]. That is, gene products regulate each other not only simultaneously but also after a certain time lag. It is necessary for taking into consideration a time lag between several ‘omics’ data. Meanwhile, investigating the responses of cells to environmental changes typically requires a system-level analysis. A key step to analyze system responses to environmental changes is identifying large state changes or ‘transitions’. Morioka et al. (2007) developed the statistical method, Linear Dynamical System (LDS), which uses internal state variables in the generative model for cellular internal state changes, and detects cellular state transitions in time series data.

Heretofore, I have established non-targeted metabolomics approach based on Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS), and showed that metabolomics data could explain cell state in *E. coli*. In this chapter, by adding time series cDNA microarray experimental data of *E. coli*, I present an integrated analysis of transcriptomics and metabolomics data, taking into consideration a time lag between transcriptomics and metabolomics in a non-targeted manner. LDS method was applied to estimate the transition point in transcriptional and metabolite levels, whether or not there was a time lag between transcriptional and metabolite levels. I examined gene-to-gene correlation analysis to determine how tightly genes on each pathway would be regulated by each other using cDNA microarray data. Furthermore, gene-to-metabolite correlation analysis was also performed, taking into consideration a time lag between transcriptomics and metabolomics data. Here I show that taking into consideration a time lag between transcriptomics and metabolomics enables integrated analysis to elucidate gene-to-metabolite networks in time series context in a more effective way.

4.2 Materials and methods

4.2.1 RNA extraction and cDNA synthesis

Cells were collected by centrifugation at 135, 150, 170, 190, 250, 420, 480, and 720 min postinoculation (which correspond to T1, T2, T3, T4, T5, T6, T7, and T8) after adding RNA protect (Qiagen) and stored –80℃, while control sample was collected at 130 min. Total RNA was isolated using the RNeasy mini kit and RNase-free DNase set according to the manufacture's instructions (Qiagen). For each labeling reaction, a total of 15 μg of RNA was used. First-strand cDNA synthesis was primed with 1.2 μg random primers (Invitrogen) in nuclease-free water (total volume: 31 μl) by heating at 70℃ for 10 min and incubating at 25℃ for an additional 10 min. Reverse transcription was performed by SuperScript III (Invitrogen) in reverse transcription buffer [1 × first-strand buffer, 10 mM DTT] in the presence of 5 mM dATP, 5 mM dUTP, 5 mM dCTP, 0.25 mM dTTP, and 0.25 mM AA-dUTP. Three amino-allyl-labeled nucleotides were incorporated into the cDNA. The reactions were incubated at 25℃ for 10 min, 37℃ for 60 min, 42℃ overnight, and quenched by heating at 70℃ for 10 min.

The RNA template was hydrolyzed by adding 20 μl of 1N NaOH followed by heating at 65℃ for 30 min. Reactions were neutralized with 20 μl of 1N HCl. cDNA was purified using a CyScribe GFX Purification Kit (GE Healthcare) according to the manufacturer's directions. NHS ester forms of Cy3 and Cy5 dyes were added to the cDNA solution and incubated for 4 h. Coupling reactions were quenched by the addition of 15 μl of 4 M hydroxylamine and incubated at room temperature for 15 min in the dark. Labeled cDNA was purified using the CyScribe GFX Purification Kit again.

4.2.2 Hybridization and spot detection

Prehybridization of the array slides was performed for 3 hr in filtered prehybridization solution [25% formamide, 5 × SSC, 10 mg BSA (fraction V), 0.1% SDS] at 42℃. Slides were briefly washed in milliQ water and 80% ethanol and dried by centrifugation at 1,000 g for 5 min. Hybridization of the probe was performed using hybridization solution (25% formamide, 5 × SSC, 0.1% SDS, 0.1 μg poly (A), 1 × Denhardt's solution and 100 pmol Cy3 and Cy5 combined probe). The hybridization solution containing the Cy-dye-labeled cDNA was heated to 95℃ for 3 min and hybridization was performed in an Advalytix hybridization machine (ArrayBooster) at 42℃ for 16 h. After hybridization, the slides were washed and dried by centrifugation at 1,000 g for 5 min and then analyzed using a Fuji FLA-8000 scanner and Array Gauge ver.2.0 software (Fuji Film).

4.2.3 Transcriptomics and metabolomics data set

After normalization of 16 sets of microarray data (twice for each of eight time points), the log ratio corresponding to each gene was averaged, and then genes with one or more missing values were removed. The remaining 3,945 genes were used for estimation of transition points in transcriptional levels. Finally, 1,162 genes were selected for which at least at one time point the expression value is more than or equal to the threshold $Mean \pm 1.5 \times SD$ determined in the context of all time point data of 3,945 genes. These highly expressed genes were used for gene-to-metabolite correlation analysis. Metabolite expression profiles consisted of 220 peaks, which were used for estimation of transition points in metabolite levels, detected by Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR/MS). Out of these, there were 174 metabolite derivative groups. To use in gene-to-metabolite correlation analysis, I determined time lagged data using linear interpolation as follows: in case of any metabolite, say m_j, metabolite quantities are measured for s time series points which are denoted by m_j (T_1), m_j (T_2), …, m_j (T_s), and let i = 1, 2, …, s. The quantity of *j*th metabolite at time *t*, m_j (t), was calculated using following equations.

In the case that *t* is in the interval between T_i and T_{i+1},

$$m_j(t) = m_j(T_i) + \frac{\{t - T_i\}\{m_j(T_{i+1}) - m_j(T_i)\}}{T_{i+1} - T_i}. \tag{4.1}$$

In the case that *t* is outside of the largest sampling point T_s $(t > T_s)$,

$$m_j(t) = m_j(T_{S-1}) + \frac{\{t - T_{S-1}\}\{m_j(T_S) - m_j(T_{S-1})\}}{T_S - T_{S-1}}. \tag{4.2}$$

Figure 4.1 shows an example of 30 minute time lag points for the ion with *m/z* = 719. 4868

(PG1) corresponding to, 165 (= 135+30), 180 (= 150+30), 200 (= 170+30), 220 (= 190+30), 280 (= 250+30), 450 (= 420+30), and 510 (= 480+30), and 750 (= 720+30) min for the eight reference measurements at times, 135, 150, 170, 190, 250, 420, 480, and 720 min. For the sampling point T1 (corresponding to 135 min), thirty minute lagged point is at 165 min (indicated by red triangle) between T2 (150 min) and T3 (170 min), and corresponds to the point on the line connecting T2 and T3. Seven lagged points corresponding to original points except for T8 were calculated as described in Equation (4.1). The lagged point corresponding to T8 was determined by expanding linearly the line connecting T7 and T8 from 720 min (T8) to 750 min (as shown by dotted line in Fig. 4.1). Here, I prepared nine sets of time lagged data of metabolite expression profiles, i.e. 10 min, 20 min, 30 min, 40 min, 50 min, 60 min, 70 min, 80 min, and 90 min lagged data.

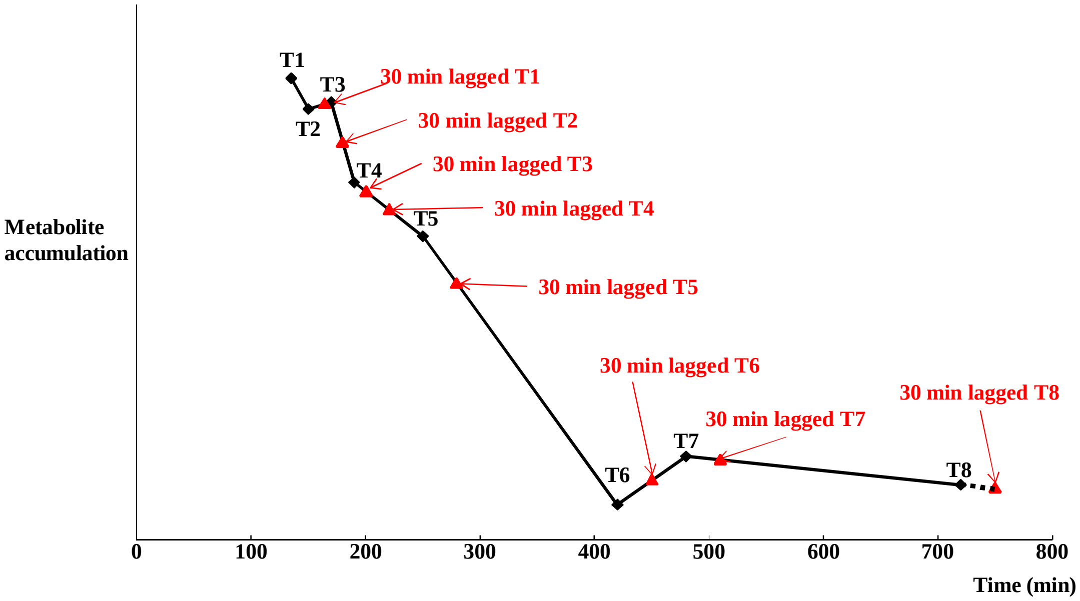


Figure 4.1. Preparing metabolomics time lagged data. Black line corresponds to one original metabolite accumulation profile. Red triangles correspond to 30 min lagged T1, T2, T3, T4, T5, T6, T7, and T8.

4.2.4 Transition points estimation by Linear Dynamical System (LDS)

LDS uses internal state variables in the generative model for cellular internal state changes. These internal states correspond to the compressed description of the observed biological system prior to adding noise factors. Observational 'omics' data is defined as $Y_{1:T} = \{\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_T}\}$. Internal state for each observational vector is defined as $X_{1:T} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_T}\}$. The proposed model is defined as follows:

Observational equation: $\mathbf{y_t} = V\mathbf{x_t} + \boldsymbol{\eta}_\mathbf{t}$ (4.3)

Transition equation: $\mathbf{x_t} = W\mathbf{x_{t-1}} + \boldsymbol{\varepsilon}_\mathbf{t}$ (4.4)

,where $t = 1, 2, \ldots, T$ is the measurement order of the time series, V is a $D \times N$ observational matrix in which D is the number of genes or metabolites, and N is the dimension of internal states, W is an $N \times N$ internal state transition matrix, D-dimensional vector $\boldsymbol{\eta}_\mathbf{t}$ is a observational noise, and N-dimensional vector $\boldsymbol{\varepsilon}_\mathbf{t}$ is a transition noise. The vectors $\mathbf{x_1}$, $\boldsymbol{\varepsilon}_\mathbf{t}$, and $\boldsymbol{\eta}_\mathbf{t}$ are generated according to the following equations:

$$\mathbf{x_1} \sim N_N\left(\mathbf{x_1} \left| \mu_1, \sigma_1^2 I_N\right.\right) \tag{4.5}$$

$$\boldsymbol{\varepsilon}_\mathbf{t} \sim N_N\left(\boldsymbol{\varepsilon}_\mathbf{t} \left| 0_N, \sigma_\varepsilon^2 I_N\right.\right) \tag{4.6}$$

$$\boldsymbol{\eta}_\mathbf{t} \sim N_D\left(\boldsymbol{\eta}_\mathbf{t} \left| 0_D, \sigma_\eta^2 I_D\right.\right). \tag{4.7}$$

$N_p\left(\mathbf{x} \left| \mathbf{m}, \Sigma\right.\right)$ is a probabilistic density function when p dimensional probabilistic vector

$\mathbf{x}$ obeys a normal distribution whose mean vector is $\mathbf{m}$, and covariance matrix Σ is as follows:

$$N_p(\mathbf{x}\,|\mathbf{m},\Sigma) \equiv (2\pi)^{-1/2}|\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})'\Sigma(\mathbf{x}-\mathbf{m})\right]. \tag{4.8}$$

I assume that the observational and internal transition noises are both Gaussian, and therefore the relationship is a first-order Markov process as follows:

$$p(\mathbf{x_t},\mathbf{y_t}\,|X_{1:t-1},Y_{1:t-1}) = p(\mathbf{y_t}\,|\mathbf{x_t})p(\mathbf{x_t}\,|\mathbf{x_{t-1}}). \tag{4.9}$$

The model parameter of (4.3)-(4.7) is defined as the parameter set θ as follows:

$$\theta = \{\mu_1, \sigma_1, W, \sigma_\varepsilon, V, \sigma_\eta\}. \tag{4.10}$$

Note that the model corresponds to a Kalman Filter when θ is known [Kalman and Bucy 1961]. The initial state $\mathbf{x_1}$ is defined as:

$$p(\mathbf{x_1}\,|\theta) = N_N(\mathbf{x_1}\,|\mu_1, \sigma_1^2 I_N). \tag{4.11}$$

From Equations (4.3) and (4.5), the following function is obtained:

$$p(\mathbf{x_t}\,|\mathbf{x_{t-1}},\theta) = N(\mathbf{x_t}\,|W\mathbf{x_{t-1}}, \sigma_\varepsilon^2 I_N). \tag{4.12}$$

From Equations (4.4) and (4.6), the following function is obtained:

$$p(\mathbf{y_t}\,|\mathbf{x_t},\theta) = N_D(\mathbf{y_t}\,|V\mathbf{x_t}, \sigma_\eta^2 I_D). \tag{4.13}$$

Using these results, the following joint probability is obtained:

$$p(Y_{1:T}, X_{1:T}\,|\theta) = N(\mathbf{x_1}\,|\theta)\left\{\prod_{t=2}^{T} p(\mathbf{x_t}\,|\mathbf{x_{t-1}},\theta)\right\}\left\{\prod_{t=1}^{T} p(\mathbf{y_t},|\mathbf{x_t},\theta)\right\}. \tag{4.14}$$

The parameter optimization follows a standard EM algorithm. Using the resulting estimated parameters, the log-likelihood with respect to the present time point t when all time points are given, is defined by Equation (4.15):

$$\log L_t = \log p\left(\mathbf{y_t} \left| Y_{1:t-1}, \theta\right.\right). \tag{4.15}$$

4.2.5 Gene-to-metabolite correlation network functional analysis

All of the gene-to-metabolite networks were constructed based on Pearson correlation coefficient (PCC) $r \geq 0.9$. Genes were functionally categorized using their Gene Ontology information with respect to ‘biological process’ [Ashburner et al. 2000], and overrepresented GO terms were identified with Fisher’s exact test. The one-tailed Fisher’s exact p-value corresponding to overrepresentation of categories have been calculated based on counts in 2 × 2 contingency tables. Counts n_{11}, n_{12}, n_{21}, and n_{22} in the contingency table refer to n_{11}, number of observations of a particular category in the first gene set; n_{12}, number of other categories in the first gene set; n_{21}, number of observations of a category in second gene set; and n_{22}, number of observations of other categories in the second gene set.

4.3 Results and discussion

4.3.1 Estimation of transition points based on Linear Dynamical System (LDS) by using transcriptomics and metabolomics data

RNA was extracted and time series (eight time points) cDNA microarray experiments of *E. coli* were performed twice, as described in Materials and methods. Samples were collected at 135, 150, 170, 190, 250, 420, 480, and 720 min postinoculation (which correspond to T1, T2, T3, T4, T5, T6, T7, and T8 as shown in Fig. 4.2a). After normalizing time series data set as described in Materials and methods, I got the expression profiles of 3,945 genes.

A key step to analyze system responses to environmental changes is identifying large state changes or ‘transitions’. In time series analysis, estimation of transition point is important for understanding living cells as biochemical systems in several stages, i.e. genomics, transcriptomics, and metabolomics. Morioka et al. (2007) developed the statistical method, Linear Dynamical System (LDS) for estimation of state transition using transcriptomics and metabolomics data. In order to compare transition points detected in comprehensive transcriptional and metabolite levels, I applied this method to microarray and metabolomics data, which was performed by using FT-ICR/MS as described in Chapter 3. I calculated the log-likelihood values for gene expression and metabolite accumulation profiles, respectively. Figure 4.2a shows the log-likelihood values calculated by using gene expression and metabolite accumulation profiles, which are indicated by red and blue curves, respectively. ‘Likelihood values’, here, means the generative probability of current data based on the condition of the past data. If this value is low, then the current data cannot be adequately explained by past data, in other words, a transition has occurred. The lowest log-likelihood values are at T4 corresponding to 190 min postinoculation in transcriptomics (-5.4×10^{-3}), and at T5 corresponding to 250 min in metabolomics data (-5.2×10^{-3}) as shown in Figure 4.2a, suggesting that transition points predicted by transcriptomics and metabolomics data are different, i.e. time lag, and transition could occur at transcriptional levels, followed by at metabolite levels according to the central dogma. The number of significantly expressed genes at each time point gradually decreases along the growth curve (Fig. 4.2b). In this case, the threshold was set to be $Mean + 1.5 \times SD$ for Cy5 intensity values after normalization. Probe intensity values were used, because mRNA abundances at each time point rather than profile changes through cell growth could affect transition points of the whole cell. The number of genes with significantly abundant mRNA starts to decrease from time point T4, indicating that this result coincides with transition point predicted by LDS analysis of transcriptomics data.

Next, in order to elucidate whether or not genes with individual metabolic pathways would be robustly coregulated, co-expression relations in time series among genes in the same functional category were examined by Pearson correlation coefficients (PCCs). Functional categories concerning individual pathways were defined based on KEGG pathways [Kanehisa et al 2008], in which there were 127 pathways with respect to *E. coli* K-12 MG1655. Out of these pathways, I used 75 pathways annotated with more than ten genes and calculated all gene-to-gene PCCs within genes of individual functional categories. Figure 4.3 shows the boxplots of PCCs for genes of 75 individual pathways. The pathways of ribosome (ribosomal proteins), fatty acid biosynthesis, and aminoacyl-tRNA biosynthesis are clearly different from other pathways. Appendix B lists genes used for individual 75 KEGG pathways.

In addition to boxplots, Figure 4.4 shows the relation between median and standard deviation of PCCs for individual categories. Standard deviations of gene-to-gene PCCs corresponding to ribosomal proteins (55 genes, i.e. 1,485 PCCs), fatty acid biosynthesis (12 genes, i.e. 66 PCCs) are low, that is, genes classified into those two categories are highly co-expressed. In addition, genes within aminoacyl-tRNA biosynthesis (25 genes, i.e. 300 PCCs), are highly correlated, but the deviation is larger than those two categories. Medians of PCCs within the genes of other categories are relatively lower than those three categories. These results suggest that three pathways are highly coregulated. Ribosomal proteins and aminoacyl-tRNA biosynthesis pathways belong to translation based on KEGG, so genes with respect to translation are particularly highly regulated through cell growth. In addition, fatty acid biosynthesis pathway within lipid metabolism is also highly regulated depending on transcriptional level, suggesting that genes with respect to fatty acid biosynthesis might control comprehensive lipid metabolism. So I could find out candidate pathways by using KEGG pathway annotations, which were highly regulated depending on transcriptional levels, although each pathway is not a closed system on itself and there are complex interacting systems within a cell. Figure 4.2c shows the relative expression levels of genes involved in translation (ribosomal proteins, aminoacyl-tRNA biosynthesis), and fatty acid biosynthesis, which are decreased at the time point T4. On the other hand, the number of detected ions are transiently decreased at the time point T5 (Fig. 4.2d) and this also coincides with transition point (T5) estimated by LDS for metabolite accumulation profiles. In consequence, there is a time lag between transcriptomics and metabolomics data.

Furthermore, I tried to calculate gene-to-gene correlations within transcription factor (TF) regulated units and sigma factor gene regulated units based on RegulonDB [Gama-Castro et al. 2008] (Appendix C and D list genes used for 99 TF regulated units and 9 sigma factor regulated units). Figure 4.5 and 4.6 show boxplots of gene-to-gene PCCs of 99 TF regulated units, which are more than five genes in individual category, and boxplots of gene-to-gene PCCs of 9 sigma factor regulated units, respectively. As shown in Figure 4.5, median of gene-to-gene PCCs within ArgP regulated unit is more than 0.9, suggesting that genes regulated by this TF are highly coexpressed through cell growth in *E. coli.* ArgP, 'arginine protein', controls the transcription of genes involved in the arginine transport system and genes involved in DNA replication [Han et al. 1998], indicating that arginine transport and DNA replication could be highly regulated in transcription levels through cell growth. According to EcoCyc [Kaseler et al. 2005], the operons of nrdAp and dnaAp1 regulated by ArgP are regulated by several other factors, e.g. DnaA and Fis. So, this result suggests that several regulators regulate these operons in concert through cell growth in *E. coli*.

As shown in Figure 4.6, sigma19 (FecI) regulated genes are highly coexpressed. FecI causes expression of genes for uptake of ferric citrate [Visca et al. 2002]. This result indicates that genes involved in iron transport system are tightly regulated through cell growth in *E. coli*. Sigma38 (RpoS) regulates more than 100 genes involved in cell survival, cross protection against various stresses and in virulence [Venturi 2003]. Genes of sigma38 regulated units (3,003 PCCs by 78 genes used for boxplot) are positively regulated through cell growth, and so this result is consistent with the fact, that sigma38 functions in stationary phase in *E. coli*.

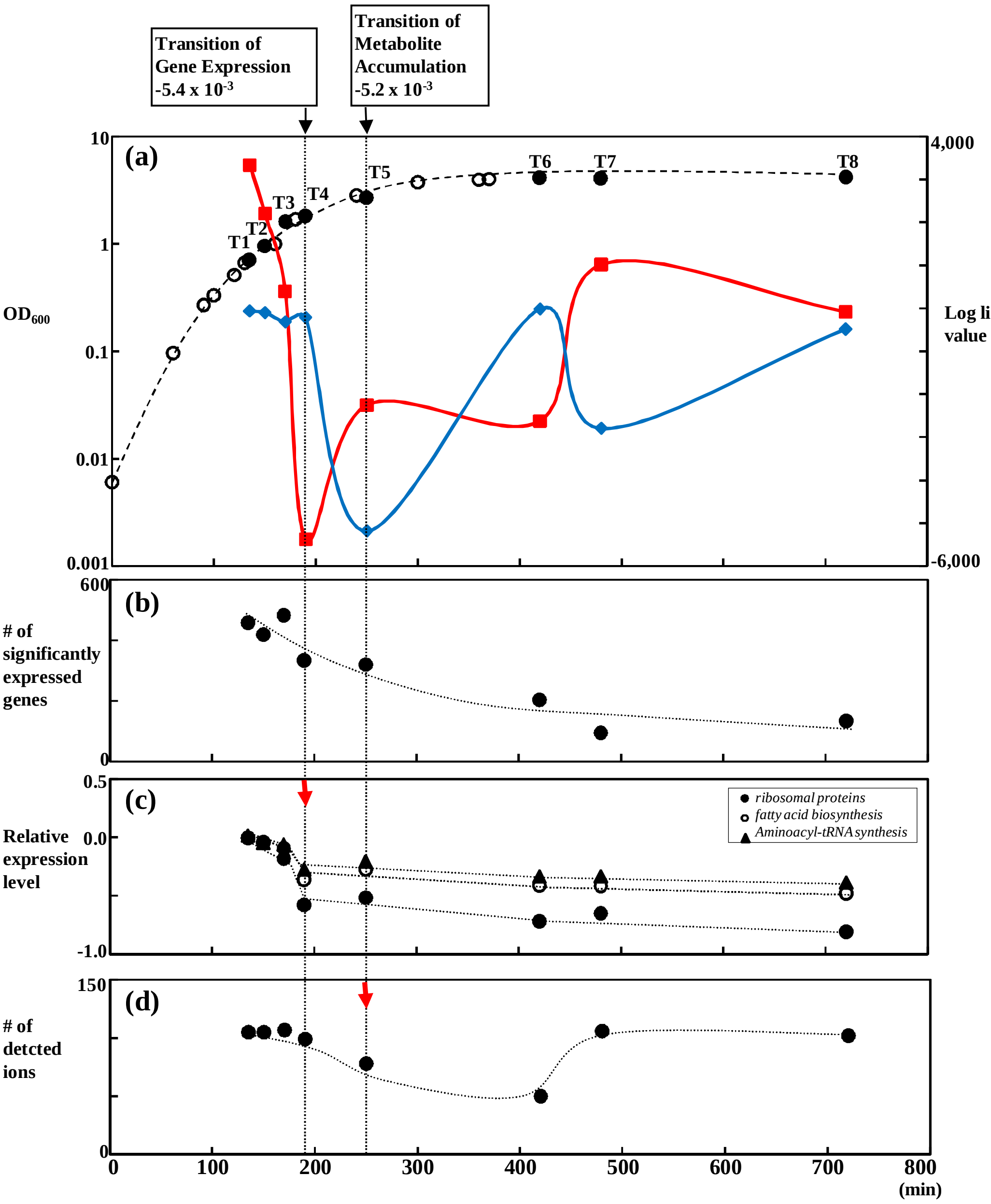


Figure 4.2. Transition point analysis. Estimated transition points by using gene expression and metabolite accumulation profiles are indicated by vertical dot lines through (a) to (d). **a** Log-likelihood values by LDS analysis. First and second axes correspond to OD_{600} values and log-likelihood values, respectively. Red and blue curves correspond to log-likelihood values calculated by LDS for gene expression and metabolite

accumulation profiles, respectively. Eight sampling points are indicated by black circles with T1, T2, T3, T4, T5, T6, T7, and T8. **b** The number of significantly expressed genes at each time point. **c** Expression profiles of three KEGG pathways. Ordinate axis corresponds to average expression values of genes within ribosomal proteins, fatty acid biosynthesis, and aminoacyl-tRNA synthesis of KEGG pathways, respectively. The red arrow shows the predicted transition point by using gene expression profiles. Black circles, white circles, and black triangles correspond to ribosomal proteins, fatty acid biosynthesis, and aminoacyl-tRNA synthesis, respectively. **d** The number of detected ions by FT-ICR/MS at each time point. The red arrow shows the predicted transition point by using metabolite accumulation profiles. Ordinate axis corresponds to the number of detected ions by FT-ICR/MS at each time point.

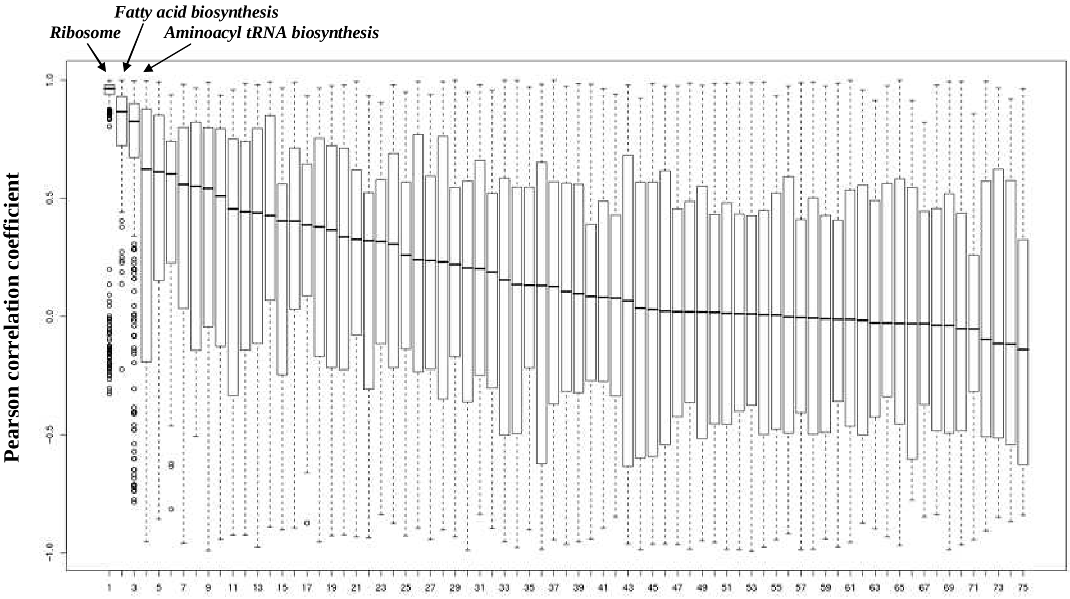


Figure 4.3. Boxplots of gene-to-gene PCCs of 75 KEGG pathways. Characterizations of each KEGG pathway based on Pearson correlation coefficients (PCCs) using gene expression profiles. The ordinate axis corresponds to the values of PCCs from minus one to plus one. The number below each boxplot corresponds to the following KEGG pathways: 1. Ribosome; 2. Fatty acid biosynthesis; 3. Aminoacyl-tRNA biosynthesis; 4. Oxidative phosphorylation; 5. Homologous recombination; 6. Fatty acid metabolism; 7. Folate biosynthesis; 8. Type III secretion system; 9. Pyrimidine metabolism; 10. One carbon pool by folate; 11. Histidine metabolism; 12. Sulfur metabolism; 13. Lipopolysaccharide biosynthesis; 14. Ubiquinone biosynthesis; 15. Lysine degradation; 16. Carbon fixation; 17. Valine, leucine and isoleucine degradation; 18. DNA replication; 19. Protein export; 20. Glycan structures - biosynthesis 2; 21. Benzoate degradation via CoA ligation; 22. Phenylalanine metabolism; 23. Selenoamino acid metabolism; 24. Type II secretion system; 25. Bacterial chemotaxis - General; 26. Reductive carboxylate cycle (CO2 fixation); 27. Phenylalanine, tyrosine and tryptophan biosynthesis; 28. Citrate cycle (TCA cycle); 29. Butanoate metabolism; 30. Bacterial chemotaxis - Organism-specific; 31. Flagellar assembly; 32. Cysteine metabolism; 33. beta-Alanine metabolism; 34. Propanoate metabolism; 35. Mismatch repair; 36. Glycine, serine and threonine metabolism;

37. Alanine and aspartate metabolism; 38. Nitrogen metabolism; 39. Purine metabolism; 40. Galactose metabolism; 41. Thiamine metabolism; 42. Ascorbate and aldarate metabolism; 43. Glycerophospholipid metabolism; 44. Peptidoglycan biosynthesis; 45. Methane metabolism; 46. Pentose and glucuronate interconversions; 47. Urea cycle and metabolism of amino groups; 48. Fructose and mannose metabolism; 49. Valine, leucine and isoleucine biosynthesis; 50. Glycolysis / Gluconeogenesis; 51. Two-component system - General; 52. ABC transporters - Organism-specific; 53. Pentose phosphate pathway; 54. Starch and sucrose metabolism; 55. Tryptophan metabolism; 56. Porphyrin and chlorophyll metabolism; 57. ABC transporters – General; 58. Two-component system - Organism-specific; 59. Aminosugars metabolism; 60. Phosphotransferase system (PTS); 61. Pyruvate metabolism; 62. Lysine biosynthesis; 63. Tyrosine metabolism; 64. Methionine metabolism; 65. Glutamate metabolism; 66. Riboflavin metabolism; 67. Glycerolipid metabolism; 68. Nucleotide sugars metabolism; 69. Glyoxylate and dicarboxylate metabolism; 70. Arginine and proline metabolism; 71. Nicotinate and nicotinamide metabolism; 72. Pantothenate and CoA biosynthesis; 73. Glutathione metabolism; 74. Base excision repair; 75. Drug metabolism - other enzymes.

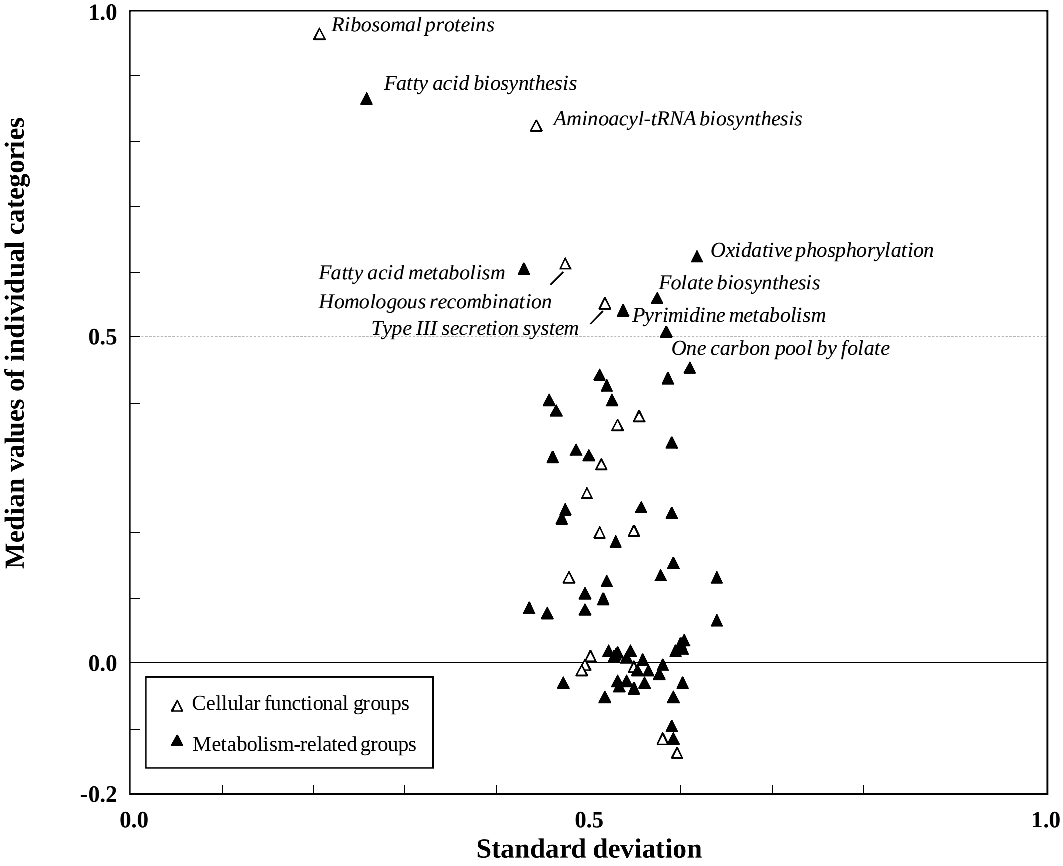


Figure 4.4. Comparison of 75 KEGG pathways by PCCs. Abscissa and ordinate axes correspond to standard deviations and medians of PCCs within individual categories. 75 KEGG pathways correspond to either cellular functional and metabolism-related groups, indicated by white and black triangles, respectively. Names of functional categories with more than 0.5 median of PCCs are indicated, i.e. ribosomal proteins, fatty acid biosynthesis, aminoacyl-tRNA biosynthesis, oxidative phosphorylation, homologous recombination, fatty acid metabolism, folate biosynthesis, type III secretion system, pyrimidine metabolism, and one carbon pool by folate.

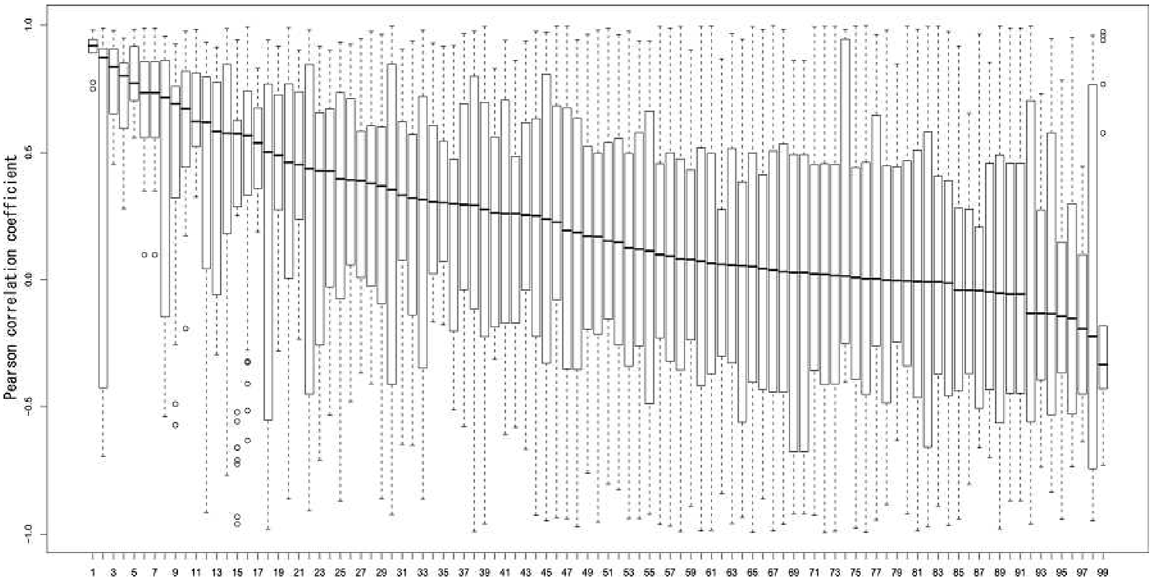


Figure 4.5. Boxplots of gene-to-gene PCCs of 99 TF regulated units. Characterizations of each TF regulated units based on Pearson correlation coefficients (PCCs) using gene expression profiles. The ordinate axis corresponds to the values of PCCs from minus one to plus one. The number below each boxplot corresponds to the following TFs: 1. ArgP; 2. GadW; 3. GlpR; 4. LsrR; 5. GatR; 6. GutM; 7. GutR; 8. PrpR; 9. CaiF; 10. DeoR; 11. GlcC; 12. CsgD; 13. ChbR; 14. MalT; 15. YiaJ; 16. PdhR; 17. HcaR; 18. RcsAB; 19. EnvY; 20. PurR; 21. AllR; 22. DnaA; 23. BaeR; 24. AraC; 25. FadR; 26. AppY; 27. RbsR; 28. NanR; 29. DcuR; 30. GadX; 31. TrpR; 32. UxuR; 33. DgsA; 34. CusR; 35. CueR; 36. FucR; 37. PaaX; 38. CytR; 39. NtrC; 40. Zur; 41. HyfR; 42. NhaR; 43. IdnR; 44. IscR; 45. RstA; 46. TorR; 47. SoxS; 48. TyrR; 49. FhlA; 50. NarP; 51. CysB; 52. CdaR; 53. ModE; 54. GntR; 55. OxyR; 56. FruR; 57. FlhDC; 58. NarL; 59. HU; 60. Fur; 61. Lrp; 62. AgaR; 63. Nac; 64. MetR; 65. Fis; 66. RutR; 67. ArcA; 68. LexA; 69. TdcA; 70. TdcR; 71. PhoB; 72. CRP; 73. IHF; 74. PspF; 75. H-NS; 76. FNR; 77. LeuO; 78. NagC; 79. ExuR; 80. Rob; 81. CpxR; 82. ArgR; 83. PhoP; 84. NsrR; 85. MetJ; 86. Cbl; 87. AlsR; 88. SgrR; 89. MarA; 90. GalR; 91. GalS; 92. GadE; 93. UlaR; 94. EvgA; 95. BirA; 96. NrdR; 97. DicA; 98. OmpR; 99. DpiA.

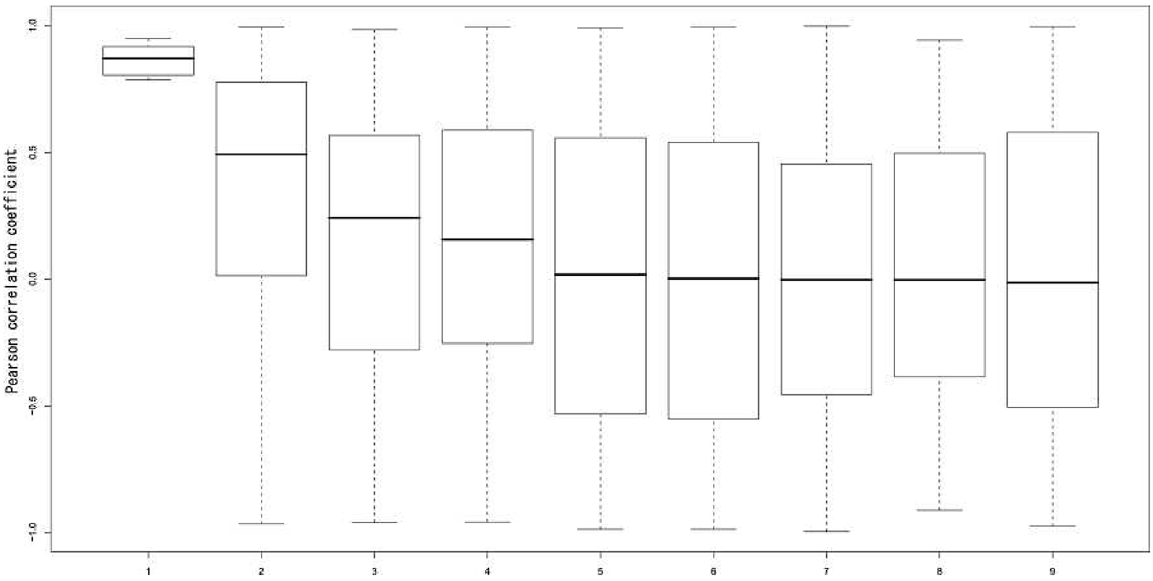


Figure 4.6. Boxplots of gene-to-gene PCCs of 9 sigma factor regulated units. Characterizations of each sigma factor regulated units based on Pearson correlation coefficients (PCCs) using gene expression profiles. The ordinate axis corresponds to the values of PCCs from minus one to plus one. The number below each boxplot corresponds to the following sigma factors: 1. Sigma19; 2. Sigma38; 3. Sigma28; 4. Sigma54; 5. Sigma24; 6. Sigma32; 7. Sigma70; 8. Sigma70, Sigma32; 9. Sigma70, Sigma38.

4.3.2 Gene-to-metabolite correlation analysis, taking into consideration a time lag between transcriptomics and metabolomics data

According to the analysis of transition points as described above, there is 60 minute time lag between transcriptomics and metabolomics data, suggesting that it is necessary to take into consideration a time lag between transcriptomics and metabolomics data, when performing the integrative analysis, i.e. gene-to-metabolite correlation analysis. First, to remove noise, I selected only significantly expressed genes and metabolites, and then calculated all gene-to-metabolite correlation pairs. 1,162 gene expression and 174 metabolite accumulation profiles were used for gene-to-metabolite correlation analysis. I set the time lag to be between 0 and 90 minutes, and calculated PCCs between gene expression profiles and each time lagged metabolite accumulation profiles obtained from experiments as described in Materials and methods.

Figure 4.7 shows the numbers of highly correlated gene-metabolite pairs ($r \geq 0.9$). The number of correlated pairs increases with time lag up to 50 min and after that it decreases (indicated by black line at the top), suggesting that many time lag specific gene-to-metabolite correlated pairs can be detected by taking into consideration the time lag between transcriptomics and metabolomics data. In order to investigate what biological processes are associated with metabolites in time lag specific manner, I determined the overrepresentation of the Gene Ontology (GO) annotations [Ashburner et al. 2000] among the genes associated to highly correlated gene-metabolite pairs corresponding to different time lagged data. Significant relations between the GO and metabolites were obtained by Fisher’s exact test. Figure 4.7 shows the GO terms under the ‘biological process’ annotation category that are significantly associated (p-value $\leq$ 0.01) to highly correlated gene-metabolite pairs determined by using gene expression profile and time lagged metabolite accumulation profile data. For example, when 30 min time lag is considered, genes involved in lipid A biosynthetic process, purine base biosynthetic process, and glutamate metabolic process, are overrepresented (indicated by black diamond or line in Fig. 4.7), while those genes are not overrepresented if no time lag is considered. Genes involved in barrier septum formation, cell division, and cell cycle are overrepresented in both 80 min and 90 min time lagged data, while no significant GO terms can be associated to 60 min time lagged data. These results suggest that I could detect some time lag specific genes, and taking into consideration a time lag between transcriptomics and metabolomics data is necessary for integrated analysis in time series experiments. Taking into consideration a time lag between transcriptomics and metabolomics data can make us elucidate direct or time lagged gene-to-metabolite relations.

In metabolomics data, I detected ten phosphatidylglycerols (PGs) as most abundant metabolites in time series analysis, i.e. the ions with *m/z* = 719.4883 (PG1 as shown in Fig. 4.7), 733.5056 (PG2), 747.5183 (PG3), 761.5293 (PG4), 691.4588 (PG5), 705.4757 (PG6), 745.5045 (PG7), 759.5242 (PG8), 773.5375 (PG9), and 787.5556 (PG10). In addition to global view of gene-to-metabolite relations, I analyzed gene-to-metabolite correlation analysis with respect to two groups of PGs, i.e. unsaturated phospholipids (called odd-numbered PGs; PG1, PG3, PG5, PG7, and PG9) and cyclopropanated phospholipids (called even-number PGs; PG2, PG4, GP6, PG8, and PG10), in order to elucidate gene-to-PG networks. In Figure 4.7 at the top, the numbers of correlated gene-to-odd-numbered PGs and gene-to-even-numbered PGs pairs are shown in red and blue line, respectively. I determined the GO terms under the ‘biological process’ annotation category that are overrepresented in the genes which are highly correlated to odd-numbered and even-numbered PGs by Fisher’s exact test. The cyclopropanation of odd-numbered PGs begins as the cells enter the stationary phase of growth [Grogan and Cronan 1997]. The cyclopropanation is thought to be involved in the long-term survival of nongrowing cells and is often associated with environmental stress such as acidic stress in *E. coli* [Grogan and Cronan 1997; Cronan 2002] and with pathogenesis in *Mycobacterium tuberculosis* [Cronan 2002]. Activity of CFA synthase is modulated by transcriptional and post-translational levels [Wang and Cronan 1994]. CFA formation is largely restricted to the transition between the late exponential- and early stationary-phase [Law 1971]. Thus, multifaceted regulation should be affected to synthesis of CFAs. The results of overrepresented GO terms in gene-to-PG correlated pairs are also shown in Figure 4.7. Genes associated with lipid biosynthetic process and fatty acid biosynthetic process are correlated with odd-numbered PGs in a time lag specific manner (indicated by red diamond or line in Fig. 4.7). These correlations, i.e. lipid metabolism correlations between transcriptional and metabolite levels, coincide with biological meaning, again indicating that taking into consideration a time lag between transcriptomics and metabolomics data is necessary for analysis of time series. Even-numbered PGs are correlated with genes associated with biofilm formation in a time lag specific manner (indicated by blue diamond or line in Fig. 4.7).

Finally, I focused on the gene-to-metabolite correlations with respect to the pathway of the phospholipids synthesis, in which PGs were synthesized. In metabolite accumulation profiles, odd-numbered PGs are accumulated in the exponential phase, and even-numbered

PGs are accumulated in the stationary phase (in Fig. 4.8). A key intermediate in phospholipid synthesis is cytosine diphosphate (CDP)-diacylglycerol (DAG), which is formed by CdsA from phosphatidic acid and cytosine triphosphate. In the biochemical regulation of phospholipid composition, the zwitterionic (phosphatidylethanolamine) and acidic (PG and cardiolipin) branches of phospholipids synthesis compete for a common pool of CDP-DAG. PGs are synthesized from CDP-DAG in two steps, by PgsA, and PgpA/B [Zhang and Rock 2008]. Figure 4.8 shows *cdsA*, *pgsA* expression profiles and average accumulation profiles of odd-numbered and even-numbered PGs, indicating that there could be the time lag between gene expression and metabolite accumulation profiles, as predicted by LDS analysis. In correlation analysis with respect to PGs, *cdsA* and *pgsA* were correlated with PG9 and all even-numbered PGs in 30 min time lagged data, respectively, while no PGs correlated with those genes when no time lag was considered. *pgpA* correlated with some even-numbered PGs in both no lag and 30 min time lagged data. Expression profiles of PG9 and *cdsA* were down-regulated through cell growth, while those of even-numbered PGs, *pgsA*, and *pgpA* were up-regulated (indicated by blue or red arrows beside gene names in Fig. 4.9). These results indicate that the content of CDP-DAG decreases according to decreasing of transcriptional level of *cdsA*, and the pathway from CDP-DAG to PGs is active with the increase of transcriptional level of *pgsA* and *pgpA*. There is another biochemical pathway from CDP-DAG, which leads to phosphatidylethanolamine. Two genes associated with this pathway are decreasing through cell growth, although the amount changes of those are not significant in microarray data, i.e. below the threshold (indicated by blue dotted arrows beside gene names in Fig. 4.9). These results indicate that the pool of CDP-DAG could be used to synthesize PGs but not phosphatidylethanolamine, when *E. coli* enters into stationary phase from exponential phase. It has been reported previously that *Bacillus subtilis* PssA, a molecularly distinct integral membrane protein, was used to replace the transiently membrane-associated PssA in *E. coli*. Amplification of *B. subtilis* PssA increases the relative and absolute amounts of phosphatidylethanolamine and impairs growth [Saha et al. 1996a; Saha et al. 1996b]. Therefore, the balance of zwitterionic (phosphatidylethanolamine) and acidic phospholipids (PGs) in *E. coli* is important. So, from this analysis, it can be said that PGs could be more responsible for membrane balance than phosphatidylethanolamine.

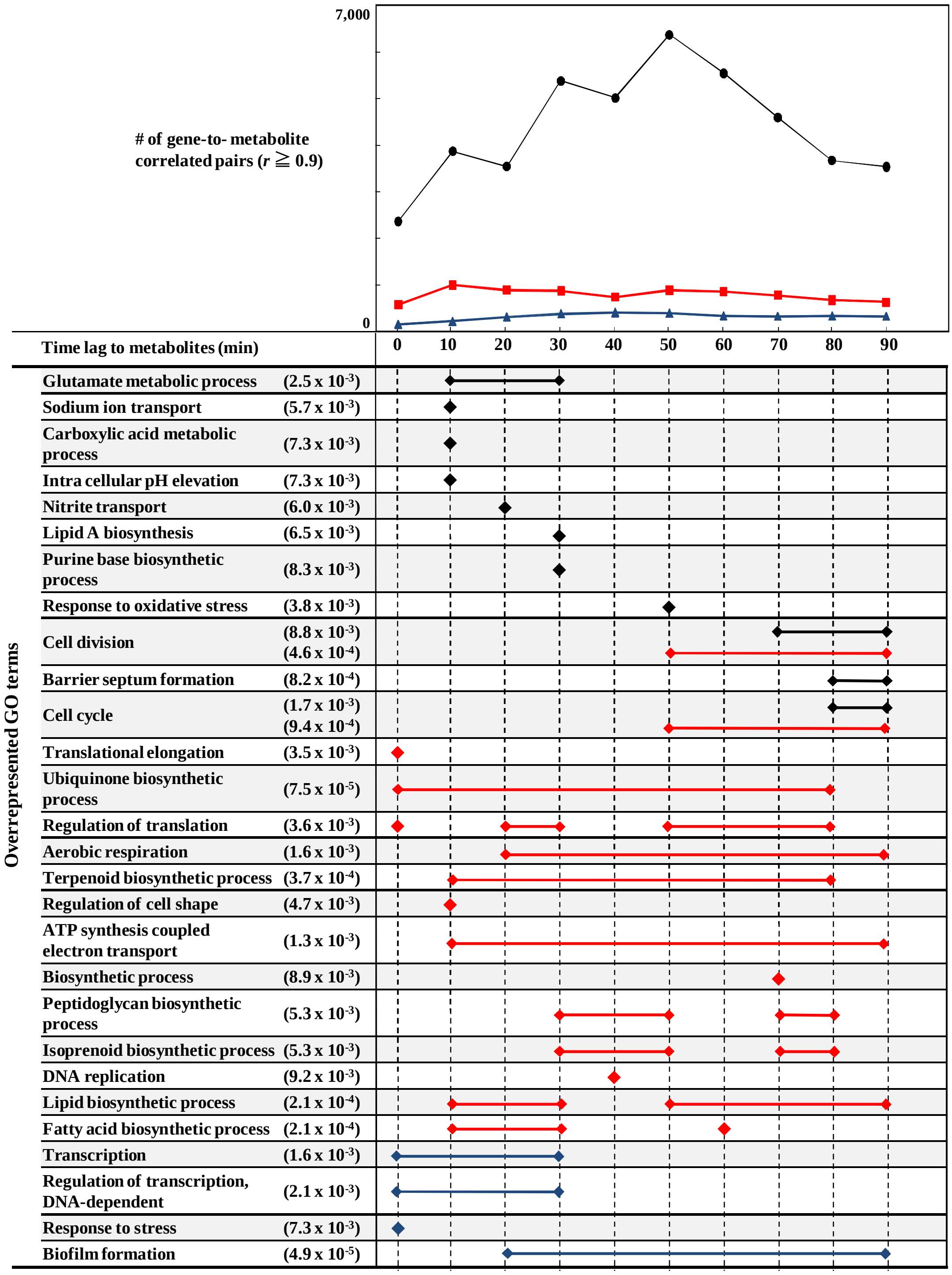


Figure 4.7. Gene-to-metabolite correlation analysis taking into consideration a time lag between

transcriptomics and metabolomics data. The plot is the number of correlated gene-to-metabolite pairs (PCC $\geq$ 0.9). Black, red, and blue indicates gene-to-metabolite, gene-to-odd-numbered PG, gene-to-even-numbered PG correlated pairs, respectively. Time lag of abscissa axis means time lag considered between transcriptomics and metabolomics, according to the procedure as described in Materials and methods. GO terms with *p*-value less than 0.01 (minimum *p*-value is indicated, if any) are indicated at the left side. Diamond and bar correspond to the significant overrepresented data by Fisher's exact test. Glutamate metabolic process, for example, is overrepresented in correlated pairs between transcriptomics and 10 min, 20 min, and 30 min time lag metabolomics data. Black, red and blue colors correspond to correlated pairs corresponding to 174 metabolites, odd-numbered PGs, and even-numbered PGs with 1,162 genes, respectively.

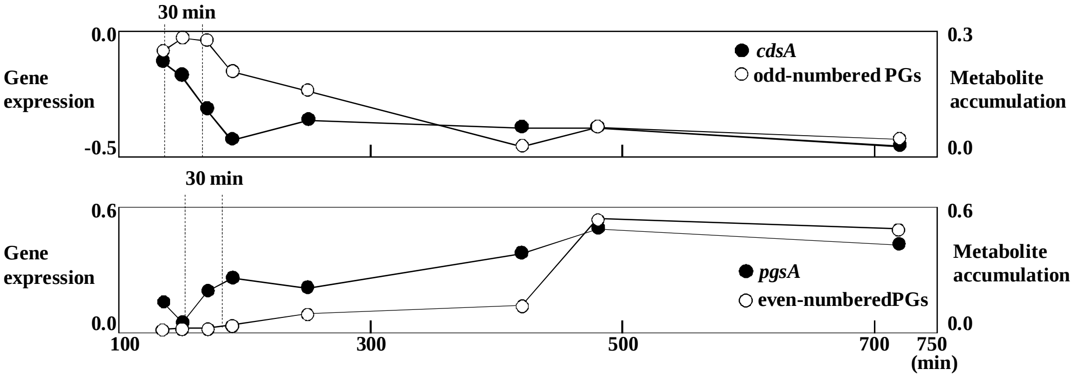


Figure 4.8. Gene expression (top: *cdsA*, bottom: *pgsA*) and metabolite accumulation profiles (top: odd-numbered PGs, bottom: even-numbered PGs). Metabolite accumulation profiles correspond to average values of each PG group.

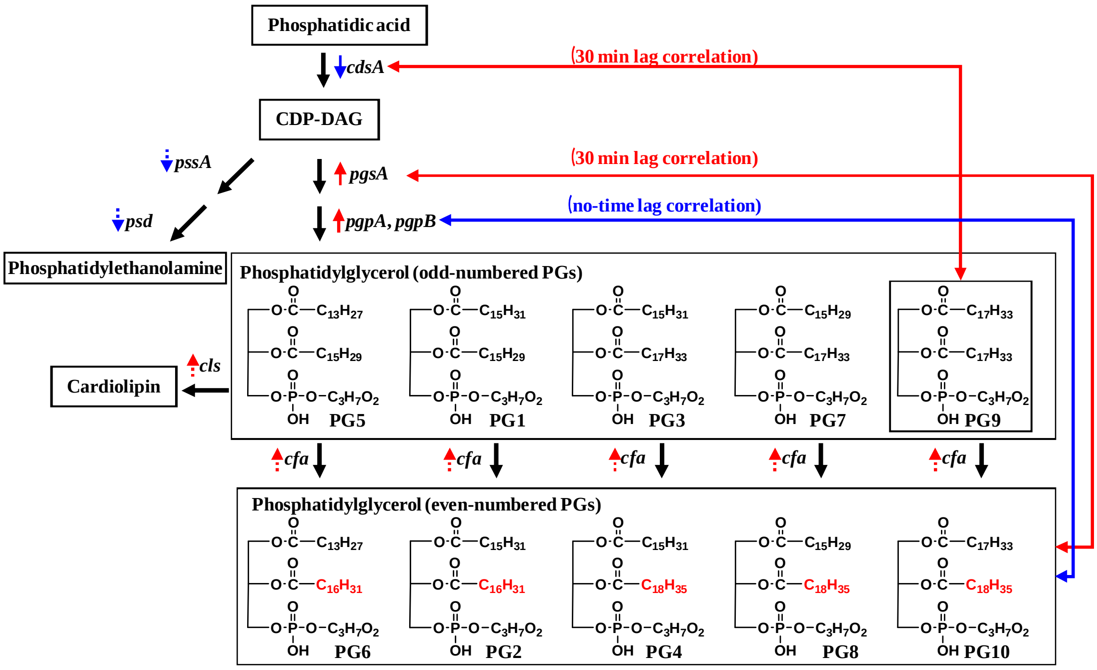


Figure 4.9. The pathways from phosphatidic acid to phosphatidylethanolamine or caldiolipin. Metabolite names are inside in the squares and gene names are beside black solid arrows, which represent chemical reactions. Arrows beside gene names indicate up and down transcriptional levels through cell growth. Dashed arrows mean the weak changes in expression. The lines between genes and metabolites indicate correlations between genes and odd-numbered or even-numbered PGs, respectively.

Chapter 5
Concluding remarks

Unlike sequence-based macrostructures such as DNA, RNA or proteins, metabolites that are small organic molecules have highly diverse chemical features. Many compounds of different chemical classes, e.g. sugars, organic acids, lipids etc. simultaneously exist in the cell. Only with the development of robust analytical technologies, adapted to a wide range of applications, the informative detection of a large number of metabolites has been made possible. Out of those, FT-ICR/MS has unique potential to facilitate metabolomics research due to its ability to detect molecular mass with high accuracy. The results of FT-ICR/MS analysis may in some cases point to several or a large number of different possible isomers. Nevertheless, even when identification of the specific isomer is not possible directly after FT-ICR/MS analysis, it provides a clue to the class of compounds to which this isomer belongs. Thus, it will narrow down the search for metabolite identity.

Since Aharoni et al. (2002) first applied FT-ICR/MS for metabolomics using strawberry ripening, to my knowledge there has been no paper which analyses bacteria metabolomics by using FT-ICR/MS. The performance characteristics of FT-ICR/MS are ideal for the types of complex mixtures encountered in high throughput metabolomics applications. However, there are several procedures for FT-ICR/MS and no standard procedures. So in Chapter 2, I have developed the platform of metabolomics informatics for FT-ICR/MS, which consists of four stages: (i) peak correction, (ii) multivariate data processing, (iii) unsupervised learning such as principal component analysis (PCA) and batch-learning SOM (BL-SOM), and (iv) supervised learning such as partial least squares (PLS) regression. This procedure could be applied to metabolomics for any organism, although I applied this procedure only to bacteria metabolomics in this dissertation.

In Chapter 3, I established non-targeted metabolomics approach to analyze growth-specific metabolites of bacteria, based on the FT-ICR/MS platform (Fig. 3.1). Bioinformatics played a crucial role for this analysis, e.g. multivariate analysis and database search. Correlation analysis has made it possible to predict unknown molecular structure using isotope ratios by way of grouping metabolite derivative ions. Though 1 ppm mass accuracy alone is insufficient for unique elemental composition assignment [Kind and Fiehn 2006], in metabolite annotation by using the mass spectrometry technology, integrated analysis based on information of isotope relation, fragmentation patterns by MS/MS analysis, and co-occurring metabolites has enabled to annotate ions as metabolites and estimate cellular conditions based on metabolite composition. PCA revealed the differences between the growth stages on the basis of 220 independent metabolites, suggesting that metabolic profiling is a useful method for distinguishing the growth stages. Using PLS regression, I constructed a linear relationship between OD_{600} values and metabolite profiles. High correlation between predicted and observed OD_{600} values certifies the correctness of the linear model. I anticipate that the method presented will be an important tool for future functional genomics research.

Cellular behavior results from the action of and interplay between the distinct networks. Studying and comparing the responses triggered by these different networks and their interrelation is of great interest [De Keersmaecker et al. 2006]. Kromer et al. (2004)

studying lysine production in *Corynebacterium glutamicum* or Lafaye et al. (2005) studying the yeast sulphur pathway integrated metabolite profiles and metabolic fluxes with high-throughput transcriptomics or proteomics data, respectively. Hirai et al. (2005) studying *Arabidopsis thaliana* integrated metabolite profiles with transcriptomics data and could identify regulatory metabolites and transcriptional factor genes. Measurement of metabolites can give and complement information on how functional proteins act.

In Chapter 4, I tried to integrate metabolite accumulation profiles with gene expression profiles by Pearson correlation coefficients and indicated that it is necessary to take into consideration the time lag between transcriptomics and metabolomics data. I detected transition point by LDS analysis. Transition points predicted by transcriptomics and metabolomics data were different, indicating that there is a time lag between transcriptomics and metabolomics data. There are several reasons for this time lag. One major reason is a consequence of the central dogma in biology, i.e. information flow from genome to protein. I performed gene-to-gene correlation analysis by using KEGG pathway annotations, in order to elucidate more global regulations instead of only functional regulations. I observed that genes related to three pathways are highly positively coregulated through cell growth (Fig. 4.3 and 4.4). This method is the effective way to characterize pathways from a global view of gene-to-gene regulations unlike regulations by operon and regulon, because cell is considered as a system. Gene-to-gene correlation analysis by using transcription factors and sigma factors based on RegulonDB was also performed (Fig. 4.5 and 4.6). Most of regulatory units are not coexpressed. One reason for this is that those units are not tightly regulated through cell growth, and another is that annotations of gene regulation are not complete yet. I performed extended gene-to-metabolite correlation analysis, i.e. using time lagged metabolomics data prepared by linear interpolation. Statistical test, i.e. Fisher's exact test, shows that there are several gene functional categories correlated with metabolites in the time lag specific manner. To my knowledge, although the time lag between expression profiles of transcription factor and regulated genes has been already reported [Redestig et al. 2007], this is the first report demonstrating that considering time lag between transcriptomics and metabolomics data is one of the effective ways to unravel the complex gene-to-metabolite networks. In gene-to-PG correlation analyses (Fig. 4.9), relations of *cdsA*, *pgsA*, *pgpA*, and *pgpB* with PGs could support the model that the pool of CDP-DAG could be used to synthesize PGs but not phosphatidylethanolamine.

Metabolomics analysis is an emerging field. Metabolomics must become a full scale discovery platform which supports and feeds into complementary, parallel activities to advance our understanding of multidimentional biological systems [Hall 2006]. The potential applications of metabolomics have already been demonstrated in a wide range of disciplines. Even though further improvements of technologies and strategies for metabolomics analyses are likely to develop, the unambiguous and simultaneous identification of all metabolites in a biological system is still a big challenge. LC-MS system can provide more information of metabolites, though I used direct-infusion FT-ICR/MS in this dissertation. High-throughput systems, database developments, and routine applications using LC-MS are promising. In addition, ^{13}C flux analyses will provide information which is not available from just metabolic snapshots. Also, for the development of metabolomics analysis, it is crucial to keep metabolomic database publicly available. The contribution of metabolomics to systems biology can become even more relevant in combination with transcriptomics and proteomics analyses for the progressive understanding of the cell.

URLs

EcoCyc: http://ecocyc.org/

GO: http://www.geneontology.org/

KEGG: http://www.genome.jp/kegg/

PubChem: http://pubchem.ncbi.nlm.nih.gov/

RegulonDB: http://regulondb.ccg.unam.mx/index.jsp

DPClus: http://kanaya.naist.jp/DPClus/

DrDMASS+: http://kanaya.naist.jp/DrDMASSplus/

KNApSAcK: http://kanaya.naist.jp/KNApSAcK/

TREBAX: http://kanaya.naist.jp/~skanaya/Web/software/trebax/trebax2.html

References

Abe, T., S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura. 2003. Informatics for unveiling hidden genome signatures. *Genome Res* **13**: 693-702.

Aharoni, A., C.H. Ric de Vos, H.A. Verhoeven, C.A. Maliepaard, G. Kruppa, R. Bino, and D.B. Goodenowe. 2002. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *Omics* **6**: 217-234.

Ali, S.T., A.J. Moir, P.R. Ashton, P.C. Engel, and J.R. Guest. 1990. Octanoylation of the lipoyl domains of the pyruvate dehydrogenase complex in a lipoyl-deficient strain of Escherichia coli. *Mol Microbiol* **4**: 943-950.

Allen, J., H.M. Davey, D. Broadhurst, J.K. Heald, J.J. Rowland, S.G. Oliver, and D.B. Kell. 2003. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol* **21**: 692-696.

Altaf-Ul-Amin, M., Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya. 2006. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* **7**: 207.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Res* **28**: 304-305.

Balasubramaniyan, R., E. Hullermeier, N. Weskamp, and J. Kamper. 2005. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics* **21**: 1069-1077.

Boecker, S., M.C. Letzel, Z. Liptak, and A. Pervukhin. 2006. WABI:12.

Boulesteix, A.L. and K. Strimmer. 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* **8**: 32-44.

Brauer, M.J., J. Yuan, B.D. Bennett, W. Lu, E. Kimball, D. Botstein, and J.D. Rabinowitz. 2006. Conservation of the metabolomic response to starvation across two divergent microbes. *Proc Natl Acad Sci U S A* **103**: 19302-19307.

Castrillo, J.I. and S.G. Oliver. 2004. Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics. *J Biochem Mol Biol.* **31**: 93-106.

Chang, Y.Y. and J.E. Cronan, Jr. 1999. Membrane cyclopropane fatty acid content is a major factor in acid resistance of Escherichia coli. *Mol Microbiol* **33**: 249-259.

Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. 1998. The transcriptional program of sporulation in budding yeast. *Science* **282**: 699-705.

Cronan, J.E., Jr. 2002. Phospholipid modifications in bacteria. *Curr Opin Microbiol* **5**: 202-205.

De Keersmaecker, S.C., I.M. Thijs, J. Vanderleyden, and K. Marchal. 2006. Integration of omics data: how well does it work for bacteria? *Mol Microbiol.* **62**: 1239-1250.

De Laeter, J.R., J.K. Bohlke, P. De Bievre, H. Hidaka, H.S. Peiser, K.J.R. Rosman, and P.D.P. Taylor. 2003. Atomic weights of the elements. Review 2000 (IUPAC Technical Report). *Pure Apply Chem* **75**: 683-800.

De Luca, V. and B. St Pierre. 2000. The cell and developmental biology of alkaloid biosynthesis. *Trends Plant Sci* **5**: 168-173.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680-686.

Dudoit, S., J. Fridlyand, and T. Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Ass.* **97**: 77-87.

Fiehn, O. 2002. Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* **48**: 155-171.

Fiehn, O., S. Kloska, and T. Altmann. 2001. Integrated studies on plant biology using multiparallel techniques. *Curr Opin Biotechnol* **12**: 82-86.

Fiehn, O., J. Kopka, P. Dormann, T. Altmann, R.N. Trethewey, and L. Willmitzer. 2000. Metabolite profiling for plant functional genomics. *Nat Biotechnol* **18**: 1157-1161.

Fiehn, O., G. Wohlgemuth, and M. Scholz. 2005. Fiehn, O., Wohlgemuth, G., Scholz, G. (2005) Setup and Annotation of Metabolomic Experiment by Intergrating Biological and Mass Spectrometric Metadata. In B. Ludascher, L. Raschid, eds, LNBI, Vol 3615.

Springer-Verlag, Berlin, Germany, pp. pp224-239.

Fisher, R. 1958. The correlation coefficient. In: Fisher RA (ed) Statistical methods for research workers, Ed. 13.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, and et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496-512.

Fridman, E. and E. Pichersky. 2005. Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Curr Opin Plant Biol* **8**: 242-248.

Galperin, M.Y. and M.J. Ellison. 2006. Systems biology: sprint or marathon? *Curr Opin Biothechnol* **17**: 437-439.

Gama-Castro, S., V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M.I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A.M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides. 2008. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36:** D120-124.

Gauthier, J.W., T.R. Trautman, and D.B. Jacobson. 1991. Sustained off-resonance irradiation for collision-activated dissociation involving Fourier transform mass spectrometry. Collision-activated dissociation technique that emulates infrared multiphoton dissociation. *Analytica Chimica Acta* **246**: 211-225.

Goto, S., Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. 2002. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* **30**: 402-404.

Grogan, D.W. and J.E. Cronan, Jr. 1997. Cyclopropane ring formation in membrane lipids of bacteria. *Microbiol Mol Biol Rev* **61**: 429-441.

Hall, R.D. 2006. Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol* **169**: 453-468.

Han, J.S., H.S. Kwon, J.B. Yim, and D.S. Hwang. 1998. Effect of IciA protein on the expression of the nrd gene encoding ribonucleoside diphosphate reductase in E. coli. *Mol Gen Genet* **259:** 610-614.

Hirai, M.Y., M. Klein, Y. Fujikawa, M. Yano, D.B. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, J. Papenbrock, and K. Saito. 2005. Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *J Biol Chem* **280**: 25590-25595.

Hirai, M.Y., M. Yano, D.B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito. 2004. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **101**: 10205-10210.

Hounsome, N., B. Hounsome, D. Tomos, and G. Edwards-Jones. 2009. Changes in antioxidant compounds in white cabbage during winter storage. *Postharvest Biology and Technology*. *in press*.

Hughes, T.R., M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102**: 109-126.

Ishinaga, M., R. Kanamoto, and M. Kito. 1979. Distribution of phospholipid molecular species in outer and cytoplasmic membrane of Escherichia coli. *J Biochem (Tokyo)* **86**: 161-165.

Ji, L. and K.L. Tan. 2005. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics* **21**: 509-516.

Joyce, A.R. and B.O. Palsson. 2006. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7**: 198-210.

Kalman, R.E. and R.S. Bucy. 1961. New results in linear filtering and prediction theory. *Trans ASME, J Basic Eng* **83**: 95-107.

Kanaya, S., M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, and T. Ikemura. 2001. Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli

O157 genome. *Gene* **276**: 89-99.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**: D480-484.

Kanehisa, M. and P. Bork. 2003. Bioinformatics in the post-sequence era. *Nat Genet* **33** Suppl: 305-310.

Kanehisa, M., S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**: D354-357.

Keseler, I.M., J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P.D. Karp. 2005. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* **33:** D334-337.

Kind, T. and O. Fiehn. 2006. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **7**: 234.

Kind, T. and O. Fiehn. 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**: 105.

Kobayashi, H., J. Akitomi, N. Fujii, K. Kobayashi, M. Altaf-Ul-Amin, K. Kurokawa, N. Ogasawara, and S. Kanaya. 2007. The entire organization of transcription units on the Bacillus subtilis genome. *BMC Genomics* **8:** 197.

Kromer, J.O., O. Sorgenfrei, K. Klopprogge, E. Heinzle, and C. Wittmann. 2004. In-depth profiling of lysine-producing Corynebacterium glutamicum by combined analysis of the transcriptome, metabolome, and fluxome. *J Bacteriol* **186:** 1769-1784.

Lafaye, A., C. Junot, Y. Pereira, G. Lagniel, J.C. Tabet, E. Ezan, and J. Labarre. 2005. Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *J Biol Chem* **280:** 24723-24730.

Laskin, J. and J.H. Futrell. 2005. Activation of large ions in FT-ICR mass spectrometry. *Mass Spectrom Rev* **24**: 135-167.

Law, J.H. 1971. Biosynthesis of cyclopropane rings. *Acc Chem Res* **4**: 199-203.

Ma, Z., H. Richard, D.L. Tucker, T. Conway, and J.W. Foster. 2002. Collaborative regulation of Escherichia coli glutamate-dependent acid resistance by two AraC-like regulators, GadX and GadW (YhiW). *J Bacteriol* **184:** 7001-7012.

Magnuson, K., S. Jackowski, C.O. Rock, and J.E. Cronan, Jr. 1993. Regulation of fatty acid biosynthesis in Escherichia coli. *Microbiol Rev* **57**: 522-542.

Marshall, A.G., C.L. Hendrickson, and S.D. Shi. 2002. Scaling MS plateaus with high-resolution FT-ICRMS. *Anal Chem* **74**: 252A-259A.

Mata, J., S. Marguerat, and J. Bahler. 2005. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci* **30**: 506-514.

Merlie, J.P. and L.I. Pizer. 1973. Regulation of phospholipid synthesis in Escherichia coli by guanosine tetraphosphate. *J Bacteriol* **116**: 355-366.

Morioka, R., S. Kanaya, M.Y. Hirai, M. Yano, N. Ogasawara, and K. Saito. 2007. Predicting state transitions in the transcriptome and metabolome using a linear dynamical system model. *BMC Bioinformatics* **8**: 343.

Nakamura, Y., A. Kimura, H. Saga, A. Oikawa, Y. Shinbo, K. Kai, N. Sakurai, H. Suzuki, M. Kitayama, D. Shibata, S. Kanaya, and D. Ohta. 2007. Differential metabolomics unraveling light/dark regulation of metabolic activities in Arabidopsis cell culture. *Planta* **227**: 57-66.

Oikawa, A., Y. Nakamura, T. Ogura, A. Kimura, H. Suzuki, N. Sakurai, Y. Shinbo, D. Shibata, S. Kanaya, and D. Ohta. 2006. Clarification of pathway-specific inhibition by Fourier transform ion cyclotron resonance/mass spectrometry-based metabolic phenotyping studies. *Plant Physiol* **142**: 398-413.

Pir, P., B. Kirdar, A. Hayes, Z.Y. Onsan, K.O. Ulgen, and S.G. Oliver. 2006. Integrative investigation of metabolic and transcriptomic data. BMC Bioinformatics 12: 203.

Polakis, S.E., R.B. Guchhait, and M.D. Lane. 1973. Stringent control of fatty acid synthesis in Escherichia coli. Possible regulation of acetyl coenzyme A carboxylase by ppGpp. *J Biol Chem* **248**: 7957-7966.

Primig, M., R.M. Williams, E.A. Winzeler, G.G. Tevzadze, A.R. Conway, S.Y. Hwang, R.W. Davis, and R.E. Esposito. 2000. The core meiotic transcriptome in budding yeasts. *Nat Genet* **26**: 415-423.

Quackenbush, J. 2002. Microarray data normalization and transformation. *Nat Genet* **32** Suppl: 496-501.

Redestig, H., D. Weicht, J. Selbig, and M.A. Hannah. 2007. Transcription factor target prediction using multiple short expression time series from Arabidopsis thaliana. *BMC Bioinformatics* **8**.

Roessner, U., A. Luedemann, D. Brust, O. Fiehn, T. Linke, L. Willmitzer, and A. Fernie. 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**: 11-29.

Roessner, U., L. Willmitzer, and A.R. Fernie. 2001. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol* **127**: 749-764.

Roy, R.S., A.M. Gehring, J.C. Milne, P.J. Belshaw, and C.T. Walsh. 1999. Thiazole and oxazole peptides: biosynthesis and molecular machinery. *Nat Prod Rep* **16:** 249-263.

Saha, S.K., Y. Furukawa, H. Matsuzaki, I. Shibuya, and K. Matsumoto. 1996. Directed mutagenesis, Ser-56 to Pro, of Bacillus subtilis phosphatidylserine synthase drastically lowers enzymatic activity and relieves amplification toxicity in Escherichia coli. *Biosci Biotechnol Biochem* **60:** 630-633.

Saha, S.K., S. Nishijima, H. Matsuzaki, I. Shibuya, and K. Matsumoto. 1996. A regulatory mechanism for the balanced synthesis of membrane phospholipid species in Escherichia coli. *Biosci Biotechnol Biochem* **60:** 111-116.

Scholz, M. and O. Fiehn. 2007. SetupX--a public study design database for metabolomic projects. *Pac Symp Biocomput*: 169-180.

Shinbo, Y., Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya. 2006. KNApSAcK: A Comprehensive Species-Metabolite Relationship Database. In Plant Metabolomics, pp. 165-181.

Soga, T., Y. Ohashi, Y. Ueno, H. Naraoka, M. Tomita, and T. Nishioka. 2003. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J Proteome Res* **2**: 488-494.

Stein, L. 2001. Genome annotation: from sequence to biology. *Nat Rev Genet* **2**: 493-503.

Suzuki, H., R. Sasaki, Y. Ogata, Y. Nakamura, N. Sakurai, M. Kitajima, H. Takayama, S. Kanaya, K. Aoki, D. Shibata, and K. Saito. 2008. Metabolic profiling of flavonoids in Lotus japonicus using liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *Phytochemistry* **69**: 99-111.

Takahashi, H., K. Kai, Y. Shinbo, K. Tanaka, D. Ohta, T. Oshima, M. Altaf-Ul-Amin, K.

Kurokawa, N. Ogasawara, and S. Kanaya. 2008. Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry. *Anal Bioanal Chem* **391:** 2769-2782.

Tautenhahn, R., C. Boettcher, and S. Neumann. 2008. Annotation of LC/ESI-MS mass signals. Proceedings of BIRD 2007-1st international conference on bioinformatics research and development.

Tohge, T., Y. Nishiyama, M.Y. Hirai, M. Yano, J. Nakajima, M. Awazuhara, E. Inoue, H. Takahashi, D.B. Goodenowe, M. Kitayama, M. Noji, M. Yamazaki, and K. Saito. 2005. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *Plant J* **42**: 218-235.

Urbanczyk-Wochniak, E., A. Luedemann, J. Kopka, J. Selbig, U. Roessner-Tunali, L. Willmitzer, and A.R. Fernie. 2003. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep* **4**: 989-993.

Venturi, V. 2003. Control of rpoS transcription in Escherichia coli and Pseudomonas: why so different? *Mol Microbiol* **49:** 1-9.

Villas-Boas, S.G., S. Rasmussen, and G.A. Lane. 2005. Metabolomics or metabolite profiles? *Trends Biotechnol* **23**: 385-386.

Vimr, E.R., K.A. Kalivoda, E.L. Deszo, and S.M. Steenbergen. 2004. Diversity of microbial sialic acid metabolism. *Microbiol Mol Biol Rev* **68**: 132-153.

Visca, P., L. Leoni, M.J. Wilson, and I.L. Lamont. 2002. Iron transport and regulation, cell signalling and genomics: lessons from Escherichia coli and Pseudomonas. *Mol Microbiol* **45:** 1177-1190.

Wang, A.Y. and J.E. Cronan, Jr. 1994. The growth phase-dependent synthesis of cyclopropane fatty acids in Escherichia coli is the result of an RpoS(KatF)-dependent promoter plus enzyme instability. *Mol Microbiol* **11**: 1009-1017.

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, W. Helmberg, Y. Kapustin, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. 2006. Database resources of the National Center for Biotechnology Information.

Nucleic Acids Res **34**: D173-180.

Wishart, D.S., D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G.E. Duggan, G.D. Macinnis, A.M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, and L. Querengesser. 2007. HMDB: the Human Metabolome Database. *Nucleic Acids Res* **35**: D521-526.

Yang, Y.H., S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**: e15.

Yano, M., S. Kanaya, M. Altaf-Ul-Amin, K. Kurokawa, M.Y. Hirai, and K. Saito. 2006. Integrated Data Mining of Transcriptome and Metabolome Based on BL-SOM. *Journal of Computer Aided Chemistry* **7**: 125-136.

Yu, U., S.H. Lee, Y.J. Kim, and S. Kim. 2004. Bioinformatics in the post-genome era. *J Biochem Mol Biol* **37**: 75-82.

Zhang, Y.M. and C.O. Rock. 2008. Membrane lipid homeostasis in bacteria. *Nat Rev Microbiol* **6**: 222-233.

Achievements

List of publications and manuscripts in preparation

1. **Hiroki Takahashi**, Kosuke Kai, Yoko Shinbo, Kenichi Tanaka, Daisaku Ohta, Taku Oshima, Md. Altaf-Ul-Amin, Ken Kurokawa, Naotake Ogasawara, Shigehiko Kanaya. Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry. *Anal Bioanal Chem.* 2008 Aug;391(8):2769-82. Epub 2008 Jun 16.
2. Takashi Oishi, Ken-ichi Tanaka, Takuya Hashimoto, Yoko Shinbo, Kanokwan Jumtee, Takeshi Baba, Eiichiro Fukusaki, Hideyuki Suzuki, Daisuke Shibata, **Hiroki Takahashi**, Hiroko Asahi, Ken Kurokawa, Yukiko Nakamura, Aki Hirai, Kensuke Nakamura, Md. Altaf-Ul-Amin, Shigehiko Kanaya. An approach to peak detection in GC-MS chromatograms and applications of KNApSAcK database in prediction of candidate metabolites. *Plant Biotechnology.* 2009. *in press*.
3. **Hiroki Takahashi**, Ryoko Morioka, Ryosuke, Ito, Taku Oshima, Naotake Ogasawara, Md. Altaf-Ul-Amin, Shigehiko Kanaya. Dynamical change of gene-to-metabolite networks in time lag. *in preparation*.

International Conferences

1. **Hiroki Takahashi**, Ryosuke Ito, Taku Oshima, Naotake Ogasawara, Md. Altaf-Ul-Amin, Shigehiko Kanaya, Ken Kurokawa. Experimental Design for Time-Series Microarray Analysis. *The Fifth Asia-Pacific Bioinformatics Conference*. Hong Kong. January, 2007 (in Chapter 3)
2. **Hiroki Takahashi**, Kosuke Kai, Yoko Shinbo, Daisaku Ohta, Taku Oshima, Kenichi Tanaka, Md. Altaf-Ul-Amin, Ken Kurokawa, Naotake Ogasawara, Shigehiko Kanaya. Bioinformatics approach toward Metabolomics: Development of the metabolic profiling tool (DrDMASS+) and the species-metabolite relation database (KNApSAcK). *The 7th International Workshop on Advanced Genomics*. Tokyo. November, 2007 (in Chapter 3)
3. **Hiroki Takahashi**, Kosuke Kai, Yoko Shinbo, Daisaku Ohta, Taku Oshima, Kenichi

Tanaka, Md. Altaf-Ul-Amin, Ken Kurokawa, Naotake Ogasawara, Shigehiko Kanaya. Elucidation of stage-specific metabolites in *Escherichia coli* based on FT-ICR/MS and bioinformatics. *The Sixth Asia-Pacific Bioinformatics Conference*. Kyoto. January, 2008 (in Chapter 3)

4. **Hiroki Takahashi**, Yoko Shinbo, Md. Altaf-Ul-Amin, Ken Kurokawa, Shigehiko Kanaya. Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry. *8th International Conference on Chemical Structures*. Netherlands. June, 2008 (in Chapter 3)
5. Kosuke Kai, **Hiroki Takahashi**, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Shigehiko Kanaya, Daisaku Ohta. Reconstitution of metabolic activities on a Van Krevlen diagram, with the aid of accurate mass measurement. *5th International Conference on Plant Metabolomics*. Yokohama. July, 2008 (in Chapter 3)
6. Shigehiko Kanaya, Ken-ichi Tanaka, **Hiroki Takahashi**, Yoko Shinbo, Md. Altaf-Ul-Amin, Hiroko Asahi, Atsushi Fukushima, Aki Hirai, Kazuki Saito, Daisaku Ohta, Daisuke Shibata. Estimation methodology of substrate-product pairs based on reaction rules of enzymes using KNApSAcK db. *5th International Conference on Plant Metabolomics*. Yokohama. July, 2008 (in Chapter 3)
7. **Hiroki Takahashi**, Ryoko Morioka, Kosuke Kai, Yoko Shinbo, Daisaku Ohta, Taku Oshima, Md. Altaf-Ul-Amin, Naotake Ogasawara, Shigehiko Kanaya. Network analysis of gene-to-gene and gene-to-metabolite on the basis of the time lag between gene and metabolite. *The 2008 Annual Conference of the Japanese Society for Bioinformatics*. Osaka. December, 2008 (in Chapter 4)

Local Conferences

1. Ryosuke Ito, **Hiroki Takahashi**, Taku Oshima, Naotake Ogasawara, Md. Altaf-Ul-Amin, Shigehiko Kanaya, Ken Kurokawa. Effect of cDNA microarray experimental design in time series (in Japanese). **情報処理学会研究報告** Okinawa. June, 2006 (in Chapter 4)
2. **Hiroki Takahashi**, Kosuke Kai, Yoko Shinbo, Daisaku Ohta, Taku Oshima, Kenichi Tanaka, Md. Altaf-Ul-Amin, Ken Kurokawa, Naotake Ogasawara, Shigehiko Kanaya.

Elucidation of stage specific metabolites in *Escherichia coli* based on FT-ICR/MS and bioinformatics. **第2回日本ゲノム微生物学会年会** Osaka. March, 2007 (in Chapter 3)

3. **Hiroki Takahashi.** Integrative analysis of transcriptomics and metabolomics in *Escherichia coli* (in Japanese). **ゲノム微生物学会若手の会** Tokyo. November, 2008 (in Chapter 3, 4)

Books

1. Yoko Shinbo, **Hiroki Takahashi**, Kenichi Tanaka, Ryo Kusaba, Md. Altaf-Ul-Amin, Azziza Kawsar Parvin, Hiroko Asahi, Aki Hirai, Ken Kurokawa, Shigehiko Kanaya. Advanced Technology of Metabolomics and its Practical Application (CMC Publishing Co., Ltd.)

Acknowledgements

First, I’d like to acknowledge Professor Shigehiko Kanaya, who has been the adviser for the research described in this dissertation. He has supported me through many kinds of efforts for the accomplishment of this dissertation. I’m also thankful to Professor Naotake Ogasawara and Professor Kotaro Minato for their worthful guidances and this dissertation’s review.

I’m deeply grateful to Professor Daisaku Ohta of Osaka Prefecture University for his support of measuring FT-ICR/MS and Professor Takaaki Nishioka of Keio University for his support of advising about metabolite extraction. I’m also thanksful to Dr. Taku Oshima for advising about doing scientific research.

I’m highly thankful to Dr. Kosuke Kai for supporting in FT-ICR/MS measurements, Dr. Ryoko Morioka for LDS analysis, Mr. Ryosuke Ito for cDNA microarray experiments and metabolite sampling, Mrs. Kaori Igarashi for sampling procedure of metabolite, and Mr. Masayoshi Kuwano for advising about MS spectrometry.

I’d like to thank Associate Professor Md. Altaf-Ul-Amin for valuable comments and reviewing my papers.

I’d like to thank all my colleagues of Comparative genomics laboratory.

Finally, I’d like to thank my parents for supporting my education throughout my life.

Appendix

A. Lists of relations between ions and cluster IDs.

B. Gene lists of 75 KEGG pathways.

C. Gene lists of 99 TF regulated units.

D. Gene lists of 9 sigma factor regulated units.

E. Instructions of DrDMASS+ software.

Appendix A

Detected m/z	Theoretical m/z	Cluster ID	Isotopic ID	Difference	Molecular formula	Exact mass	Error	Candidates	Speices	Candidate ID
344.7563	345.7635	1-1								
450.2637	451.2710	1-1								
547.0756	548.0829	1-1	1	1.0027	$C_{16}H_{26}N_2O_{15}P_2$	548.0808	0.0020	dTDP-L-rhamnose	*Escherichia coli*	M-1
548.0783	549.0856	1-1	1	1.0027	$C_{16}H_{26}N_2O_{15}P_2$	548.0808	0.0020	dTDP-L-rhamnose	*Escherichia coli*	M-1
645.4527	646.4600	1-1								
687.4817	688.4890	1-1								
742.5399	743.5472	1-1								
860.2281	861.2354	1-1								
864.2546	865.2619	1-1								
686.4800	687.4873	1-1,1-2								
186.8848	187.8921	1-1,1-4								
660.4630	661.4703	1-2	2	1.0035						
661.4666	662.4738	1-2	2	1.0035						
691.4588	692.4660	1-2	3	1.0044	$C_{36}H_{69}O_{10}P$	692.4628	0.0032	PG5		
692.4631	693.4704	1-2	3	1.0044	$C_{36}H_{69}O_{10}P$	692.4628	0.0032	PG5		
693.4750	694.4822	1-2								
717.4737	718.4810	1-2	4	1.0073						
718.4810	719.4883	1-2	4	1.0073						
673.4643	674.4716	1-3	5	0.5040						
673.9683	674.9756	1-3	5	0.5040						
248.5080	249.5153	1-3	6		$C_{40}H_{75}O_{10}P$	746.5098		PG7		
372.7604	373.7677	1-3	6	0.5018	$C_{40}H_{75}O_{10}P$	746.5098		PG7		
373.2622	374.2695	1-3	6	0.5018	$C_{40}H_{75}O_{10}P$	746.5098		PG7		
745.5045	746.5118	1-3,1-4,	6	1.0028	$C_{40}H_{75}O_{10}P$	746.5098	0.0020	PG7		

		1-5								
746.5072	747.5145	1-4,1-5	6	1.0028	$C_{40}H_{75}O_{10}P$	746.5098	0.0020	PG7		
386.7779	387.7852	1-3								
786.4712	787.4785	1-3			$C_{41}H_{65}N_5O_{10}$	787.4731	0.0054	BE 32030B	*Nocardia sp.* *A32030*	M-2
773.5375	774.5448	1-4	7	1.0035	$C_{42}H_{79}O_{10}P$	774.5411	0.0037	PG9		
774.5410	775.5482	1-4	7	1.0035	$C_{42}H_{79}O_{10}P$	774.5411	0.0037	PG9		
775.5453	776.5526	1-4	7	1.0044	$C_{42}H_{79}O_{10}P$	774.5411	0.0037	PG9		
249.1802	250.1875	1-5	8	0.3344	$C_{40}H_{77}O_{10}P$	748.5254		PG3		
249.5146	250.5219	1-4,1-5	8	0.3344	$C_{40}H_{77}O_{10}P$	748.5254		PG3		
373.7679	374.7752	1-4,1-5	8	0.5019	$C_{40}H_{77}O_{10}P$	748.5254		PG3		
374.2697	375.2770	1-4,1-5	8	0.5019	$C_{40}H_{77}O_{10}P$	748.5254		PG3		
747.5183	748.5256	1-4,1-5	8	1.0044	$C_{40}H_{77}O_{10}P$	748.5254	0.0001	PG3		
748.5227	749.5300	1-4,1-5	8	1.0044	$C_{40}H_{77}O_{10}P$	748.5254	0.0001	PG3		
749.5249	750.5322	1-4	8	1.0026	$C_{40}H_{77}O_{10}P$	748.5254	0.0001	PG3		
750.5275	751.5348	1-4	8	1.0026	$C_{40}H_{77}O_{10}P$	748.5254	0.0001	PG3		
239.8353	240.8426	1-5,1-6	9	0.3348	$C_{38}H_{73}O_{10}P$	720.4941		PG1		
240.1701	241.1774	1-6	9	0.3348	$C_{38}H_{73}O_{10}P$	720.4941		PG1		
359.7514	360.7587	1-5,1-6	9	0.5018	$C_{38}H_{73}O_{10}P$	720.4941		PG1		
360.2532	361.2605	1-5	9	0.5018	$C_{38}H_{73}O_{10}P$	720.4941		PG1		
719.4868	720.4941	1-5,1-6	9	1.0048	$C_{38}H_{73}O_{10}P$	720.4941	0.0000	PG1		
720.4917	721.4990	1-5,1-6	9	1.0048	$C_{38}H_{73}O_{10}P$	720.4941	0.0000	PG1		
721.5007	722.5080	1-6	9	1.0090	$C_{38}H_{73}O_{10}P$	720.4941	0.0000	PG1		
179.8764	180.8837	1-6								
229.5043	230.5116	1-6								
344.2545	345.2618	1-6								
360.7582	361.7655	1-6								
690.5029	691.5102	1-6								
694.4790	695.4863	1-6								
714.5114	715.5187	1-6	10	1.0033						

715.5148	716.5220	1-6	10	1.0033						
183.3814	184.3887	2-1								
705.4757	706.4830	2-1	11	1.0051	$C_{37}H_{71}O_{10}P$	706.4785	0.0045	PG6		
706.4808	707.4881	2-1	11	1.0051	$C_{37}H_{71}O_{10}P$	706.4785	0.0045	PG6		
244.5091	245.5164	2-1	12	0.3348	$C_{39}H_{75}O_{10}P$	734.5098		PG2		
244.8439	245.8512	2-1	12	0.3348	$C_{39}H_{75}O_{10}P$	734.5098		PG2		
366.7615	367.7688	2-1	12	0.5020	$C_{39}H_{75}O_{10}P$	734.5098		PG2		
367.2635	368.2708	2-1	12	0.5020	$C_{39}H_{75}O_{10}P$	734.5098		PG2		
733.5056	734.5129	2-1	12	1.0032	$C_{39}H_{75}O_{10}P$	734.5098	0.0031	PG2		
734.5087	735.5160	2-1	12	1.0032	$C_{39}H_{75}O_{10}P$	734.5098	0.0031	PG2		
735.4982	736.5055	2-1	12	0.9895	$C_{39}H_{75}O_{10}P$	734.5098	0.0031	PG2		
297.8384	298.8456	2-1								
308.5415	309.5488	2-1								
253.8546	254.8619	2-1	13		$C_{41}H_{79}O_{10}P$	762.5411		PG4		
380.7791	381.7864	2-1	13	0.5024	$C_{41}H_{79}O_{10}P$	762.5411		PG4		
381.2815	382.2888	2-1	13	0.5024	$C_{41}H_{79}O_{10}P$	762.5411		PG4		
761.5293	762.5365	2-1	13	1.0047	$C_{41}H_{79}O_{10}P$	762.5411	0.0045	PG4		
762.5340	763.5412	2-1	13	1.0047	$C_{41}H_{79}O_{10}P$	762.5411	0.0045	PG4		
763.5494	764.5566	2-1	13	1.0154	$C_{41}H_{79}O_{10}P$	762.5411	0.0045	PG4		
606.6584	607.6657	2-1								
607.1607	608.1680	2-1								
618.0897	619.0970	2-1			$C_{17}H_{27}N_5O_{16}P_2$	619.0928	0.0042	ADP-L-glycero-beta-D-manno-heptopyranose	*Escherichia coli*	M-3
674.4814	675.4886	2-1								
707.4910	708.4983	2-1								
728.5293	729.5365	2-1								
759.5242	760.5315	2-1	14	0.9994	$C_{41}H_{77}O_{10}P$	760.5254	0.0060	PG8		
760.5236	761.5309	2-1,2-2	14	0.9994	$C_{41}H_{77}O_{10}P$	760.5254	0.0060	PG8		

787.5556	788.5628	2-1			$C_{43}H_{81}O_{10}P$	788.5567	0.0061	PG10		
397.1092	398.1165	2-1,2-2								
309.5084	310.5157	2-2								
396.7746	397.7819	2-2								
742.0735	743.0808	2-2								
595.6651	596.6724	2-2,2-3	15	0.5015						
596.1667	597.1739	2-3	15	0.5015						
596.6681	597.6753	2-3	15	0.5014						
143.1080	144.1153	3			$C_8H_{16}O_2$	144.1150	0.0003	Octanoic acid	*Escherichia coli*	M-4
273.0346	274.0419	3								
273.5383	274.5456	3								
282.0215	283.0287	3								
302.5353	303.5426	3								
321.0506	322.0579	3			$C_{10}H_{15}N_2O_8P$	322.0566	0.0013	dTMP	*Escherichia coli K12*	M-5
359.7333	360.7405	3								
507.1397	508.1470	3								
563.8505	564.8578	3								
565.0503	566.0576	3			$C_{15}H_{24}N_2O_{17}P_2$	566.0550	0.0025	UDP-D-glucose UDP-D-galactose	*Escherichia coli*	M-6
587.0306	588.0378	3								
606.0775	607.0848	3			$C_{17}H_{27}N_3O_{17}P_2$	607.0816	0.0032	UDP-*N*-acetyl-D-mannosamine UDP-*N*-acetyl-D-glucosamine	*Escherichia coli*	M-7

628.0576	629.0649	3								
672.3815	673.3888	3								
687.9839	688.9911	3								
178.9633	179.9706	4								
288.1211	289.1284	4								
339.0471	340.0544	4								
367.7652	368.7725	4								
397.4434	398.4507	4								
401.0168	402.0241	4			$C_{10}H_{16}N_2O_{11}P_2$	402.0229	0.0012	dTDP	*Escherichia coli*	M-8
417.0807	418.0880	4								
464.2796	465.2869	4								
495.1039	496.1112	4			$C_{24}H_{20}N_2O_{10}$	496.1118	0.0006	Kinamycin A Kinamycin C	*Streptomyces murayamaensis sp. nov.*	M-9
505.9908	506.9981	4			$C_{10}H_{16}N_5O_{13}P_3$	506.9957	0.0023	ATP dGTP	*Escherichia coli*	M-10
626.1268	627.1341	4								
627.6249	628.6321	4								
725.3865	726.3938	4								
731.4910	732.4983	4								
755.4897	756.4970	4								
253.2137	254.2210	5			$C_{16}H_{30}O_2$	254.2246	0.0036	omega-Cycloheptanenonanoic acid	*Alicyclobacillus acidocaldarius*	M-11
281.2444	282.2516	5			$C_{18}H_{34}O_2$	282.2559	0.0042	Oleic acid cis-11-Octadecanoic acid omega-Cycloheptylunde	*Escherichia coli* *Lactobacillus plantarum* *Alicyclobacillus acidocaldarius*	M-12

								canoic acid		
366.6251	367.6323	5								
374.7723	375.7796	5								
517.9586	518.9659	5								
711.8751	712.8823	5								
737.4933	738.5006	5								
752.4853	753.4925	5								
777.5083	778.5156	5	16	0.9973						
778.5056	779.5128	5	16	0.9973						
779.5014	780.5086	5	16	0.9958						
789.5380	790.5453	5								
220.7068	221.7141	6								
426.0237	427.0310	6			$C_{10}H_{15}N_5O_{10}P_2$	427.0294	0.0016	Adenosine 3',5'-bisphosphate ADP dGDP	*Escherichia coli*	M-13
540.0557	541.0630	6								
331.0586	332.0659	6	17		$C_{21}H_{27}N_7O_{14}P_2$	663.1091		NAD	*Escherichia coli*	M-14
662.1037	663.1109	6	17	1.0044	$C_{21}H_{27}N_7O_{14}P_2$	663.1091	0.0018	NAD	*Escherichia coli*	M-14
663.1080	664.1153	6	17	1.0044	$C_{21}H_{27}N_7O_{14}P_2$	663.1091	0.0018	NAD	*Escherichia coli*	M-14
626.6241	627.6314	7	18	0.4990						
627.1231	628.1304	7	18	0.4990						
808.9615	809.9688	7								
402.9962	404.0035	8			$C_9H_{14}N_2O_{12}P_2$	404.0022	0.0013	UDP	*Escherichia coli*	M-15
535.0862	536.0935	8								
664.1095	665.1168	8			$C_{21}H_{29}N_7O_{14}P_2$	665.1248	0.0080	NADH	*Escherichia coli*	M-16
180.1271	181.1344	9								
438.5855	439.5928	9								
772.4812	773.4885	9								

763.5153	764.5226	10	19	1.0009						
764.5162	765.5235	10	19	1.0009						
382.5525	383.5598	11								
454.0391	455.0464	11			$C_{20}H_{19}Cl_2NO_7$	455.0539	0.0075	Antibiotic MI 178-34F18A2 Antibiotic MI 178-34F18C2	*Actinomadura spiralis MI178-34F18* *Actinomadura spiralis MI178-34F18*	M-17
72.9878	73.9951	-			$C_2H_2O_3$	74.0004	0.0053	Glyoxylic acid	*Escherichia coli*	M-18
85.0774	86.0847	-								
171.1013	172.1085	-								
199.1676	200.1749	-								
221.8013	222.8086	-								
234.1784	235.1856	-								
240.5056	241.5129	-								
241.2141	242.2214	-								
250.1419	251.1492	-								
253.2185	254.2258	-			$C_{16}H_{30}O_2$	254.2246	0.0012	omega-Cycloheptanenonanoic acid	*Alicyclobacillus acidocaldarius*	M-19
256.2350	257.2422	-								
284.2678	285.2751	-								
297.2410	298.2482	-			$C_{18}H_{34}O_3$	298.2508	0.0026	alpha-Cycloheptaneundecanoic acid	Alicyclobacillus acidocaldarius	M-20
297.2467	298.2540	-			$C_{18}H_{34}O_3$	298.2508	0.0032	alpha-Cyclo	*Alicyclobacillus*	M-21

								heptaneundecanoic acid	*acidocaldarius*	
297.2516	298.2589	-			$C_{18}H_{34}O_3$	298.2508	0.0081	alpha-Cycloheptaneundecanoic acid	*Alicyclobacillus acidocaldarius*	M-22
299.2605	300.2678	-								
310.7338	311.7410	-								
312.7327	313.7399	-								
314.7306	315.7379	-								
338.5455	339.5528	-								
346.0570	347.0643	-			$C_{10}H_{14}N_5O_7P$	347.0631	0.0012	AMP 3'-AMP dGMP	*Escherichia coli*	M-23
358.7457	359.7530	-								
363.5739	364.5812	-								
369.3025	370.3098	-								
383.3194	384.3266	-								
403.5576	404.5648	-								
403.6630	404.6703	-								
409.2364	410.2436	-								
414.6508	415.6580	-								
425.3658	426.3731	-								
425.6668	426.6741	-								
429.0263	430.0336	-								
452.2799	453.2872	-								
453.3973	454.4046	-								
457.3200	458.3273	-								
457.7755	458.7828	-								
458.0933	459.1005	-								
458.1112	459.1185	-			$C_{15}H_{22}N_7O_8P$	459.1267	0.0083	Phosmidosin	*Streptomyces sp.*	M-24

								e B	*strain RK-16*	
493.6398	494.6471	-								
499.3673	500.3746	-								
515.9629	516.9702	-								
533.6234	534.6307	-								
563.3385	564.3458	-								
565.3207	566.3280	-								
569.0570	570.0643	-								
579.3351	580.3423	-								
580.3373	581.3446	-								
659.4702	660.4775	-								
663.4834	664.4907	-								
665.4413	666.4486	-								
672.0622	673.0695	-								
672.8809	673.8882	-								
674.4098	675.4171	-								
679.3691	680.3763	-								
722.5075	723.5148	-								
723.5126	724.5199	-								
728.1586	729.1659	-								
732.5021	733.5094	-								
736.4996	737.5069	-								
741.4729	742.4801	-			$C_{32}H_{62}N_{12}O_8$	742.4814	0.0012	Argimicin A	*Sphingomonas sp.*	M-25
751.4800	752.4873	-								
758.0944	759.1017	-								
765.4990	766.5063	-								
788.5285	789.5358	-								
808.6293	809.6366	-								
809.2969	810.3042	-								
853.3166	854.3239	-			$C_{41}H_{46}N_{10}O_9S$	854.3170	0.0069	Argyrin G	*Archangium*	M-26

									gephyra Ar 8082	
					$C_{45}H_{56}Cl_2N_2O_{10}$	854.3312	0.0073	Decatromicin B	*Actinomadura sp. MK73-NF4*	
					$C_{39}H_{50}N_8O_{12}S$	854.3269	0.0030	Napsamycin C	*Streptomyces sp. HIL Y-82,11372*	
860.7291	861.7364	-								
861.2335	862.2408	-								
890.3419	891.3491	-								
891.3315	892.3388	-								
941.2571	942.2644	-								

Appendix B

eco00010	Glycolysis / Gluconeogenesis	*aceE, aceF, acs, adhE, adhP, agp, ascB, ascF, bglA, bglB, chbF, crr, eno, fbaA, fbp, frmA, galM, gapA, glk, glpX, glvC, gpmA, gpmM, lpd, malX, pfkA, pfkB, pgi, pgk, pgm, ptsG, pykA, pykF, tpiA, yccX, ytjC*
eco00020	Citrate cycle (TCA cycle)	*acnA, acnB, citD, cite, citF, frdA, frdB, frdC, frdD, fumA, fumB, fumC, gltA, icd, lpd, mdh, pck, sdhA, sdhB, sdhC, sdhD, sucA, sucB, sucC, sucD, ybhJ, ybiC*
eco00030	Pentose phosphate pathway	*deoB, deoC, eda, edd, fbaA, fbp, gcd, glpX, gnd, gntK, idnK, kdgK, pfkA, pfkB, pgi, pgl* *pgm, prs, rbsK, rpe, rpiA, rpiB, talA, talB, tktA, tktB, zwf*
eco00040	Pentose and glucuronate interconversions	*araA, araB, araD, eda, galF, galU, kdgK, kduD, kduI, lyx, rhaB, rhaD, rpe, sgbE, sgbH, sgbU, ugd, uidA, uxaA, uxaB, uxaC, uxuA, xylA, xylB, yiaK*
eco00051	Fructose and mannose metabolism	*aceK, cmtA, cmtB, cpsB, cpsG, fbaA, fbp, fcl, fruA, fruB, fruK, frvA, frvB, frwC, frwD, fryA, fryB, fryC, fucA, fucI, fucK, glpX, gmd, manA, manX, manY, manZ, mtlA, mtlD, pfkA, pfkB, ptsA, rffT, rhaA, rhaB, rhaD, srlA, srlB, srlD, srlE, tpiA, xylA, yfdH, yniC*
eco00052	Galactose metabolism	*agaB, agaC, agaD, agaI, agaV, dgoA, dgoK, ebgA, galE, galF, galK, galT, galU, gatA, gatC, gatD, gatY, gatZ, glk, kbaY, kbaZ, lacZ, malZ, melA, pfkA, pfkB, pgm, sgcA, sgcC*
eco00053	Ascorbate and aldarate metabolism	*garD, garL, gudD, gudX, lyx, ugd, ulaB, ulaC, ulaD, ulaE, ulaF, ulaG, yadI, yiaK*
eco00061	Fatty acid biosynthesis	*accA, accB, accC, accD, fabA, fabB, fabD, fabF, fabG, fabH, fabI, fabZ*
eco00071	Fatty acid metabolism	*aas, adhE, adhP, atoB, fadA, fadB, fadD, fadE, fadI, fadJ, frmA, hcaD, paaF, paaG*
eco00130	Ubiquinone biosynthesis	*entC, menA, menB, menC, menD, menE, menF, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, ubiA, ubiB, ubiD, ubiE, ubiF, ubiG, ubiH, ubiX, yfbB*

eco00190	Oxidative phosphorylation	*appB, appC, atpA, atpB, atpC, atpD, atpE, atpF, atpG, atpH, cydA, cydB, cyoA, cyoB, cyoC, cyoD, cyoE, frdA, frdB, frdC, frdD, ndh, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, ppa, ppk, sdhA, sdhB, sdhC, sdhD*
eco00220	Urea cycle and metabolism of amino groups	*adiA, argA, argB, argC, argD, argE, argF, argG, argH, argI, mtn, pepD, proA, proB, puuA, puuB, puuC, puuD, speA, speB, speC, speD, speE, speF, speG, ydcW, ygjG*
eco00230	Purine metabolism	*add, ade, adk, allA, allB, allC, allD, amn, apaH, apt, cpdB, cyaA, cysC, cysD, cysN, deoA, deoB, deoD, dgt, dnaE, dnaN, dnaQ, dnaX, gmk, gpp, gpt, gsk, guaA, guaB, guaC, guaD, holA, holB, holC, holD, holE, hpt, mazG, ndk, nrdA, nrdB, nrdD, nrdE, nrdF, nudF, pnp, polA, prs, purA, purB, purC, purD, purE, purF, purH, purK, purL, purM, purN, purT, pykA, pykF, rdgB, relA, rihB, rihC, spoT, surE, xdhA, xdhB, yagR, yahI, ybcF, yfbR, yjjG, yqeA*
eco00240	Pyrimidine metabolism	*carA, carB, cdd, cmk, codA, cpdB, dcd, deoA, deoD, dnaE, dnaN, dnaQ, dnaX, dut, holA, holB, holC, holD, holE, mazG, ndk, nrdA, nrdB, nrdD, nrdE, nrdF, pnp, polA, pyrB, pyrC, pyrD, pyrE, pyrF, pyrG, pyrH, pyrI, rihB, surE, tdk, thyA, tmk, trxB, udk, udp, upp, yfbR, yjjG*
eco00251	Glutamate metabolism	*adiA, aspC, carA, carB, gabD, gabT, gadA, gadB, gdhA, glmS, glnA, glnS, gltB, gltD, gltX, gor, gshA, gshB, guaA, murI, nadE, nagK, purF, putA, puuE, speA, yahI, ybaS, ybcF, ybdK, yneH, yqeA*
eco00252	Alanine and aspartate metabolism	*aceE, aceF, alaS, alr, ansA, ansB, argG, argH, asnA, asnB, asnS, aspA, aspC, aspS, dadX, gabT, gadA, gadB, iaaA, lpd, nadB, panD, pepD, purA, purB, puuE, pyrB, pyrI*
eco00260	Glycine, serine and threonine metabolism	*asd, betA, betB, dsdA, garK, gcvP, gcvT, glxK, glyA, glyQ, glyS, ilvA, kbl, lpd, ltaE, lysC, metL, psd, pssA, sdaA, sdaB, serA, serB, serC, serS, tdcB, tdcG, tdh, thrA, thrB, thrC, thrS, tynA, usg*
eco00271	Methionine metabolism	*dcm, fmt, luxS, malY, metA, metB, metC, metE, metG, metH, metK, mmuM, mtn, speD, speE, tyrB*
eco00272	Cysteine metabolism	*aspC, cysE, cysK, cysM, cysS, dcyD, malY, metB, metC, sdaA, sdaB, sseA, tdcG*

eco00280	Valine, leucine and isoleucine degradation	*atoB, fadA, fadB, fadI, fadJ, gabT, ilvE, lpd, paaF, paaG, scpA*
eco00290	Valine, leucine and isoleucine biosynthesis	*aceE, avtA, ileS, ilvA, ilvB, ilvC, ilvD, ilvE, ilvH, ilvI, ilvM, ilvN, leuA, leuB, leuD, leuS, tdcB, valS*
eco00300	Lysine biosynthesis	*argD, asd, dapA, dapB, dapD, dapE, dapF, lysA, lysC, lysS, lysU, metL, murE, murF, poxA, thrA, usg, yagE*
eco00310	Lysine degradation	*atoB, cadA, fadB, fadJ, ldcC, paaF, paaG, rzoD, rzoR, rzpD, rzpR, sucA, sucB*
eco00330	Arginine and proline metabolism	*argF, argG, argH, argI, argS, aspC, astA, astB, astC, astD, astE, eda, proC, proS, putA, yahI, ybcF, yqeA*
eco00340	Histidine metabolism	*hisA, hisB, hisC, hisD, hisF, hisG, hisH, hisI, hisS, pepD, rsmF, tynA*
eco00350	Tyrosine metabolism	*adhE, adhP, aspC, frmA, gabD, hisC, pagP, rsmF, tynA, tyrB, ydcK*
eco00360	Phenylalanine metabolism	*aspC, dadA, hcaB, hcaE, hcaF, hisC, mhpC, mhpD, mhpE, paaK, pagP, tynA, tyrB, ydcK*
eco00380	Tryptophan metabolism	*atoB, fadB, fadJ, katE, katG, paaF, paaG, rsmF, sucA, tnaA, trpS, tynA*
eco00400	Phenylalanine, tyrosine and tryptophan biosynthesis	*aroA, aroB, aroC, aroD, aroE, aroF, aroG, aroH, aroK, aroL, aspC, hisC, pheA, pheS, pheT, trpA, trpB, trpC, trpD, trpE, tyrA, tyrB, tyrS, ydiB*
eco00410	beta-Alanine metabolism	*fadB, fadJ, gabT, gadA, gadB, paaF, paaG, panC, panD, pepD, puuE, speE, tynA, ydcW*
eco00450	Selenoamino acid metabolism	*cysC, cysD, cysK, cysM, cysN, ggt, malY, metB, metC, metG, metK, rsmF, selA, selD, sufS*
eco00480	Glutathione metabolism	*btuE, ggt, gor, gshA, gshB, gsp, gst, icd, pepN, pepT, ybdK, yncG, zwf*
eco00500	Starch and sucrose metabolism	*amyA, bcsA, bcsZ, bglX, galF, galU, glgA, glgB, glgC, glgP, glk, malP, malQ, malS, malX, malZ, murP, otsA, otsB, pgi, pgm, rhlE, treA, treB, treC, treF, ugd, uidA, yagH, ycjM, ycjU, yoaA*

eco00520	Nucleotide sugars metabolism	*arnA, arnB, arnC, galE, galF, galT, galU, rfbA, rfbB, rfbC, rfbD, rffG, rffH, ugd, yagH*
eco00530	Aminosugars metabolism	*aceK, chiA, glmM, glmS, glmU, murA, murB, nagA, nagB, nagE, nagK, nagZ, nanE, nanK, rffD, rffE, yniC*
eco00540	Lipopolysaccharide biosynthesis	*diaA, gmhB, kdsA, kdsB, kdsC, lpcA, lpxA, lpxB, lpxC, lpxD, lpxH, lpxK, rfaB, rfaC, rfaE, rfaF, rfaG, rfaI, rfaJ, rfaL, rfaP, rfaQ, rfaY, waaA, waaU*
eco00550	Peptidoglycan biosynthesis	*amiA, amiB, amiC, amiD, bacA, ddlA, ddlB, ftsI, glnA, mraY, mrcB, murC, murD, murE, murF, murG, ybjG*
eco00561	Glycerolipid metabolism	*dgkA, dhaK, dhaL, ebgA, garK, gldA, glpK, glxK, lacZ, mdoB, melA, plsB, plsC*
eco00564	Glycerophospholipid metabolism	*aas, cdh, cdsA, cls, dgkA, eutA, eutB, eutC, glpA, glpB, glpC, glpD, glpQ, gpsA, pagP, pgpA, pgpB, pgsA, pldA, plsB, plsC, psd, pssA, ybhO, ydcK, ynbB*
eco00620	Pyruvate metabolism	*accA, accB, accC, accD, aceB, aceE, aceF, ackA, acs, adhE, aldA, aldB, atoB, dld, fucO, ghrA, glcB, gloA, gloB, leuA, lldD, lpd, maeA, maeB, mdh, mgsA, mhpF, mqo, pck, pflB, pflD, poxB, ppc, pps, pta, pykA, pykF, tdcE, ybiC, ybiW, yccX*
eco00630	Glyoxylate and dicarboxylate metabolism	*aceA, aceB, acnA, acnB, aldA, eda, fdnG, fdnH, fdnI, fdoG, fdoH, fdoI, folD, fucO, garK, garR, gcl, ghrA, glcB, gltA, glxK, glxR, gph, hyi, mdh, oxc, purU, ttdA, ttdB, ybhJ, ybiC, ydeP, yeaU*
eco00632	Benzoate degradation via CoA ligation	*atoA, atoB, atoD, fadB, fadJ, fadK, frdA, frdB, frdC, frdD, lsrK, mak, paaF, paaG, paaH, pagP, sdhA, sdhB, sdhC, sdhD, yccX, ydcK*
eco00640	Propanoate metabolism	*accA, accB, accC, accD, ackA, acs, atoA, atoB, atoD, fadB, fadJ, fadK, gabT, paaF, paaG, pflB, pflD, prpB, prpC, prpD, prpE, pta, puuE, scpA, scpB, sucC, sucD, tdcD, tdcE, ybiW*
eco00650	Butanoate metabolism	*aceE, adhE, atoA, atoB, atoD, fadB, fadJ, frdA, frdB, frdC, frdD, gabD, gabT, gadA, gadB, ilvB, ilvH, ilvI, ilvM, ilvN, mhpF, paaF, paaG, paaH, pflB, pflD, puuE, sdhA, sdhB, sdhC, sdhD, tdcE, ybiW, yeaU*
eco00670	One carbon pool by folate	*fmt, folA, folD, folM, gcvT, glyA, metF, metH, purH, purN, purT, purU, thyA*

eco00680	Methane metabolism	*fdnG, fdnH, fdnI, fdoG, fdoH, fdoI, frmA, glyA, katE, katG, metF, ydeP*
eco00710	Carbon fixation	*aspC, fbaA, fbp, glpX, maeB, mdh, pck, pgk, ppc, prkB, pykA, pykF, rpe, rpiA, rpiB, tktA, tktB, tpiA, ybiC*
eco00720	Reductive carboxylate cycle (CO2 fixation)	*acnA, acnB, acs, frdA, frdB, frdC, frdD, fumA, fumB, fumC, icd, mdh, ppc, pps, sdhA, sdhB, sdhC, sdhD, sucC, sucD, ybhJ, ybiC*
eco00730	Thiamine metabolism	*aceK, iscS, rdgB, sufS, thiC, thiD, thiE, thiF, thiG, thiH, thiI, thiK, thiL, thiM, yniC*
eco00740	Riboflavin metabolism	*aceK, aphA, appA, cobT, ribA, ribC, ribD, ribE, ribF, yniC*
eco00760	Nicotinate and nicotinamide metabolism	*nadA, nadB, nadC, nadD, nadE, nadK, nadR, pncA, pncB, rihC, sthA*
eco00770	Pantothenate and CoA biosynthesis	*acpH, coaA, coaD, coaE, dfp, ilvB, ilvC, ilvD, ilvE, ilvH, ilvI, ilvM, ilvN, panB, panC, panE*
eco00790	Folate biosynthesis	*folA, folB, folC, folE, folK, folM, pabA, pabB, pabC, phoA, rhlE, sscR, yoaA*
eco00860	Porphyrin and chlorophyll metabolism	*btuR, cobC, cobS, cobT, cobU, cyoE, cysG, eutT, fre, gltX, hemA, hemB, hemC, hemD, hemE, hemF, hemG, hemH, hemL, hemN, hemX, uidA, yggW*
eco00910	Nitrogen metabolism	*ansA, ansB, asnA, asnB, aspA, can, cynS, cynT, dadA, gcvT, gdhA, glnA, gltB, gltD, iaaA, malY, metC, napA, narG, narH, narI, narJ, narV, narW, narY, narZ, nirB, nirD, nrfA, tnaA, yahI, ybaS, ybcF, yneH, yqeA*
eco00920	Sulfur metabolism	*cysC, cysD, cysE, cysH, cysI, cysJ, cysK, cysM, cysN, malY, metA, metB, metC*
eco00970	Aminoacyl-tRNA biosynthesis	*alaS, argS, asnS, aspS, cysS, fmt, glnS, gltX, glyQ, glyS, hisS, ileS, leuS, lysS, lysU, metG, pheS, pheT, poxA, proS, serS, thrS, trpS, tyrS, valS*
eco00983	Drug metabolism - other enzymes	*cdd, deoA, guaA, guaB, hpt, pyrE, tdk, udk, udp, uidA*
eco01031	Glycan structures - biosynthesis 2	*rfaB, rfaC, rfaF, rfaG, rfaI, rfaJ, rfaP, rfaQ, rfaY, waaA, waaU*

eco02010	ABC transporters - General	*afuB, afuC, alsA, alsB, alsC, araF, araG, araH, argT, artI, artJ, artM, artP, artQ, btuC, btuD, btuF, ccmA, ccmB, ccmC, cydC, cydD, cysA, cysP, cysU, ddpA, ddpB, ddpC, ddpD, ddpF, dppA, dppB, dppC, dppD, dppF, fecB, fecC, fecD, fecE, fepB, fepC, fepD, fepG, fhuB, fhuC, fhuD, fliY, ftsX, glnH, glnP, glnQ, gltI, gltJ, gltK, gltL, gsiA, gsiB, gsiC, gsiD, hisJ, hisM, hisP, hisQ, livF, livG, livH, livJ, livK, livM, lolC, lolD, lolE, lsrA, lsrB, lsrC, lsrD, macB, malE, malF, malG, malK, mdlB, metI, metN, metQ, mglA, mglB, mglC, modA, modB, modC, modF, mppA, msbA, nikC, nikD, nikE, nlpA, oppA, oppB, oppC, oppD, oppF, osmF, phnC, phnD, phnE, phnK, potA, potB, potC, potD, potF, potG, potH, potI, proV, proW, proX, pstA, pstB, pstC, pstS, rbsA, rbsB, rbsC, rbsD, sapA, sapB, sapC, sapD, sapF, sbp, ssuB, ssuC, tauA, tauB, tauC, tbpA, thiP, thiQ, ugpA, ugpC, xylG, yadG, yadH, ybbA, ybbL, ybbM, ybbP, ybhR, yddA, yecC, yecS, yehW, yehY, yejA, yejB, yejE, yejF, ygiS, yhdW, yhdX, yhdY, yhdZ, yhhJ, ynjB, ynjC, ynjD, yojI, yrbC, yrbE, yrbF, znuA, znuB, znuC*
eco02011	ABC transporters - Organism-specific	*afuB, afuC, alsA, alsB, alsC, araF, araG, araH, argT, artI, artJ, artM, artP, artQ, btuC, btuD, btuE, ccmA, ccmB, ccmC, ccmE, cysA, cysP, cysU, ddpA, ddpB, ddpC, ddpD, ddpF, dppA, dppB, dppC, dppD, dppF, fecA, fecB, fecC, fecD, fecE, fepA, fepB, fepC, fepD, fepG, fhuA, fhuB, fhuC, fhuD, fliY, glnH, glnP, glnQ, gltI, gltJ, gltK, gltL, gsiA, gsiB, gsiC, gsiD, hisJ, hisM, hisP, hisQ, livF, livG, livH, livJ, livK, livM, lsrA, lsrB, lsrC, lsrD, malE, malF, malG, malK, malM, metI, metN, metQ, mglA, mglB, mglC, modA, modB, modC, modF, mppA, nikC, nikD, nikE, oppA, oppB, oppC, oppD, oppF, osmF, phnC, phnD, phnE, potA, potB, potC, potD, potF, potG, potH, potI, proV, proW, proX, pstA, pstB, pstC, pstS, rbsA, rbsB, rbsC, rbsD, sapA, sapB, sapC, sapD, sapF, sbp, ssuB, ssuC, ssuD, ssuE, tauA, tauB, tauC, tauD, tbpA, thiP, thiQ, ugpA, ugpC, xylG, ybbL, ybbM, ycjN, ycjO, ydcS, ydcT, ydcU, ydcV, yecC, yecS, yehW, yehY, yejA, yejB, yejE, yejF, ygiS, yhdW, yhdX, yhdY, yhdZ, yjfF, ynjB, ynjC, ynjD, yrbE, yrbF, ytfQ, ytfR, ytfT, znuA, znuB, znuC*
eco02020	Two-component	*aer, ampC, ampH, appY, arcA, arcB, arnB, atoA, atoB, atoC, atoD, atoE,*

	system - General	*atoS, baeR, baeS, barA, basR, basS, cheA, cheB, cheW, cheY, citC, citD, citE, citF, citG, citT, citX, cpxA, cpxR, creB, creC, csrA, cusR, cusS, dcuR, dcuS, degP, dpiA, dpiB, emrK, emrY, envZ, evgA, evgS, fdnG, fdnH, fdnI, fimZ, flhC, flhD, fliA, fliC, frdA, frdB, frdC, frdD, glnA, glnB, glnD, glnG, glnL, kdpA, kdpB, kdpC, kdpD, kdpE, mdtA, mdtB, mdtC, motA, narG, narH, narI, narJ, narL, narP, narQ, narX, ompC, ompF, ompR, phoA, phoB, phoP, phoQ, phoR, qseB, qseC, rcsA, rcsB, rcsC, rcsD, rcsF, rstA, rstB, sdiA, tap, tar, torA, torC, torD, torR, torS, trpA, trpB, trpC, trpD, trpE, trpL, tsr, uhpA, uhpB, uhpC, uhpT, uvrY, yedV, yedW, yehT, yehU, yfhA, yfhK, ypdA, zraR, zraS*
eco02021	Two-component system - Organism-specific	*appA, arcA, arcB, atoC, atoS, baeR, baeS, barA, basR, basS, cpxA, cpxR, creB, creC, cusR, cusS, dcuR, dcuS, dpiA, dpiB, envZ, evgA, evgS, fdnG, fdnH, fdnI, fimZ, frdA, frdB, frdC, frdD, ftsZ, glnA, glnB, glnD, glnG, glnL, kdpA, kdpB, kdpC, kdpD, kdpE, narG, narH, narI, narJ, narL, narP, narQ, narX, ompC, ompF, ompR, phoA, phoB, phoP, phoQ, phoR, qseB, qseC, rcsA, rcsB, rcsC, rcsD, rssB, rstA, rstB, torA, torC, torD, torR, torS, uhpA, uhpB, uhpC, uhpT, uvrY, yedV, yedW, yehT, yehU, yfhA, yfhK, ypdA, ypdB, zraR, zraS*
eco02030	Bacterial chemotaxis - General	*aer, cheA, cheB, cheR, cheW, cheY, cheZ, fliG, fliM, fliN, lafU, motA, tap, tar, tsr*
eco02031	Bacterial chemotaxis - Organism-specific	*aer, cheA, cheB, cheR, cheW, cheY, cheZ, dppA, fliG, fliM, fliN, lafU, malE, mglB, motA, rbsB, tap, tar, tsr*
eco02040	Flagellar assembly	*flgA, flgC, flgD, flgE, flgF, flgG, flgH, flgI, flgK, flgL, flgM, flgN, flhA, flhB, flhC, flhD, fliC, fliD, fliE, fliF, fliG, fliH, fliI, fliJ, fliK, fliM, fliN, fliP, fliQ, fliR, fliS, fliT, lafU, lfhA, motA*
eco02060	Phosphotransferase system (PTS)	*agaB, agaC, agaD, agaV, ascF, bglF, chbB, chbC, cmtA, cmtB, crr, fruA, fruB, frvA, frvB, frwC, frwD, fryA, fryB, fryC, gatA, gatC, glvC, malX, manX, manY, manZ, mtlA, murP, nagE, npr, ptsA, ptsG, ptsH, ptsI, ptsN, ptsP, sgcA, sgcC, srlA, srlB, srlE, treB, ulaB, ulaC, yadI*
eco03010	Ribosome	*rplA, rplB, rplC, rplD, rplE, rplF, rplI, rplJ, rplK, rplL, rplM, rplN, rplO, rplP, rplQ, rplR, rplS, rplT, rplU, rplV, rplW, rplX, rplY, rpmA, rpmB,*

		rpmC, rpmD, rpmE, rpmF, rpmG, rpmH, rpmI, rpmJ, rpsA, rpsB, rpsC, rpsD, rpsE, rpsF, rpsG, rpsH, rpsI, rpsJ, rpsK, rpsL, rpsM, rpsN, rpsO, rpsP, rpsQ, rpsR, rpsS, rpsT, rpsU, ykgM
eco03030	DNA replication	*dnaB, dnaE, dnaG, dnaN, dnaQ, dnaX, holA, holB, holC, holD, holE, ligA, ligB, polA, rnhB, ssb*
eco03060	Protein export	*ffh, ftsY, lepB, lspA, secA, secB, secD, secE, secF, secG, secY, tatA, tatB, tatC, tatE, yajC, yidC*
eco03070	Type III secretion system	*flhA, flhB, fliF, fliH, fliI, fliN, fliP, fliQ, fliR, lfhA*
eco03090	Type II secretion system	*gspA, gspC, gspD, gspE, gspF, gspH, gspI, gspJ, gspK, gspL, gspM, gspO, hofC, hofQ, ppdA, ppdB, ppdC, ppdD, pppA, yghD, yghE, yghF*
eco03410	Base excision repair	*alkA, ligA, ligB, mug, mutM, mutY, nei, nfo, nth, polA, recJ, tag, ung, xthA*
eco03430	Mismatch repair	*dam, dnaE, dnaN, dnaQ, dnaX, exoX, holA, holB, holC, holD, holE, ligA, ligB, mutH, mutL, mutS, recJ, sbcB, ssb, uvrD, xseA, xseB*
eco03440	Homologous recombination	*dnaE, dnaN, dnaQ, dnaT, dnaX, holA, holB, holC, holD, holE, polA, priA, priB, priC, recA, recB, recC, recD, recF, recG, recJ, recO, recR, ruvA, ruvB, ruvC, ssb*

Appendix C

AgaR	*agaA, agaB, agaC, agaD, agaI, agaR, agaS, agaV, agaW, kbaY, kbaZ*
AllR	*allA, allB, allS, gcl, glxK, glxR, hyi, ybbW, ybbY*
AppY	*appA, appB, appC, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF*
AraC	*araA, araB, araC, araD, araE, araF, araG, araH, araJ*
ArcA	*aceA, aceB, aceE, aceF, aceK, ackA, acnA, acnB, aldA, appA, appB, appC, betA, betB, betI, betT, cadA, cadB, caiA, caiB, caiC, caiE, caiT, cydA, cydB, cydC, cydD, cyoA, cyoB, cyoC, cyoD, cyoE, dctA, dcuC, fadA, fadB, fadD, fadE, fadI, fadJ, fadL, fnr, focA, fumA, fumB, fumC, gadA, gadB, gadX, gatA, gatC, gatD, gatY, gatZ, glcA, glcB, glcD, glcE, glcG, glpA, glpB, glpC, glpD, gltA, hemA, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF, hybA, hybB, hybC, hybD, hybE, hybF, hybG, hybO, icd, lldD, lldP, lldR, lpd, mdh, moeA, moeB, ndh, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, oppA, oppB, oppC, oppD, oppF, pflB, prfA, prmC, ptsG, rhaT, rplB, rplC, rplD, rplP, rplV, rplW, rpmC, rpoS, rpsC, rpsJ, rpsQ, rpsS, rutA, rutB, rutD, rutE, rutG, sdhA, sdhB, sdhC, sdhD, sodA, ssb, sucA, sucB, sucC, sucD, tpx, treB, treC, ubiA, uvrA, xylR, ydeA, yfiD*
ArgP	*dnaA, dnaN, nrdA, nrdB, recF*
ArgR	*argA, argB, argC, argD, argE, argF, argG, argH, argI, argR, artI, artJ, artM, artP, artQ, astA, astB, astC, astD, astE, carA, carB, gltB, gltD, gltF, hisJ, hisM, hisP, hisQ, infB, nusA, pnp, rbfA, rpsO, truB, yhbC*
BaeR	*acrD, baeR, baeS, mdtA, mdtB, mdtC, mdtD, spy, ycaC*
BirA	*bioA, bioB, bioC, bioD, bioF*
CaiF	*caiA, caiB, caiC, caiE, caiT, fixC, fixX*
Cbl	*ssuB, ssuC, ssuD, ssuE, tauA, tauB, tauC, tauD*
CdaR	*cdaR, garD, garK, garL, garR, gudD, gudP, gudX*
ChbR	*chbB, chbC, chbF, chbG, chbR*
CpxR	*acrD, aroG, bacA, baeR, baeS, cheA, cheW, cpxA, cpxP, cpxR, csgA, csgC, csgD, csgE, csgF, csgG, degP, dsbA, dsbC, fabZ, ftnB, hha, lpxA, lpxD, mdtA, mdtB, mdtC, mdtD, motA, ompC, ompF, ppiA, ppiD, psd, rdoA, rpoE, rpoH, rseA, rseB, rseC, skp, spy, tsr, ung, ybaJ, yccA, ydeH, yebE, yidQ, yjeP, yqjA, yqjB*
CRP	*aceA, aceB, aceE, aceF, aceK, acnA, acnB, acs, actP, aer, agaA, agaV, agaW, agp, aldA, aldB,*

	ansB, araA, araB, araC, araD, araE, araF, araG, araH, araJ, argG, aroA, aspA, bglB, bglF, bglG, caiA, caiB, caiC, caiE, caiF, caiT, cdd, chbB, chbC, chbF, chbG, chbR, chpA, chpR, cirA, cpdB, crp, crr, csgD, csgE, csgF, csgG, csiE, cstA, cyaA, cyoA, cyoB, cyoC, cyoD, cyoE, cysG, cytR, dadA, dadX, dctA, dcuA, dcuR, deoA, deoB, deoC, deoD, dgsA, dksA, dsdA, dsdX, dusB, ebgA, ebgC, entA, entB, entC, entD, entE, envZ, epd, exuT, fadD, fadL, fbaA, feaR, fecA, fecB, fecC, fecD, fecE, fepA, fis, fiu, fixC, fixX, flhC, flhD, focA, fucA, fucI, fucK, fucO, fucP, fucR, fucU, fumA, fumB, fur, gabD, gabP, gabT, gadA, gadB, gadC, gadE, gadX, galE, galK, galM, galP, galS, galT, gapA, gatA, gatC, gatD, gatY, gatZ, gcd, gdhA, glcC, glgA, glgC, glgP, glgS, glnA, glnG, glnL, glpA, glpB, glpC, glpD, glpE, glpF, glpG, glpK, glpQ, glpR, glpT, glpX, gltA, gltB, gltD, gltF, gntK, gntP, gntT, gntU, gntY, guaA, guaB, gutM, gutQ, gyrA, hpt, hupA, hupB, hyfA, hyfD, hyfE, hyfF, hyfG, hyfH, hyfI, hyfJ, hyfR, idnD, idnK, idnO, idnR, idnT, ilvB, ilvN, infB, ivbL, kbaZ, lacA, lacY, lacZ, lamB, lpd, lsrA, lsrB, lsrC, lsrD, lsrF, lsrG, lyx, malE, malF, malG, malI, malK, malM, malS, malT, malX, malY, manX, manY, manZ, maoC, marA, marR, mdh, mdtE, mdtF, melA, melB, melR, metK, mglA, mglB, mglC, mhpC, mhpD, mhpE, mhpF, modA, modB, modC, mpl, mtlA, mtlD, mtlR, nagA, nagB, nagC, nagD, nagE, nanC, nanE, nanK, nanM, nirB, nirD, nmpC, nrdA, nrdB, nupC, nupG, nusA, ompA, ompF, ompR, osmY, oxyR, paaA, paaB, paaC, paaD, paaE, paaF, paaG, paaH, paaI, paaJ, paaK, pdhR, pflB, pgk, pncB, pnp, ppiA, proP, prpB, prpC, prpD, prpE, prpR, psiE, ptsG, ptsH, ptsI, putP, rbfA, rbsA, rbsB, rbsC, rbsD, rbsK, rbsR, relA, rhaA, rhaB, rhaD, rhaR, rhaS, rhaT, rpoH, rpoS, rpsO, sdhA, sdhB, sdhC, sdhD, serA, serC, sfsA, sgbE, sgbH, sgbU, sodA, sodB, sohB, speC, srlA, srlB, srlD, srlE, srlR, sucA, sucB, sucC, sucD, tdcA, tdcB, tdcC, tdcD, tdcE, tdcG, tnaA, tnaB, treB, treC, truB, trxA, tsx, ubiG, udp, ugpA, ugpC, uhpT, uidA, uidB, uidC, ulaB, ulaC, ulaD, ulaE, ulaF, uxaA, uxaB, uxaC, uxuA, uxuR, xseA, xylA, xylB, xylG, xylR, ybdB, ychH, yfiD, ygaF, yhbC, yhcH, yhfA, yiaJ, yiaK, yiaL, yiaM, yiaN, yiaO, yjcH, ynfK, zraR, zraS
CsgD	*adrA, csgA, csgC, csgD, csgE, csgF, csgG, iraP, pepD*
CueR	*copA, cueO, moaA, moaB, moaC, moaE*
CusR	*cusA, cusB, cusC, cusF, cusR, cuss*
CysB	*cbl, cysA, cysC, cysD, cysH, cysI, cysJ, cysK, cysM, cysN, cysP, cysU, ssuB, ssuC, ssuD, ssuE, tauA, tauB, tauC, tauD*
CytR	*cdd, cytR, deoA, deoB, deoC, deoD, nupC, nupG, ppiA, rpoH, tsx, udp*
DcuR	*dctA, frdA, frdB, frdC, frdD, fumB*

DeoR	*deoA, deoB, deoC, deoD, nupG, tsx*
DgsA	*crr, dgsA, malT, manX, manY, manZ, ptsG, ptsH, ptsI, ynfK*
DicA	*dicB, insD, intQ, ydfD, ydfE*
DnaA	*aldA, dnaA, dnaN, guaA, guaB, nrdA, nrdB, polA, recF, rpoH*
DpiA	*appY, citC, citD, citE, citF, citG, citX*
EnvY	*moaA, moaB, moaC, moaE, ompC, ompF*
EvgA	*acrD, emrK, emrY, evgA, evgS, frc, gadE, mdtE, mdtF, ydeO, ydeP, yfdX*
ExuR	*exuR, exuT, uxaA, uxaB, uxaC, uxuA, uxuR*
FadR	*fabA, fabB, fadA, fadB, fadD, fadE, fadI, fadJ, fadL, iclR, uspA*
FhlA	*fhlA, hycA, hycB, hycC, hycD, hycE, hycF, hycG, hycH, hycI, hydN, hyfA, hyfD, hyfE, hyfF, hyfG, hyfH, hyfI, hyfJ, hyfR, hypA, hypB, hypE, hypF*
Fis	*acnB, acs, actP, adhE, aldB, ansB, apaG, apaH, bglB, bglF, bglG, carA, carB, chpA, chpR, crp, cspI, cysG, deoA, deoB, deoC, deoD, dmsA, dmsB, dmsC, dusB, fadA, fadB, fis, flxA, fumB, gadA, gadB, gadC, gadX, glcC, glnA, glnG, glnL, glnQ, glpA, glpB, glpC, glpQ, glpT, guaA, guaB, gyrA, gyrB, hns, hupA, hupB, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF, infB, katE, ksgA, lpd, malE, malF, malG, marA, marR, mazG, mglA, mglC, msrA, mtlA, mtlD, mtlR, nanE, nanK, narG, narH, narI, narJ, narK, ndh, nirB, nirD, nrdA, nrdB, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, nusA, osmE, osmY, pdxA, pflB, pnp, proP, ptsG, pyrD, queA, rbfA, rpsO, topA, tpr, trmA, truB, tufA, tufB, xylG, xylR, yfiD, ygjG, yhbC, yhcH, yjcH*
FlhDC	*ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, flgA, flgC, flgD, flgE, flgF, flgG, flgH, flgI, flgJ, flgM, flgN, flhA, flhB, fliA, fliD, fliE, fliF, fliG, fliH, fliI, fliJ, fliK, fliL, fliM, fliN, fliP, fliQ, fliR, fliS, fliT, fliY, fliZ, glpA, glpB, glpC, gltI, gltJ, gltK, gltL, hydN, hypF, mdh, mglA, mglB, mglC, napA, napB, napC, napD, napF, napG, napH, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, ppdA, ppdB, ppdC, recC, ycgR, yecR, ygbK, yhjH*
FNR	*aceE, aceF, ackA, acnA, acrE, adhE, aer, aldA, ansB, arcA, aspA, bcsB, bcsZ, cadC, caiA, caiB, caiC, caiE, caiF, caiT, ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, cheB, cheR, cheY, cheZ, cydA, cydB, cydC, cydD, cyoA, cyoB, cyoC, cyoD, cyoE, cysG, dcuA, dcuC, dcuR, dcuS, dmsA, dmsB, dmsC, dmsD, dppA, dppB, dppC, dppD, dppF, emrK, emrY, entF, fdnG, fdnH, fdnI, feoA, feoB, fepE, fes, fhlA, fixC, fixX, fnr, focA, frdA, frdB, frdC, frdD, fumA, fumB, gadA, gadB, gadW, gadX, garK, garL, garR, gcvH, gcvP, gcvT, glpA, glpB, glpC, glpQ, glpT, gltB, gltD, gltF,*

	hcp, hcr, hemA, hmp, hyfA, hyfD, hyfE, hyfF, hyfG, hyfH, hyfI, hyfJ, hyfR, hypB, hypE, katG, lpd, malP, malQ, moaA, moaB, moaC, moaE, moeA, moeB, napA, napB, napC, napD, napF, napG, napH, narG, narH, narI, narJ, narK, narL, narX, ndh, nikC, nikD, nikE, nikR, nirB, nirD, norV, norW, nrdD, nrdG, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, ompW, ompX, pdhR, pepT, pflB, pheM, phoU, pitA, prfA, prmC, pstA, pstB, pstC, pstS, purM, purN, rimM, rplB, rplC, rplD, rplM, rplP, rplS, rplT, rplV, rplW, rpmC, rpsC, rpsI, rpsJ, rpsP, rpsQ, rpsS, scpA, scpB, scpC, sdhA, sdhB, sdhC, sdhD, sodA, ssuB, ssuC, ssuD, ssuE, sucA, sucB, sucC, sucD, tap, tar, tdcA, tdcB, tdcC, tdcD, tdcE, tdcG, tpx, trmD, ubiA, upp, uraA, uxaA, uxaC, xdhA, xdhB, xdhC, ycaC, ychO, ydhT, ydhU, ydhV, ydhW, ydhX, ydhY, yecR, yeiL, yfiD, ygbA, yhjA, yjiD, ynfE, ynfF, ynfG, ynfH, yqjI, ysgA, ytfE
FruR	*aceA, aceB, aceK, acnA, acnB, adhE, crr, cydA, cydB, cysG, eda, edd, eno, epd, fbaA, fruA, fruB, fruK, gapA, glk, hypF, icd, mtlA, mtlD, mtlR, nirB, nirD, pck, pfkA, pgk, pps, ptsH, ptsI, pykF, yahA*
FucR	*fucA, fucI, fucK, fucO, fucP, fucR, fucU*
Fur	*cirA, cyoA, cyoB, cyoC, cyoD, cyoE, entA, entB, entC, entD, entE, entF, entS, exbB, exbD, fecA, fecB, fecC, fecD, fecE, fecI, fecR, feoA, feoB, fepA, fepB, fepC, fepD, fepE, fepG, fes, fhuA, fhuB, fhuC, fhuD, fhuE, fhuF, fiu, flhC, flhD, fumB, fur, gpmA, hmp, metH, metJ, mntH, nohA, nohB, nrdE, nrdF, nrdH, nrdI, ompF, purR, rcnA, rcnR, sdhA, sdhB, sdhC, sdhD, sodA, sodB, sucA, sucB, sucC, sucD, sufA, sufB, sufC, sufE, sufS, tfaD, tfaQ, tfaR, tonB, ybdB, ydfN, ygaC, yhhY, yodA*
GadE	*cadA, cadB, cyoA, cyoB, cyoC, cyoD, cyoE, fabZ, fliC, gadA, gadB, gadC, gadE, gadW, gadX, gltB, gltD, gltF, gnd, hdeA, hdeB, hdeD, lpxA, lpxD, lrp, mdtE, mdtF, purA, rcsA, skp, yhiD*
GadW	*gadA, gadB, gadC, gadE, gadW, gadX, mdtE, mdtF*
GadX	*amtB, asnB, cadA, cadB, gadA, gadB, gadC, gadE, gadX, glnK, hdeA, hdeB, hdeD, hns, lon, mdtE, mdtF, rpoS, ybaS, ybaT, yhiD*
GalR	*galE, galK, galM, galP, galR, galS, galT, mglA, mglB, mglC*
GalS	*galE, galK, galM, galP, galR, galS, galT, mglA, mglB, mglC*
GatR	*gatA, gatC, gatD, gatY, gatZ*
GlcC	*glcA, glcB, glcC, glcD, glcE, glcG*
NtrC	*amtB, argT, astA, astB, astC, astD, astE, cbl, ddpA, ddpB, ddpC, ddpD, ddpF, ddpX, glnA, glnG, glnH, glnK, glnL, glnP, glnQ, hisJ, hisM, hisP, hisQ, nac, potF, potG, potH, potI, rutA, rutB, rutD, rutE, rutG, yeaG, ygjG, yhdW, yhdX, yhdY, yhdZ*

GlpR	*glpA, glpB, glpC, glpD, glpF, glpK, glpQ, glpT, glpX*
GntR	*eda, edd, gntK, gntT, gntU, gntY, idnD, idnK, idnO, idnR, idnT*
GutM	*gutM, gutQ, srlA, srlB, srlD, srlE, srlR*
HcaR	*hcaB, hcaC, hcaD, hcaE, hcaF, hcaR*
H-NS	*adiA, appY, bglB, bglF, bglG, bglJ, bolA, cadA, cadB, caiF, chiA, chpA, chpR, csiE, cspD, cydA, cydB, cysA, cysG, cysM, cysP, cysU, degP, entF, fepE, fes, fimA, fimB, fimD, fimE, fimF, fimG, fimH, fimI, flhC, flhD, fliA, fliC, fliY, fliZ, gabD, gabP, gabT, gadA, gadB, gadW, gadX, galE, galK, galM, galT, garK, garL, garR, gspA, gspB, gspC, gspD, gspE, gspF, gspH, gspI, gspJ, gspK, gspL, gspM, gspO, gutM, gutQ, hchA, hdeA, hdeB, hdeD, hisJ, hisM, hisP, hisQ, hns, ilvH, ilvI, lacA, lacY, lacZ, leuO, mukB, mukE, mukF, nhaA, nhaR, nirB, nirD, osmC, proV, proW, proX, rcsA, relA, smtA, sodB, srlA, srlB, srlD, srlE, srlR, stpA, yciE, yciF, ygaF, yhiD, yjjQ*
HU	*galE, galK, galM, galT, mtr, pgm, seqA, tyrP*
HyfR	*hyfA, hyfD, hyfE, hyfF, hyfG, hyfH, hyfI, hyfJ, hyfR*
IdnR	*gntK, gntU, idnD, idnK, idnO, idnR, idnT*
IHF	*aceA, aceB, aceK, acs, actP, adiA, amiA, atoA, atoB, atoD, atoE, caiA, caiB, caiC, caiE, caiT, carA, carB, cysG, cysH, cysI, cysJ, dcuD, dmsA, dmsB, dmsC, dppA, dppB, dppC, dppD, dppF, dps, dusB, envZ, fhlA, fimA, fimB, fimD, fimF, fimG, fimH, fimI, fis, flhC, flhD, focA, folA, gcd, glcA, glcB, glcD, glcE, glcG, glnH, glnP, glnQ, glpQ, glpT, gltA, gltB, gltD, gltF, hemA, hemF, hipA, hipB, hpt, htrE, hycA, hycB, hycC, hycD, hycE, hycF, hycG, hycH, hycI, hypA, hypB, hypE, ihfA, ihfB, ilvA, ilvD, ilvE, ilvL, ilvM, lyx, maoC, mtr, narG, narH, narI, narJ, narK, ndh, nirB, nirD, nmpC, norV, norW, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, ompC, ompF, ompR, osmE, osmY, paaA, paaB, paaC, paaD, paaE, paaF, paaG, paaH, paaI, paaJ, paaK, pflB, phoU, prfA, prmC, pspA, pspB, pspC, pspD, pspE, pspG, pstA, pstB, pstC, pstS, rpoH, rtcA, rtcB, sgbE, sgbH, sgbU, sodA, sodB, ssuB, ssuC, ssuD, ssuE, sucA, sucB, sucC, sucD, sufA, sufB, sufC, sufE, sufS, tdcA, tdcB, tdcC, tdcD, tdcE, tdcG, tyrP, ubiA, ulaB, ulaC, ulaD, ulaE, ulaF, ulaG, uspA, uspB, yeiL, ygjG, yiaJ, yiaK, yiaL, yiaM, yiaN, yiaO, yjbE, yjbF, yjbG, yjbH, yjcH*
IscR	*erpA, gntY, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF, iscA, iscR, iscS, iscU, napA, napB, napC, napD, napF, napG, napH, sufA, sufB, sufC, sufE, sufS, ydiU*
LeuO	*bglB, bglF, bglG, bglJ, cadC, leuA, leuB, leuD, leuL, leuO, yjjQ*
LexA	*ddlB, dinF, dinG, dnaG, ftsA, ftsI, ftsK, ftsL, ftsQ, ftsW, ftsZ, insK, lexA, lpxC, mraY, murC, murD,*

	murE, murF, murG, phr, polB, recA, recN, recX, rpoD, rpsU, ruvA, ruvB, ssb, sulA, symE, umuC, umuD, uvrA, uvrB, uvrC, uvrD, uvrY, ydjM
Lrp	*aidB, aroA, dadA, dadX, fimA, fimD, fimE, fimF, fimG, fimH, fimI, gabD, gabP, gabT, gcvH, gcvP, gcvT, gltB, gltD, gltF, hdeA, hdeB, ilvA, ilvD, ilvE, ilvH, ilvI, ilvL, ilvM, kbl, livF, livG, livH, livJ, livK, livM, lrp, lysU, malT, ompC, ompF, oppA, oppB, oppC, oppD, oppF, osmC, osmY, sdaA, serA, serC, stpA, tdh, yeiL, ygaF, yhiD*
LsrR	*lsrA, lsrB, lsrC, lsrD, lsrF, lsrG*
MalT	*lamB, malE, malF, malG, malK, malM, malP, malQ, malS, malZ*
MarA	*acrA, acrB, dctR, fpr, fumC, hdeA, hdeB, inaA, marA, marR, nfo, nfsB, poxB, pqiA, pqiB, purA, putA, rob, slp, sodA, yhiD, zwf*
MetJ	*ahpF, metA, metB, metC, metE, metF, metI, metK, metL, metN, metQ, metR*
MetR	*glyA, hmp, metA, metE, metH, metR*
ModE	*ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, deoA, deoB, deoC, deoD, dmsA, dmsB, dmsC, hycA, hycB, hycC, hycD, hycE, hycF, hycG, hycH, hycI, moaA, moaB, moaC, moaE, modA, modB, modC, napA, napB, napC, napD, napF, napG, napH, narL, narX, oppA, oppB, oppC, oppD, oppF*
Nac	*asnC, codA, codB, gabD, gabP, gabT, gdhA, gltB, gltD, gltF, mioC, mnmG, nac, nupC, serA*
NagC	*chbB, chbC, chbF, chbG, chbR, fimB, glmS, glmU, manX, manY, manZ, nagA, nagB, nagC, nagD, nagE, nanC, nanM*
NanR	*fimB, nanC, nanE, nanK, nanM, yhcH*
NarL	*adhE, aspA, caiF, ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, cydC, cydD, cysG, dcuA, dcuR, dcuS, dmsA, dmsB, dmsC, fdnG, fdnH, fdnI, focA, frdA, frdB, frdC, frdD, fumB, hcp, hcr, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF, hybA, hybB, hybC, hybD, hybE, hybF, hybG, hybO, moeA, moeB, napA, napB, napC, napD, napF, napG, napH, narG, narH, narI, narJ, narK, nikC, nikD, nikE, nikR, nirB, nirD, norV, norW, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, pflB, torA, torC, torD, ubiA, ydhT, ydhU, ydhV, ydhW, ydhX, ydhY, yeaR, yoaG, ytfE*
NarP	*ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, cysG, fdnG, fdnH, fdnI, hcp, hcr, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF, napA, napB, napC, napD, napF, napG, napH, nirB, nirD, norV, norW, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, ydhT, ydhU, ydhV, ydhW, ydhX, ydhY, yeaR, yoaG, ytfE*

NhaR	*nhaA, nhaR, osmC, pgaA, pgaB, pgaC, pgaD*
NrdR	*nrdA, nrdB, nrdD, nrdE, nrdF, nrdG, nrdH, nrdI*
NsrR	*hcp, hcr, hmp, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, tehA, tehB, yeaR, ygbA, yoaG, ytfE*
OmpR	*bolA, csgD, csgE, csgF, csgG, fadL, flhC, flhD, nmpC, ompC, ompF, tppB*
OxyR	*ahpF, dps, fur, gor, grxA, hemH, katG, oxyR, sufA, sufB, sufC, sufE, sufS, trxC, yhjA*
PaaX	*maoC, paaA, paaB, paaC, paaD, paaE, paaF, paaG, paaH, paaI, paaJ, paaK*
PdhR	*aceE, aceF, cyoA, cyoB, cyoC, cyoD, cyoE, fecA, fecB, fecC, fecD, fecE, hemL, hha, lpd, ndh, pdhR, ybaJ, yfiD*
PhoB	*amn, argP, asr, eda, phnC, phnD, phnE, phnF, phnG, phnH, phnI, phnJ, phnK, phnM, phnO, phnP, phoA, phoB, phoE, phoH, phoR, phoU, pitB, psiE, psiF, pstA, pstB, pstC, pstS, ugpA, ugpC, yibD*
PhoP	*acrA, acrB, argD, borD, dcuD, fadL, hemL, malS, metB, metL, mgrB, mgtA, nagA, pagP, phoP, phoQ, purD, purH, rstA, rstB, rutA, rutB, rutD, rutE, rutG, slyB, treR, ybjG, yrbL*
PrpR	*prpB, prpC, prpD, prpE, prpR*
PspF	*pspA, pspB, pspC, pspD, pspE, pspF, pspG*
PurR	*carA, carB, codA, codB, cvpA, gcvH, gcvP, gcvT, glnB, glyA, guaA, guaB, hflD, prs, purA, purB, purC, purD, purE, purF, purH, purK, purL, purM, purN, purR, pyrC, pyrD, speA, speB, ubiX*
RbsR	*rbsA, rbsB, rbsC, rbsD, rbsK, rbsR*
RcsAB	*bdm, csgD, csgE, csgF, csgG, flhC, flhD, ftsA, ftsZ, osmB, osmC, rcsA, wcaA, wcaB, wzb, wzc, yjbE, yjbF, yjbG, yjbH*
Rob	*acrA, acrB, aslB, fumC, inaA, marA, marR, nfo, sodA, ybiS, zwf*
AlsR	*alsA, alsB, alsC, alsE, rpiB*
RstA	*asr, csgD, csgE, csgF, csgG, narG, narH, narI, narJ, ompF*
RutR	*carA, carB, gadW, gadX, gmr, rutA, rutB, rutD, rutE, rutG, rutR*
SgrR	*setA, tbpA, thiP, thiQ, yfdZ*
SoxS	*acrA, acrB, fldA, fldB, fpr, fumC, fur, inaA, marA, marR, nfo, nfsA, pgi, poxB, pqiA, pqiB, ptsG, ribA, rimK, sodA, soxS, ybjC, ybjN, yodA, zwf*
GutR	*gutM, gutQ, srlA, srlB, srlD, srlE, srlR*
TdcA	*tdcA, tdcB, tdcC, tdcD, tdcE, tdcG*
TdcR	*tdcA, tdcB, tdcC, tdcD, tdcE, tdcG*
TorR	*gadA, gadB, gadX, hdeA, hdeB, tnaA, tnaB, torA, torC, torD, torR, yhiD*

TrpR	*aroH, aroL, mtr, trpA, trpB, trpC, trpD, trpE, trpL, trpR, yaiA*
TyrR	*aroF, aroG, aroL, aroP, folA, mtr, tyrA, tyrB, tyrP, tyrR, yaiA*
UlaR	*ulaB, ulaC, ulaD, ulaE, ulaF, ulaG*
UxuR	*gntP, uidA, uidB, uidC, uxuA, uxuR*
YiaJ	*lyx, sgbE, sgbH, sgbU, yiaK, yiaL, yiaM, yiaN, yiaO*
Zur	*ykgM, yodA, znuA, znuB, znuC*

Appendix D

Sigma19	*fecA, fecB, fecC, fecD, fecE*
Sigma28	*aer, cheA, cheB, cheR, cheW, cheY, cheZ, flgK, flgL, flgM, flgN, fliA, fliC, fliD, fliE, fliF, fliG, fliH, fliI, fliJ, fliK, fliL, fliM, fliN, fliP, fliQ, fliR, fliS, fliT, fliY, fliZ, flxA, modA, modB, modC, motA, ppdA, ppdB, ppdC, recC, tap, tar, tsr, ycgR, yecF, ygbK, yhiL, yhjH, yjcS, ynjH*
Sigma70, Sigma32	*can, clpB, dgsA, dsbC, glnS, groS, hepA, lipB, lon, pyrF, recJ, xerD, ybeD, yceI, yceJ, yciH, yciM, yciS, ynfK*
Sigma70, Sigma38	*acnA, adhE, aidB, appA, appB, appC, bolA, btuF, cbpM, cfa, csgA, csgC, csiE, dnaN, dps, ftsA, ftsQ, ftsZ, gadA, gadB, gadC, gadX, galE, galK, galM, galT, hdeA, hdeB, hyaA, hyaB, hyaC, hyaD, hyaE, hyaF, mglA, mglB, mglC, mpl, mtn, osmB, osmE, osmY, phoU, pqiA, pqiB, proV, proW, proX, pstA, pstB, pstC, pstS, recF, sohB, topA, yadS, yhiD*
Sigma70	*accA, accB, accC, accD, aceA, aceB, aceE, aceF, aceK, acnA, acnB, acrA, acrB, acs, actP, ada, adiA, adrA, agaA, agaB, agaC, agaD, agaI, agaR, agaS, agaV, agaW, ahpF, aldA, alkA, alkB, alsA, alsB, alsC, alsE, amiA, amiB, ampC, amyA, ansB, apt, araA, araB, araC, araD, araE, araF, araG, araH, araJ, arcA, argB, argC, argD, argE, argF, argG, argH, argI, argR, aroA, aroB, aroF, aroG, aroH, aroK, aroL, aroP, arsB, arsC, artI, artJ, artM, artP, artQ, ascB, ascF, ascG, asnA, asnC, aspA, asr, astA, astB, astC, astD, astE, atpA, atpB, atpC, atpD, atpE, atpF, atpG, atpH, bamC, bcp, betA, betB, betI, betT, bglJ, bglX, bioA, bioB, bioC, bioD, bioF, bolA, btuB, cadA, cadB, cadC, caiA, caiB, caiC, caiE, caiF, caiT, carA, carB, cbl, ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH, cdd, cedA, chiA, chpA, chpB, chpR, chpS, cirA, clpA, cls, cmk, coaD, cobS, cobT, cobU, codA, codB, corA, cpdB, cpxA, cpxP, cpxR, creA, creB, creC, creD, crp, crr, cspA, cspD, cspE, cstA, cusA, cusB, cusC, cusF, cusR, cusS, cutA, cvpA, cyaA, cydA, cydB, cynR, cynS, cynT, cynX, cyoA, cyoB, cyoC, cyoD, cyoE, cysA, cysC, cysD, cysG, cysH, cysI, cysJ, cysK, cysM, cysN, cysP, cysU, cytR, dadA, dadX, dam, damX, dapA, dapB, dapD, dapE, dctA, dctR, dcuA, dcuD, dcuR, dcuS, ddlB, def, deoA, deoB, deoC, deoD, dgsA, dicB, dinF, dinG, dksA, dmsA, dmsB, dmsC, dnaG, dnaN, dppA, dppB, dppC, dppD, dppF, dsbA, dsdA, dsdC, dsdX, dusB, dut, dxs, ebgA, ebgC, eda, edd, efeU, efp, entA, entB, entC, entD, entE, entF, entS, envZ, epd, era, evgA, evgS, exbB, exbD, fabA, fadA, fadB, fadD, fadI, fadJ, fadL, fbaA, fdnG, fdnH, fdnI, fdx, fecI, fecR, fepA, fepB, fepC, fepD, fepE, fepG, fes, fhuA, fhuB, fhuC, fhuD, fhuF, fimA, fimB, fimD, fimE, fimF, fimG, fimH, fimI, fis, fiu, fixC, fixX, fkpB, fldB, flgA, flgC, flgD, flgE, flgF, flgG, flgH, flgI, flgJ,*

	flgM, flgN, flhC, flhD, fliA, fliD, fliE, fliF, fliG, fliH, fliI, fliJ, fliK, fliL, fliM, fliN, fliP, fliQ, fliR, fliS, fliT, fliY, fliZ, fmt, fnr, focA, folA, folK, fpr, frc, frdA, frdB, frdC, frdD, frsA, fruA, fruB, fruK, ftnB, ftsA, ftsI, ftsK, ftsL, ftsQ, ftsW, ftsZ, fucA, fucI, fucK, fucO, fucP, fucR, fucU, fumA, fumB, fur, gabD, gabP, gabT, galE, galK, galM, galP, galR, galS, galT, gapA, gatA, gatC, gatD, gatY, gatZ, gcd, gcvA, gcvH, gcvP, gcvT, gdhA, glcA, glcB, glcC, glcD, glcE, glcG, glgA, glgC, glgP, glmS, glmU, glnA, glnB, glnG, glnL, gloA, glpA, glpB, glpC, glpD, glpF, glpK, glpQ, glpT, glpX, gltA, gltB, gltD, gltF, gltI, gltJ, gltK, gltL, gltX, glyA, glyQ, glyS, gmr, gnd, gntK, gntP, gntR, gntT, gntU, gntY, gph, gpmA, gpt, gspA, gspB, gspC, gspD, gspE, gspF, gspH, gspI, gspJ, gspK, gspL, gspM, gspO, guaA, guaB, gutM, gutQ, gyrA, gyrB, hchA, hdeD, hemA, hemF, hemH, hemN, hepA, hflB, hflC, hflD, hflK, hflX, hfq, hha, hipA, hipB, hisA, hisB, hisF, hisH, hisI, hisJ, hisM, hisP, hisQ, hisS, hns, hokD, hpf, hpt, hscA, hscB, htrE, hupB, hybA, hybB, hybC, hybD, hybE, hybF, hybG, hybO, hypF, icd, idnK, ileS, ilvA, ilvB, ilvC, ilvD, ilvE, ilvH, ilvI, ilvL, ilvM, ilvN, ilvY, imp, inaA, infA, infB, infC, insD, intQ, iraP, iscA, iscR, iscS, iscU, iscX, ispA, ispH, ivbL, katG, kbaY, kbaZ, kbl, kdpA, kdpB, kdpC, kdsA, kdsB, kdsC, kdsD, lacA, lacI, lacY, lacZ, lamB, leuA, leuB, leuD, leuL, leuO, lexA, livF, livG, livH, livJ, livK, livM, lldD, lldP, lldR, lpd, lpxC, lrp, lspA, lysA, lysC, lysP, lysR, lysU, lyx, malE, malF, malG, malI, malK, malM, malP, malQ, malS, malT, malX, malY, manX, manY, manZ, marA, marR, mdh, mdoG, mdoH, melA, melB, melR, menA, menB, menC, menE, metA, metB, metC, metF, metH, metI, metJ, metK, metL, metN, metQ, mfd, mgtA, mhpC, mhpD, mhpE, mhpF, mhpR, miaA, mioC, mngR, mnmG, modA, modB, modC, moeA, moeB, mpl, mprA, mraW, mraY, mraZ, mreB, mreC, mreD, mtlA, mtlD, mtlR, mtr, mukB, mukE, mukF, murC, murD, murE, murF, murG, murI, mutL, nadB, nagA, nagB, nagC, nagD, nagE, nanE, nanK, napA, napB, napC, napD, napF, napG, napH, narG, narH, narI, narJ, narK, narL, narU, narX, ndh, nfo, nfsA, nfsB, nhaA, nhaR, nirB, nirD, nohA, nohB, npr, nrdA, nrdB, nrdD, nrdG, nrdR, nrfA, nrfB, nrfC, nrfD, nrfE, nrfF, nrfG, nudB, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, nuoN, nupC, nupG, nusA, nusB, ompA, ompC, ompF, ompR, oppA, oppB, oppC, oppD, oppF, osmC, otsA, otsB, oxyR, paaA, paaB, paaC, paaD, paaE, paaF, paaG, paaH, paaI, paaJ, paaK, paaX, paaY, panB, panC, panF, pck, pcm, pcnB, pdhR, pdxJ, pepD, pfkA, pflB, pgi, pgk, pgpA, pheP, phnC, phnD, phnE, phnF, phnG, phnH, phnI, phnJ, phnK, phnM, phnO, phnP, phoA, phoB, phoE, phoH, phoR, pitB, pncB, pnp, pntA, pntB, polB, ppiA, ppiD, pps, prfA, prmA, prmC, proP, proS, proV, proW, proX, prpR, prs, psiE, psiF, pspF, pth, ptsH, ptsI, ptsN, purA, purB, purC, purD, purE, purF, purH, purK, purL, purM, purN, purR, putA, putP, pykF, pyrC,

	pyrD, qseB, qseC, rbfA, rbsA, rbsB, rbsC, rbsD, rbsK, rbsR, rcnA, rcnR, rcsA, rdoA, recA, recF, recN, recO, recX, relA, relB, relE, rfaB, rfaC, rfaF, rfaG, rfaI, rfaJ, rfaL, rfaP, rfaQ, rfaS, rfaY, rfaZ, rhaA, rhaB, rhaD, rhaR, rhaS, rhaT, ribA, ribD, ribE, ribF, rihB, rimK, rimM, rluA, rnb, rnc, rne, rpe, rpiB, rplB, rplC, rplD, rplP, rplQ, rplS, rplT, rplV, rplW, rpmC, rpmI, rpoA, rpoD, rpoE, rpoH, rpoN, rpoS, rpsA, rpsC, rpsD, rpsJ, rpsK, rpsM, rpsO, rpsP, rpsQ, rpsS, rpsU, rraA, rrmJ, rsd, rseA, rseB, rseC, rutR, ruvC, sbcC, sbcD, sdaB, sdaC, sdhA, sdhB, sdhC, sdhD, secG, serA, serC, setA, sfsA, sgbE, sgbH, sgbU, sixA, slmA, slp, smtA, sodA, sodB, sohA, sohB, soxR, soxS, speA, speB, speC, speD, speE, spy, srlA, srlB, srlD, srlE, srlR, ssb, ssuB, ssuC, ssuD, ssuE, stpA, sucA, sucB, sucC, sucD, sufA, sufB, sufC, sufE, sufS, sulA, surE, symE, tbpA, tdh, tfaD, tfaQ, tfaR, thiL, thiP, thiQ, thrA, thrB, thrC, thrL, tig, tnaA, tnaB, tonB, topA, torA, torC, torD, tppB, tpr, tpx, treB, treC, treR, trmD, trpA, trpB, trpC, trpD, trpE, trpL, trpR, trpS, truB, trxC, tsr, tsx, tufA, tufB, tyrA, tyrB, tyrP, tyrR, ubiA, ubiG, ubiX, udp, ugpA, ugpC, uhpT, ulaB, ulaC, ulaD, ulaE, ulaF, umuC, umuD, upp, uraA, uspA, uvrA, uvrB, uvrD, uxuA, valS, ves, waaA, waaU, wcaA, wcaB, wzb, wzc, xapB, xapR, xseA, xseB, yacC, yahA, yaiA, yajO, ybaJ, ybdB, ybjC, ybjN, ycaR, yccA, ycdN, ychA, ychF, ychH, ychQ, yciU, ydeH, ydeO, ydeP, ydfD, ydfE, ydfN, ydjM, yeaR, yebB, yebC, yebE, yejA, yejB, yejE, yejF, yfdX, yfdZ, yffB, yggG, yhaV, yhbC, yhbJ, yhcH, yhdT, yhfA, yhhY, yhjA, yiaJ, yiaK, yiaL, yiaM, yiaN, yiaO, yibD, yjaB, yjbE, yjbF, yjbG, yjbH, yjcH, yjeE, yjeF, yjjQ, ykgM, yoaE, yoaG, yodA, ypfN, yqjA, yqjB, yrbG, yrbK, zntA, znuB, znuC, zwf
Sigma24	*ahpF, apaG, apaH, bacA, bamA, bamB, bamC, bamD, cca, cutC, degP, der, dnaE, dsbC, fabZ, fkpA, ftnB, fusA, greA, gspA, gspB, hcp, hcr, hpf, htrG, imp, ksgA, lhr, lon, lptA, lpxA, lpxB, lpxD, lyx, malQ, narV, narW, pdxA, plsB, prfB, psd, ptsN, recJ, recR, rfaC, rfaF, rfaL, rnhB, rpoD, rpoH, rpoN, rseA, rseB, rseC, rseP, rutR, rzoD, rzoR, rzpD, rzpR, sbmA, sgbE, sgbH, sgbU, skp, surA, tufA, tufB, uspD, wzb, wzc, yaiW, ybaB, ydhI, ydhJ, ydhK, yeaY, yfeK, yfeS, yfeX, yfeY, yfgC, yfgD, yfjO, yggN, yghF, yhbJ, yhjJ, yiaK, yiaL, yiaM, yiaN, yiaO, yicI, yidQ, yieE, yieF, yiiS, yjeP, yqjA, yqjB, yraP, ytfJ*
Sigma32	*bssS, creA, creB, creC, fkpB, fxsA, gapA, gntY, hflB, hflC, hflK, hflX, hfq, holC, hspQ, ileS, ispH, lnt, lspA, macB, metA, miaA, mutL, mutM, narP, nusB, osmF, pgpA, phoP, phoQ, pphA, ppiD, prlC, raiA, rdgB, rfaC, rfaF, rfaL, ribE, rnlA, rpmE, rpoD, rrmJ, sdaA, thiL, topA, trmA, tyrR, valS, yafD, yafE, yafU, ybeX, ybeY, ybeZ, yccE, ycjF, ycjX, ydeO, yeaD, yehR, yehW, yehY, yfjV, ygaD, ygbF, ygbT, ygcH, ygcI, yggW, yhdN, yhiQ, yiaA, yibA, yjaZ, yjhG, yjhH, yjhI, yjiT, yrdA, yrfG, zntR*

Sigma54	*amtB, argT, astA, astB, astC, astD, astE, atoA, atoB, atoD, atoE, chaC, dcuD, ddpA, ddpB, ddpC, ddpD, ddpF, ddpX, fhlA, glnA, glnG, glnH, glnK, glnL, glnP, glnQ, gltI, gltJ, gltK, gltL, hisJ, hisM, hisP, hisQ, hycA, hycB, hycC, hycD, hycE, hycF, hycG, hycH, hycI, hydN, hyfA, hyfD, hyfE, hyfF, hyfG, hyfH, hyfI, hyfJ, hyfR, hypA, hypB, hypE, hypF, kch, nac, norV, norW, potF, potG, potH, potI, prpB, prpC, prpD, prpE, pspA, pspB, pspC, pspD, pspE, pspG, puuP, rpoH, rtcA, rtcB, rtcR, rutA, rutB, rutD, rutE, rutG, yaiS, ybhK, yeaG, yfhK, ygjG, yhdW, yhdX, yhdY, yhdZ, zraR, zraS*
Sigma38	*adhE, aldB, ansP, artI, artM, artP, artQ, astA, astB, astC, astD, astE, blc, cfa, csgD, csgE, csgF, csgG, fic, fliY, ftsQ, fumC, gabD, gabP, gabT, gadE, gadW, gadX, glgS, gor, hchA, hmp, htrE, ihfA, ihfB, katE, ldcC, lsrA, lsrB, lsrC, lsrD, lsrF, lsrG, mdtE, mdtF, msyB, narU, nhaA, nhaR, osmB, osmC, osmF, otsA, otsB, pfkB, phr, poxB, proP, rraA, rsd, rssB, talA, tktB, treA, uspB, xthA, ybgA, ybjP, yehW, yehY, yeiL, ygaF, yggE, yiaG, yihG, ytfK*

Appendix E

Introduction

Software DrDMASS+ has been developed to effectively analyze mass spectral data based on multivariate analysis. Figure 1 shows a flow diagram of Data Processing consisting of four stages, (i) Peak Correction, (ii) Multivariate Data Preprocessing, (iii) Unsupervised Learning, and (iv) Supervised Learning. In Peak Correction process, we can correct experimental *m/z* values based on the relation between experimental and desired values of internal mass calibrants (**IMC**s). A multivariate data is consisting of a data set of multiple samples. In Multivariate Data Preprocessing, we can assess reproducibility of samples with iterative measurement, and select useful peaks for separating groups of samples and so on. In Unsupervised Learning, we can visualize the multivariate data by using multivariate analysis method such as principal component analysis (PCA) and Batch-learning self-organizing map (BL-SOM). In Supervised Learning, we can get the regression equation by using Partial Least Squares Regression (PLS).

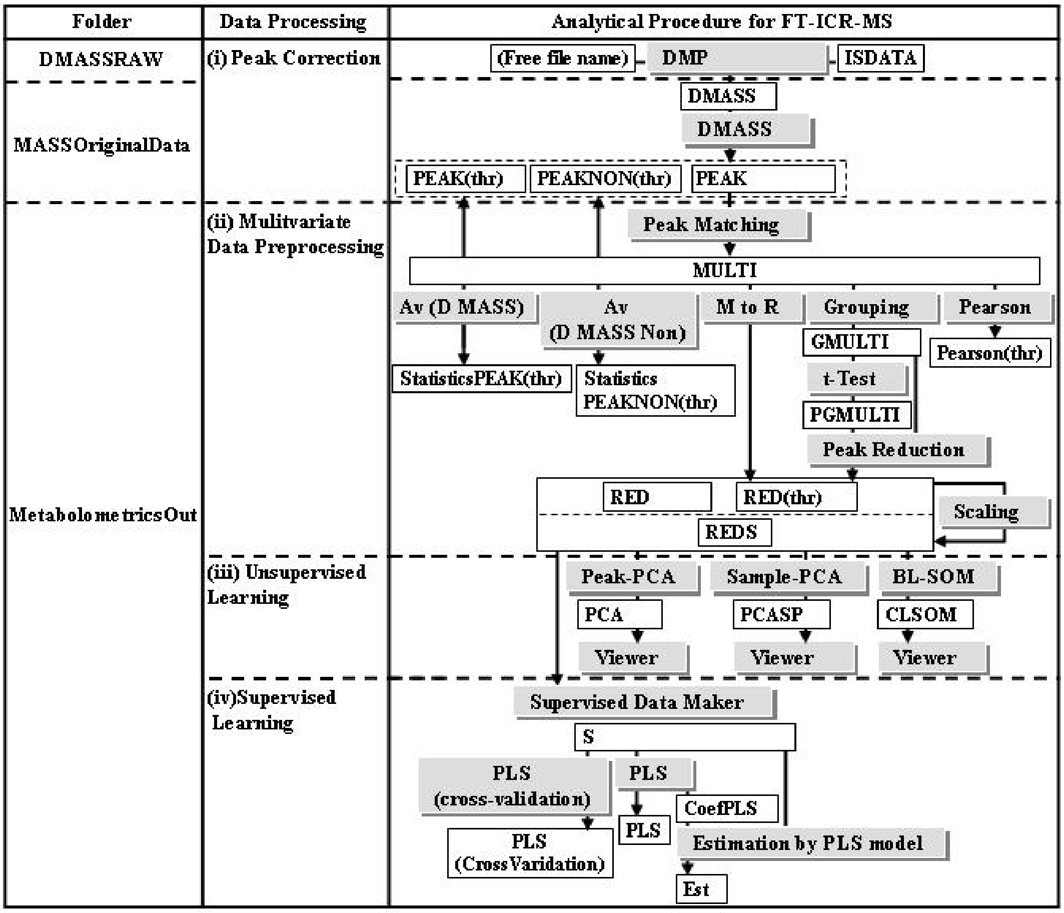
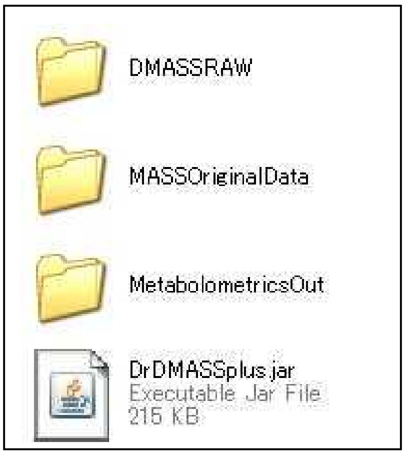


Figure 1. Flow Diagram of Data Processing in DrDMASS+ (Silver boxes correspond to individual processes, and white boxes correspond to prefix in input/output file names).

1. Execution of DrDMASS+

Java j2sdk-1.4.2 is required to be installed in the user’s computer. First, the compressed file, DrDMASSplus.zip is to be downloaded from http://kanaya.naist.jp/DrDMASSplus/. Under the ‘DrDMASSplus’ folder, there are three folders ‘DMASSRAW’, ‘MASSOriginalData’, and ‘MetabolometricsOut’, and an executable file ‘DrDMASSplus.jar’.



1.1 Starting data files

Put digital mass spectral data and an internal mass calibrant data to ‘DMASSRAW’ folder. The calibrant data file name should start with ‘ISDATA’. These file formats are as follows.

Digital mass spectral data

Digital mass spectral data from an IonSpec Explorer FT-ICR (IonSpec Inc., Lake Forest, CA) equipped with a 8 tesla actively shielded super conducting magnet is a text file separated by tabs. The first to fifth columns correspond to *m/z*, Frequency, Amplitude, Relative abundant and Resolution, respectively.

m/z	Frequency	Amplitude	Rel.Abund.	Resolution
72.9895	1475427.93	0.0832	1.81	144100
73.6554	1462089.917	0.045	0.98	189100
……				
……				

ISDATA

ISDATA consists of *m/z* values for internal mass calibrants (**IMC**s).

```
218.96212
348.10235
613.38820
829.32078
```

1.2 Execution of DrDMASS+

User can start by clicking the file DrDMASSplus.jar. The main window is shown in Panel 1. The button names correspond to those in Figure 1. DrDMASS+ consists of 18 Data processing modules corresponding to the button names. In the present system, the processed single mass spectral data is put into 'MASSOriginalDATA', and the multivariate data is put into 'MetabolometricsOut' folder. Prefix for each input/output file is described in Figure 1. Each process is explained in detail in the next section and the summary of the processes is given below.

1. DMP Selection of *m/z* values for IMCs from Digital mass spectral data.

2. DMASS Correction of *m/z* values for all peaks by those of IMCs.

3. Peak Matching (D MASS) Matrix construction for multiple samples. *m/z* value and intensity for each sample is arranged in the matrix.

4. Av (D MASS) Calculation of average intensity for all samples.

5. Av (D Mass Non) Calculation of average intensity for samples with non-zero intensity.

6. M to R Construction of multivariate data consisting of *m/z* values and the intensity of multiple samples.

7. Group Definition of categories for individual samples.

8. *t*-Test Estimation of *p*-values by *t*-statistics for the difference between the average intensities for pairs of groups.

9. Peak Reduction Selection of peaks with the group differences by p-values.

10. Scaling Scaling data.

11. Pearson correlation Pearson correlations of the intensities for pairs of *m/z* larger than the threshold set by the user are list up.

12. Peak-PCA and its Viewer Principal component analysis for peaks and visualization of its results.

13. Sample-PCA and its Viewer Principal component analysis for samples and visualization of its results.

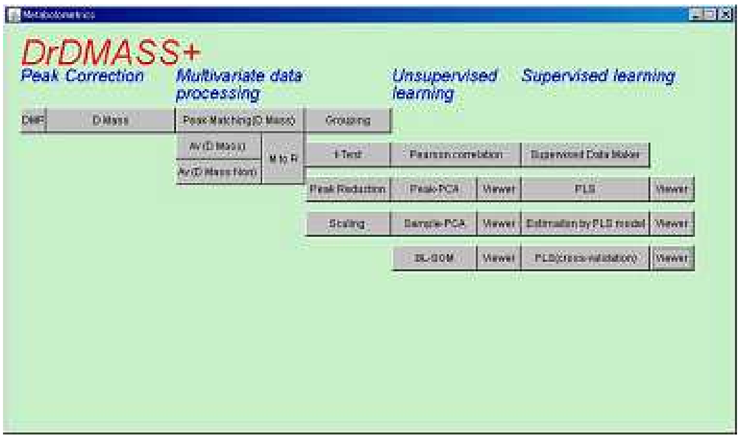
14. BL-SOM and its Viewer Batch-learning SOM for peaks and visualization of its results.

15. Supervised Data Maker Construction of multivariate data for PLS.

16. PLS and its Viewer Calculation of regression equation by using Partial Least Squares Regression and verifying calculations.

17. Estimation by PLS model and its Viewer
Estimation by PLS model and verifying calculations.

18. PLS (cross-validation) Calculation of the optimum number of components by using cross validation.



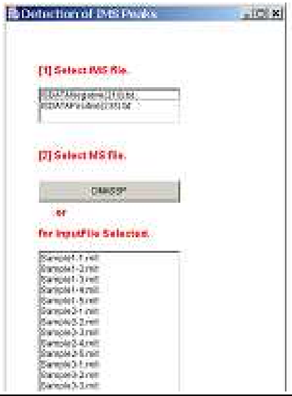
Panel 1. The main window.

2 Explanation of individual processes

(i) Peak Correction

2.1. DMP

DMP process is a selection process of *m/z* values for IMCs from Digital mass spectral data. The nearest *m/z* value in the digital mass spectral to those for IMCs is selected.

Input file	(i) Digital spectral data (its naming is free), and (ii) **ISDATA**
Output file	**DMASS**
Execution	**[1] Click DMP button**, so the following panel is displayed. **[2] Select** a suitable **IMC file** consisting of *m/z* values in internal mass calibrants. **[3] Click DMASSP button** if the selection process of *m/z* values for IMCs is carried out for all MS files, or **click an inputfile name** (for example Sample1-1.mit) if the selection process is carried out for a targeted input file. In the demonstration data, we select ISDATANegative(218).txt and click 'DMASSP' button.  Detection of IMS Peaks [1] Select IMS file. ISDATANegative(218).txt ISDATAPositive(238).txt [2] Select MS file. DMASSP or for inputFile Selected. Sample1-1.mit Sample1-2.mit Sample1-3.mit Sample1-4.mit Sample1-5.mit Sample2-1.mit Sample2-2.mit Sample2-3.mit Sample2-4.mit Sample2-5.mit Sample3-1.mit Sample3-2.mit Sample3-3.mit
Output file format (DMASS)	>Sample1-1.mit Standard 218.9664 218.96212 348.1119 348.10235 613.4106 613.3882 829.3563 829.32078 // AllData 72.9895 0.0832 73.6554 0.045 …. …. 976.4644 0.0313 // 1st line represents inputfile name. From 2nd line to '//' (7th line): *m/z* values for IMCs are listed, that is, experimental and theoretical values for individual IMCs correspond to the first and second columns. From AllData to '//' (final line), *m/z* and its intensity are arranged.

2.2 DMASS

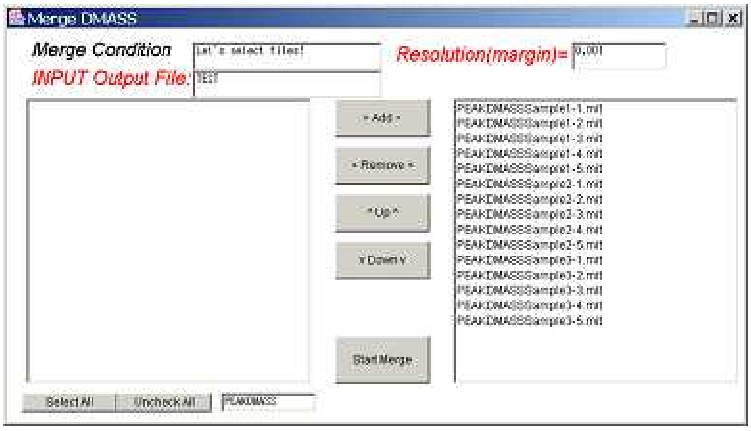
All *m/z* values are corrected by linear relationship between theoretical and experimental values in the interval defined by the nearest *m/z* values for IMCs.

Input file	**DMASS**
Output file	**PEAK**
Execution	**[1]** Select samples used by **clicking any number of filenames** or **'Select All' button**. **[2] Click 'Start Correct' button**. Output files started with 'PEAK' are obtained.
Output file format (PEAK)	>Sample1-1.mit Standard 218.96212 218.96212 348.10235 348.10235 613.3882 613.3882 829.32078 829.32078 // AllData 72.99117683367211 0.0832 73.65704966049918 0.045 … … 976.4199422981981 0.0313 // 1st line represents inputfile name. From 2nd line to '//' (7th line), *m/z* values shown for IMCs are listed, that is, the theoretical values for individual IMCs correspond to both columns. From AllData to '//' (final line), corrected *m/z* and its intensity are arranged.

(ii) Multivariate Data Preprocessing

2.3 Peak Matching (D MASS)

According to *m/z* values, peaks for multiple samples are arranged to matrix.

<table>
<tr><td><b>Input file</b></td><td><b>PEAK</b></td></tr>
<tr><td><b>Output file</b></td><td><b>MULTI</b></td></tr>
<tr><td><b>Execution</b></td><td>


Let's select files!
0.001
TEST
PEAKDMASSSample1-1.mit
PEAKDMASSSample1-2.mit
PEAKDMASSSample1-3.mit
PEAKDMASSSample1-4.mit
PEAKDMASSSample1-5.mit
PEAKDMASSSample2-1.mit
PEAKDMASSSample2-2.mit
PEAKDMASSSample2-3.mit
PEAKDMASSSample2-4.mit
PEAKDMASSSample2-5.mit
PEAKDMASSSample3-1.mit
PEAKDMASSSample3-2.mit
PEAKDMASSSample3-3.mit
PEAKDMASSSample3-4.mit
PEAKDMASSSample3-5.mit
PEAKDMASS

<b>[1]</b> Input <b>Output file name.</b>
<b>[2]</b> Input <b>Resolution (margin).</b> This parameter determines the region of <i>m/z</i> as identical positions.
<b>[3]</b> Select samples used by <b>clicking any number of filenames</b> or <b>'Select All' button</b>.
<b>[4] Click 'Add' button</b>, then the selected filenames are moved from left to right side.
<b>[5]</b> Prepare the order of filenames by <b>selecting a filename</b> and using <b>'Up' and 'Down' buttons</b>. The order of filenames corresponds to the order from left to right in the constructed matrix.
<b>[4] Click 'Start Merge' button</b>. Output files started with 'MULTI' are obtained.
</td></tr>
<tr><td><b>Output file format (MULTI)</b></td><td>

```
Resolution=0.0010
>Sample1-1.mit
>Sample1-2.mit
...
>Sample1-1.mit
standard[0]        218.96212     218.96212     218.96212     218.96212     ...
…                  ....          ...           ...           ...           ...
standard[3]        829.32078     829.32078     829.32078     829.32078     ...
63.67552147        0.0           0.0           0.0           0.0           ...
72.99098211        72.991176     0.0832        72.98792      0.0912        ...
...
...
976.3983668        976.41994     0.0313        976.3946      0.0353        ...
//
```

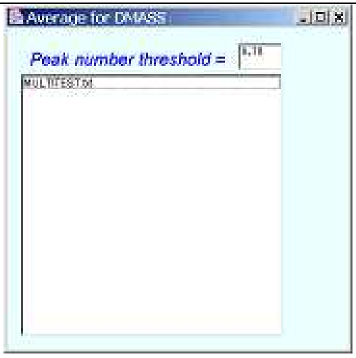
1st line represents Resolution. Lines started with '>' are sample names analyzed. Standard [0] to standard [3] represent corrected <i>m/z</i> for IMCs. After IMC line, pairs of <i>m/z</i> and its intensities are arranged according to the order of sample names.
</td></tr>
</table>

Reproducibility of iterative measurements

Reproducibility of iterative measurements can be checked by statistical assessment for ‘Av (D MASS)’ and ‘Av (D Mass Non)’.

2.4 Av (D MASS)

Calculation of average intensity for all samples.

<table>
<tr><td><b>Input file</b></td><td><b>MULTI</b></td></tr>
<tr><td><b>Output file</b></td><td><b>(i) PEAK(thr), (ii) StatisticsPEAK(thr)</b></td></tr>
<tr><td><b>Execution</b></td><td>

Peak number threshold = 0.18
MULTITEST.txt
<b>[1] Set peak number threshold</b> by relative value. For each set of <i>m/z</i> values, the average for their intensities are calculated when the sample number of non-zero intensities for each <i>m/z</i> is larger than the product of relative value and total sample number.
<b>[2] Click filename</b>, statistics for the reproducibility in iterative measurements is output to StatisticsPEAK(thr) file, and <i>m/z</i> values and its corrected average intensities by those for IMCs are output to a ‘PEAK(thr)’ file.</td></tr>
<tr><td><b>Output file format</b></td><td><b>PEAK(thr)</b>
Format of PEAK(thr) files is the same as that of PEAK files described in section 2.2. The intensity for each <i>m/z</i> is corrected by the average intensities for IMCs.
The word ‘thr’ in the parentheses represents threshold set in execution of Av (D MASS).
<b>StatisticsPEAK(thr)</b>
Statistical information for selected input file is obtained as follows. Details are described in the following subsection entitled ‘Statistics of Av (D MASS)’.
<table>
<tr><td></td><td>&gt;Sample1-1.mit</td><td>…</td><td>&gt;Sample1-5.mit</td><td>Av</td><td>SD</td><td>Av/SD</td></tr>
<tr><td>TOTAL</td><td>32.7891</td><td>…</td><td>34.7244</td><td>31.4475</td><td>2.32830</td><td>13.506</td></tr>
<tr><td>CRTOTAL</td><td>9.88717</td><td>…</td><td>10.1118</td><td>10.5488</td><td>1.36673</td><td>7.7182</td></tr>
<tr><td>AvRef</td><td>3.31632</td><td>…</td><td>3.43402</td><td>3.02552</td><td>0.46837</td><td>6.4595</td></tr>
<tr><td>Threshold=0.7</td><td>200</td><td>…</td><td>193</td><td></td><td></td><td></td></tr>
<tr><td></td><td>&gt;Sample1-1.mit</td><td>…</td><td>&gt;Sample1-5.mit</td><td>nonzero</td><td>Av</td><td>correctAv</td></tr>
<tr><td>72.9903883</td><td>0.0832</td><td>…</td><td>0.0836</td><td>5</td><td>0.08486</td><td>0.02887</td></tr>
<tr><td>73.65618938</td><td>0.0450</td><td>…</td><td>0.0496</td><td>5</td><td>0.05088</td><td>0.01731</td></tr>
<tr><td>…</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
</table>

Statistics of Av (D MASS)

In M iterative measurements, the intensities for IMCs and *m/z* are represented by data matrices **Y** and **X**, respectively. Here, the number of IMCs and peaks are denoted by S and N, respectively.

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{iM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{S1} & \cdots & y_{Sj} & \cdots & y_{SM} \end{pmatrix} \qquad (2.4.1)$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i'1} & \cdots & x_{i'j} & \cdots & x_{i'M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Nj} & \cdots & x_{NM} \end{pmatrix} \qquad (2.4.2)$$

Line that stars with ‘AvRef’ in StatisticsPEAK(thr) corresponds to the average of IMCs for *j*th measurement represented by Eq. (2.4.3).

$$\bar{y}_j = \frac{\sum_{i=1}^{S} y_{ij}}{S} \qquad (2.4.3)$$

The values under the columns Av, SD, Av/SD correspond to $\bar{y}$ represented by Eq. (2.4.4), $SD(\bar{y})$ represented by Eq. (2.4.5), and $\bar{y}/SD(\bar{y})$, respectively.

$$\bar{y} = \frac{\sum_{j=1}^{M} \bar{y}_j}{M} \qquad (2.4.4)$$

$$SD(\bar{y}) = \frac{\sqrt{\sum_{j=1}^{M} (\bar{y}_j - \bar{y})^2}}{M-1} \qquad (2.4.5)$$

Line that starts with ‘TOTAL’ corresponds to the statistical parameters for the intensities in *j*th measurement in a set of *m/z* whose intensities for all measurements are not zero is denoted by $\bar{x}_{ij}'$.

Concretely, $total\,(x'_j)$ corresponds to total intensities for *j*th.

Line ‘CRTOTAL’ corresponds to statistical parameters for corrected intensities by the average of IMCs, that is,

$$cav(x'_j) = \frac{av(\overline{x}'_j)}{\overline{y}_j} \quad . \tag{2.4.6}$$

Here, $av(x'_j) = \frac{\sum_{j=1}^{M} total\,(x'_j)}{N'}$.

In ’Threshold’ line, threshold set by the user and the number of peaks satisfied by this condition is represented. From the following line to the last line, *m/z* value, and corrected intensity for each sample are calculated as represented in Eq. (2.4.7).

$$c(x'_{ij}) = \frac{x'_{ij}}{\overline{y}_j} \tag{2.4.7}$$

The column ‘nonzero’ corresponds to the number of intensities larger than zero. CorrectAv corresponds to the average of $c(x'_{ij})$ represented by Eq. (2.4.8)

$$av(x'_i) = \frac{\sum_j c(x'_{ij})}{M'_i} \tag{2.4.8}$$

Here M’$_i$ represents the number of measurements larger than zero for *i*th *m/z*. The *m/z* and its correctedAv are arranged in the other output file (PEAK(thr)).

2.5 Av (D MASS Non)

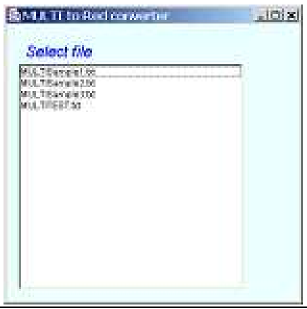
Calculation of average intensity for samples with non-zero intensity.

Input file	**MULTI**
Output file	**PEAKNON(thr), (ii) StatisticsPEAKNON(thr)**
Execution	Execution of Av (D MASS Non) is the same as that of Av (D MASS). **[1] Set peak number threshold** by relative value. For each set of *m/z* values, the average for their intensities is calculated when the sample number of non-zero intensities for each *m/z* is larger than the product of relative value and total sample number. **[2] Click filename**, statistics for the reproducibility in iterative measurements is output to StatisticsPEAK(thr) file, and *m/z* values and its average intensities are output to a 'PEAK(thr)' file.

```
               >Sample1-1.mit  …  >Sample1-5.mit  Av        SD       Av/SD
TOTAL          32.7891         …  34.724          31.4475   2.3283   13.5066
CRTOTAL        9.8871          …  10.111          10.5488   1.3667   7.71827
AvRef          3.3163          …  3.4340          3.0255    0.4683   6.45957
Threshold=0.7  200             …  193
               >Sample1-1.mit  …  >Sample1-5.mit  nonzero   Av       correctAv
72.9903        0.1699          …  0.1760          5         0.1815   0.0616
73.6561        0.0918          …  0.1044          5         0.1093   0.0371
…
…
```

2.6 M to R

In ‘M to R’, a multivariate data matrix consisting of average *m/z* values and the intensities for multiple measurements is constructed.

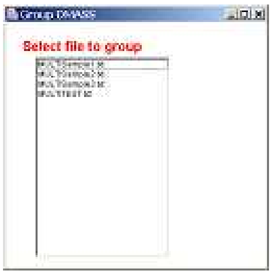
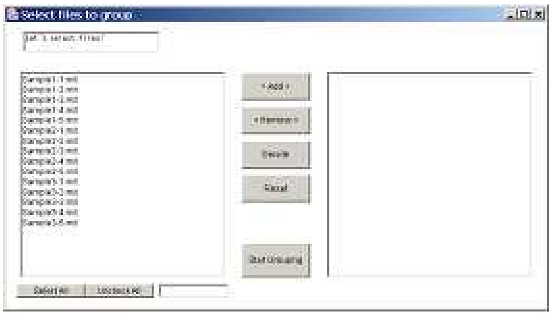
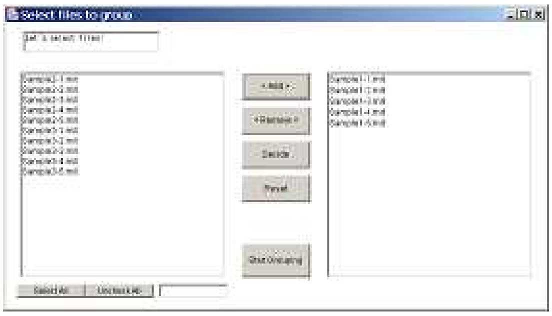
Input file	**MULTI**
Output file	**RED** (data format for multivariate analyses in DrDMASS+ system)
Execution	**Click filename**, then format in MULTI is exchanged to that in RED file.  MULTI to Red converter Select file MULTISample1.txt MULTISample2.txt MULTISample3.txt MULTITEST.txt
Output file format (RED)	1st line corresponds to inputfile name, and 2nd and 3rd lines correspond to group index and merged filenames, respectively. 4th to the last lines corresponds to *m/z* and intensities for individual measurements. >MULTITEST.txt >no. 1 2 3 … >filename >Sample1-1.mit >Sample1-2.mit >Sample1-3.mit … 72.99098211 0.0832 0.0912 0.0963 … 73.65656016 0.0450 0.0535 0.0571 … … …

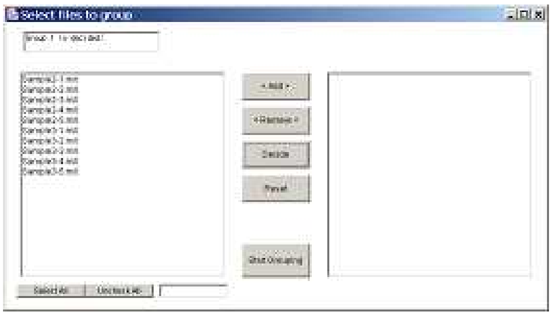
Feature (m/z) selection

Peak selection of the statistical significant differences in intensities between groups is carried out by the sequential processes entitled Group (Section 2.7), *t*-Test (Section 2.8), and Peak Reduction (Section 2.9).

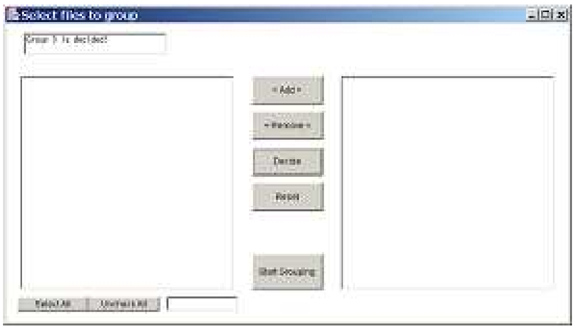
2.7 Group

Definition of categories for individual samples.

Input file	**MULTI**
Output file	**GMULTI**
Execution	**[1] Click Grouping button**, then Group DMASS panel is displayed. **[2] Select filename**, then 'Select files to group' panel is displayed.   Select files to group < Add > < Remove > Decide Reset Start Grouping Select All Uncheck All **[3] Select files** belonging to the same group**,** and move these files to right side by **clicking 'Add' button.**  Select files to group < Add > < Remove > Decide Reset Start Grouping Select All Uncheck All **[4] Click 'Decide'** button. So selected file is removed and their group is assigned to output file started with 'G'.



[5] Continue the manipulation of [3] and [4] until groups for all files are assigned, that is, all files are removed.
[6] Click 'Start Grouping' button, then grouping process is started.



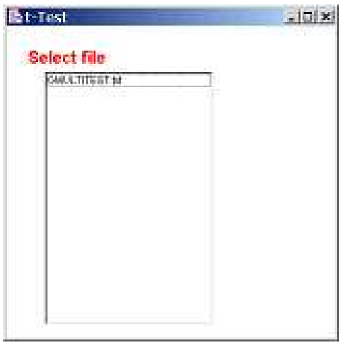
Output file format

Lines started with '>group' correspond to the numbers of individual groups. The line with '>no' corresponds to group ID, the line with 'filename' corresponds to the labels for individual measurements, and the *m/z* values and their intensities are arranged.

```
>group1      5
>group2      5
>group3      5
>no.         1               …        1          2            …     2               …
>filename    Sample1-1.mit            …          Sample1-5.mit      Sample2-1.mit   …
72.99098211  0.0832          …        0.0836     0.0635       …     0.0704          …
73.65656016  0.0450          …        0.0496     0.0477       …     0.0419          …
77.07104731  0               …        0          0            …     0               …
….
….
```

2.8 *t*-Test

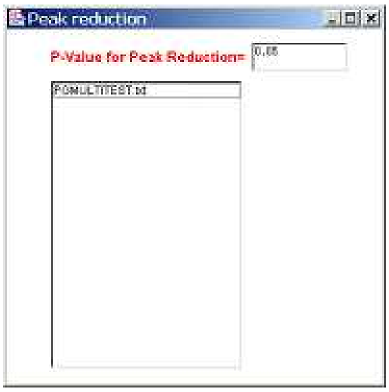
Estimation of *p*-values by *t*-statistics for the difference between the average intensities for pairs of groups.

Input file	**GMULTI**
Output file	**PGMULTI**
Execution	**Click filename**, then *p*-values by *t*-statistics for the difference between the average intensities for pairs of groups are calculated for individual *m/z*. 
Output file format	Lines started with '>group' correspond to measurements belonging to the same groups. Line with 'combination' represents pairs of groups. The *m/z* values and two statistical parameters, *t*-values and *p*-values, for all pairs of groups are arranged.

```
>group1     Sample1-1.mit   Sample1-2.mit   Sample1-3.mit  Sample1-4.mit   Sample1-5.mit
>group2     Sample2-1.mit   Sample2-2.mit   Sample2-3.mit  Sample2-4.mit   Sample2-5.mit
>group3     Sample3-1.mit   Sample3-2.mit   Sample3-3.mit  Sample3-4.mit   Sample3-5.mit
>combination         1 vs 2                  1 vs 3                 2 vs 3
72.99098211          3.903      0.0022       3.817      0.0025      0.938      0.1878
73.65656016          1.236      0.1257       2.612      0.0154      1.892      0.0475
…
…
```

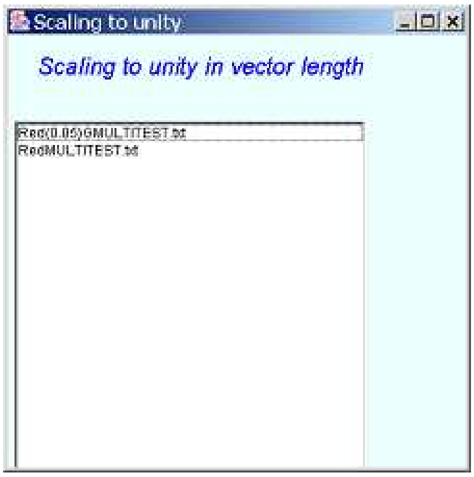
2.9 Peak Reduction

The *m/z* for statistically significant differences of the intensity between pairs of groups with *p*-value smaller than the threshold is selected. Thus, noisy intensities can be removed from the multivariate analysis.

Input file	**PGMULTI, GMULTI**
Output file	**RED(thr)**; thr corresponds to *p*-value set by user.
Execution	**[1] Input threshold of *p*-value**. **[2] Click filename.**  Peak reduction P-Value for Peak Reduction= 0.05 PGMULTITEST.txt
Output file format	(RED format described in 2.6 ‘M to R’)

2.10. Scaling

For each *m/z*, peak intensities for multiple measurements are normalized to unity in sum of square.

Input file	**RED**
Output file	**REDS**
Execution	**[1] Click filename.**  Scaling to unity Scaling to unity in vector length Red(0.05)GMULTITEST.txt RedMULTITEST.txt
Output file format	(RED format described in 2.6 'M to R')

In M measurements, the intensities for individual *m/z* values are represented by a data matrix **X**. Here, the number of IMCs and is denoted by S are denoted by S and N, respectively.

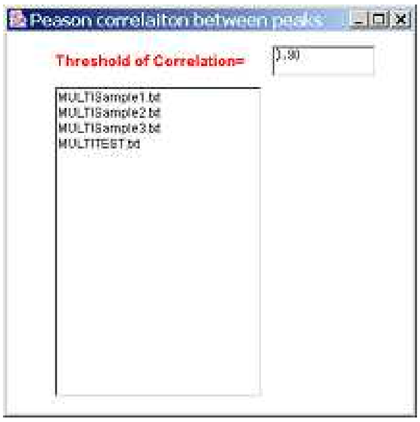
$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ y_{N1} & \dots & y_{Nj} & \dots & y_{NM} \end{pmatrix} \tag{2.10.1}$$

For each *m/z*, peak intensities for multiple measurements are scaled by using Eq. (2.10.2).

$$x''_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{M} x_{ij}}} \tag{2.10.2}$$

2.11. Pearson correlation

Pearson correlations of the intensities for pairs of *m/z* larger than the threshold set by the user are list up.

Input file	**MULTI**
Output file	**Pearson(thr)**
Execution	**[1] Input threshold of Correlation**. **[2] Click filename.**  Peason correlaiton between peaks Threshold of Correlation= 0.90 MULTISample1.txt MULTISample2.txt MULTISample3.txt MULTITEST.txt
Output file format	1st line started with '>Threshold' corresponds to the threshold set by the user. 2nd to the last lines correspond to pairs of peaks (*m/z* values) and their Pearson correlations. The numbers in parentheses represent index number for peaks in the input file. The index numbers of peaks (*m/z*'s) are assigned from up to down in the input file. >Threshold>=0.9 72.99283739615221(1) 109.48510305120904(7) 0.9802398346382768 72.99283739615221(1) 153.35659328616688(17) 0.9455749879863627 72.99283739615221(1) 166.05649834186937(19) 0.9856522752844742 ….. …..

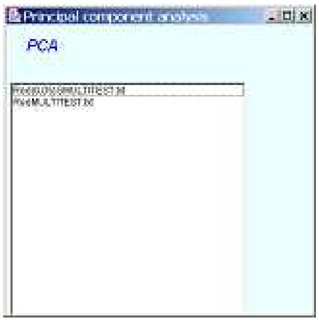
(iii) Multivariate Analysis

2.12. Peak-PCA and its Viewer

Principal component analysis for peaks and visualization of its results. In PCA, three types of parameters, Score, Factor loading, and %Var are used for interpreting multivariate data. In Peak-PCA, variables correspond to experiments, and objects correspond to peak intensities for individual *m/z* values. So Score is calculated for the peak intensities for each *m/z*, and Factor loading (correlation between *s*th principal component and *t*th variable) is calculated for pairs between a vector for *s*th principal components and a vector corresponding to *t*th experiment.

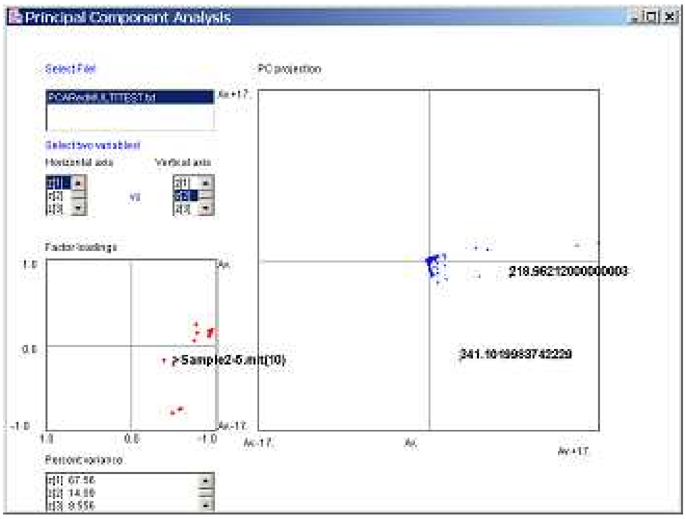
<table>
<tr><td><b>Input file</b></td><td><b>RED</b></td></tr>
<tr><td><b>Output file</b></td><td><b>PCA</b></td></tr>
<tr><td><b>Execution</b></td><td>

[1] Click filename.



[2] After finishing execution of Principal component analysis, the results of PCA can be visualized by **clicking 'viewer'** button on right side of 'Peak-PCA' button.

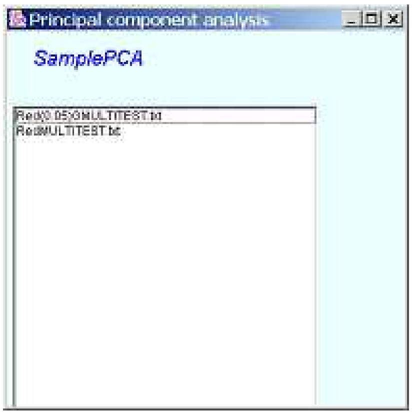
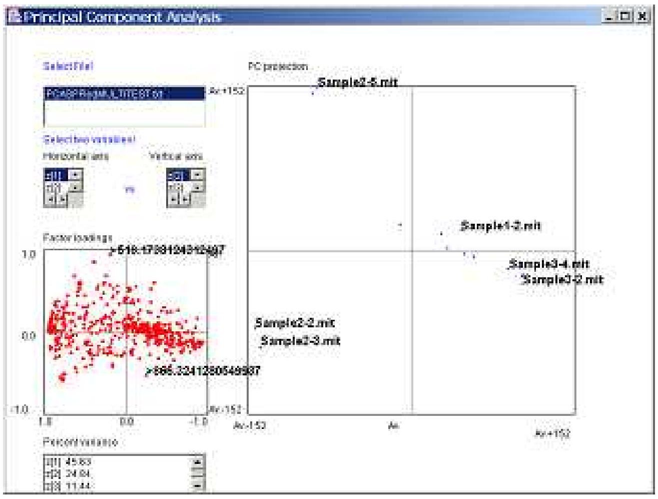
[3] Select file and two variables, then we can obtain PC projection for *m/z* and Factor loadings for measurements. Information of *m/z* and measurements are obtained by clicking dots in the plots, PC projection and Factor loadings, respectively. Percent variance is also listed up in the left of the downside.



</td></tr>
</table>

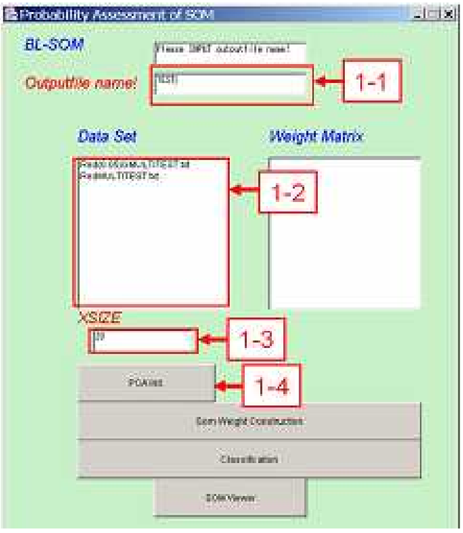
2.13. Sample-PCA and its Viewer

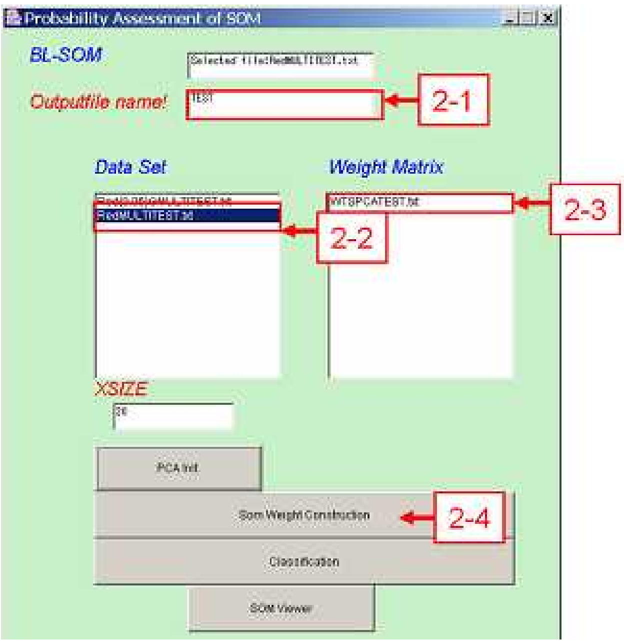
In Sample-PCA, variables correspond to *m/z*, and objects correspond to peak intensities for individual experiments. So Score is calculated for the peak intensities for each experiment, and Factor loading is calculated for pairs between a vector for *s*th principal components and a vector corresponding to *t*th *m/z*.

Input file	**RED**
Output file	**PCASP**
Execution	**[1] Click filename.**  Red(0.05)GMULTITEST.txt RedMULTITEST.txt **[2]** After finishing execution of Principal component analysis, the results of PCA can be visualized by **clicking 'viewer'** button on right side of 'Peak-PCA' button. [3] Select file and two variables, then we can obtain PC projection for measurements and Factor loadings for *m/z*. Information of measurements and *m/z* are obtained by clicking dots in the plots, PC projection and Factor loadings, respectively. Percent variance is also listed up in the bottom left.  Principal Component Analysis Select File! PCASPRedMULTITEST.txt Select two variables! Horizontal axis Vertical axis vs Factor loadings 1.0 0.0 -1.0 1.0 0.0 -1.0 >516.4708124312497 >865.3241280549937 PC projection Av+152 Av-152 Av-152 Av Av+152 Sample2-5.mit Sample1-2.mit Sample3-4.mit Sample3-2.mit Sample2-2.mit Sample2-3.mit Percent variance [1] 45.63 [2] 24.84 [3] 11.44

2.14. BLSOM and its Viewer

Batch-learning SOM for peaks and visualization of its results

Input file	**RED**
Output file	**CLSOM**
Execution	**I. Construction of Self-organizing map** **1. Constructing of initial weight vectors by PCA** **[1-1] Input Outputfile name.** **[1-2] Select Inputdata** **[1-3] Set the number of weights in x size**, then y size is automatically determined by variance ratio of the first and second principal components determined by PCA. **[1-4] Click 'PCA Init.' button.**  Probability Assessment of SOM BL-SOM Please INPUT output file name! Outputfile name! TEST 1-1 Data Set Weight Matrix 1-2 XSIZE 1-3 PCA Init 1-4 Som Weight Construction Classification SOM Viewer After execution of 'PCA Init.', a weight matrix file whose name is given by user and automatically added by 'WTSPCA' in the head is constructed in Weight Matrix. **II Learning process by Data Set and Initial Weight Matrix** **[2-1] Input outputfile name.** **[2-2, 2-3] Select an inputfile and its corresponding initial weight matrix** in Data Set and Weight Matrix, respectively. **[2-4] Click 'Som Weight Construction' button.**



Selected file:RedMULTITEST.txt

Outputfile name! TEST

Data Set: Red0.05SOMULTITEST.txt, RedMULTITEST.txt

Weight Matrix: WTSPCATEST.txt

XSIZE: 20

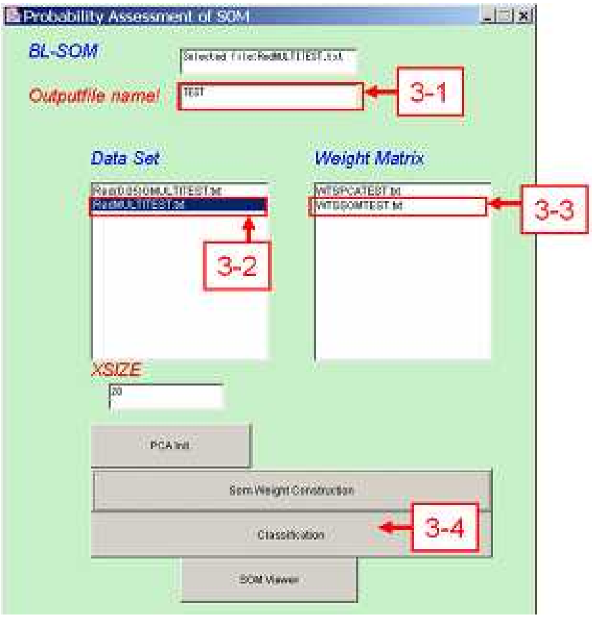
After execution, weight vectors optimized by input vectors are constructed in a filename with WTSSOM.

III Classification of objects (*m/z*)

[3-1] Input inputfile name.

[3-2, 3-3] Select an inputfile and its corresponding weight matrix started with WTSSOM in Data Set and Weight Matrix, respectively.

[3-4] Click 'Classification' button.



Selected file:RedMULTITEST.txt

Outputfile name! TEST

Data Set: Red0.05SOMULTITEST.txt, RedMULTITEST.txt

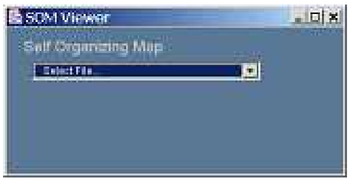
Weight Matrix: WTSPCATEST.txt, WTSSOMTEST.txt

XSIZE: 20

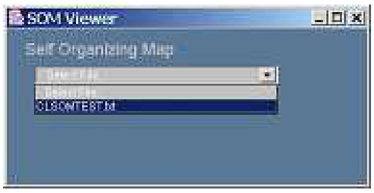
IV Visualization of Classification of objects (*m/z*)

(This process is the same as that of Viewer on the right of BL-SOM.

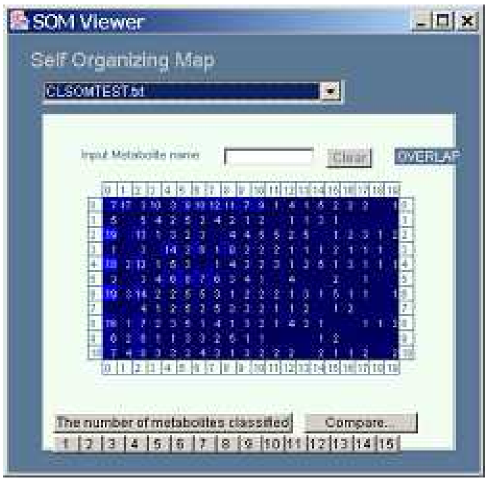
[4-1] Click ‘SOM Viewer’.



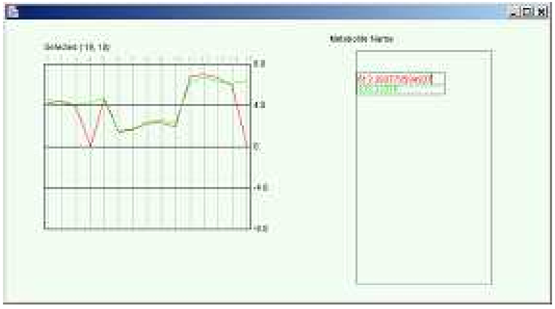
[4-2] Select file.



Then, SOM Viewer is displayed.

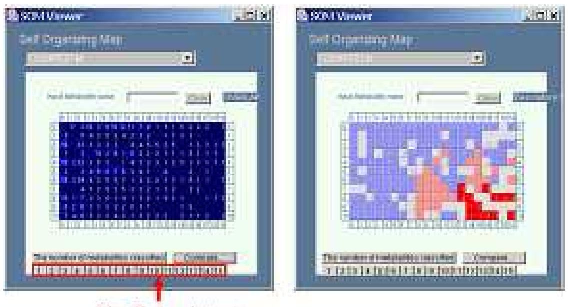


Profile analysis
Click a square, *m/z* values with similar profiles in multiple measurements are displayed. The following example is the profiles in the square at X=19 and Y=10. Two profiles in *m/z* with 255.2329 and 348.1023 are very similar in the multiple measurements.



Characterization of individual measurements

When a user wants to know high and low levels corresponding to individual experiments, he/she should click experiment ID. In this example the 5th experiment has been selected. Pink and Red lattices include only objects with measurements larger than the average for the selected experiment. Sky blue and Blue lattices include only objects with measurements smaller than the average for the selected experiment. A red lattice indicates that at least one of the objects belonging to it is with a measurement value larger than the average plus the standard deviation and a blue lattice indicates that at least one of the objects belonging to it is with a measurement value smaller than the average minus the standard deviation.



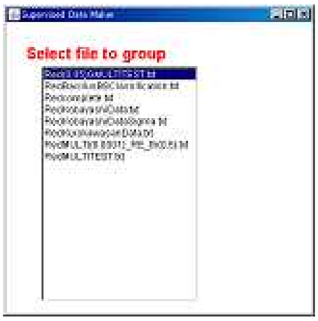
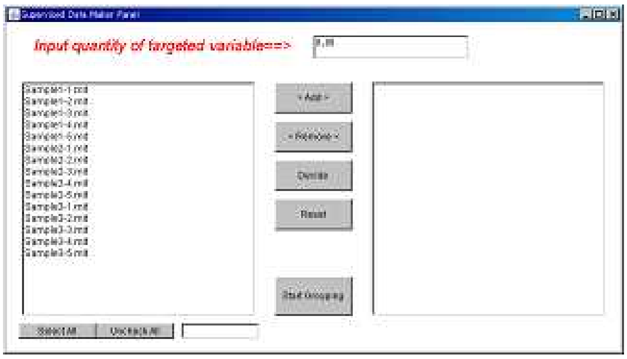
Click Feature button!

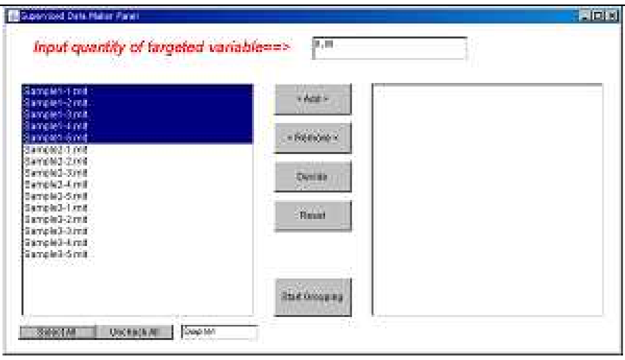
BL-SOM package is also available in our laboratory http://kanaya.naist.jp/SOM/, and is applied to several works as bioinformatics tool as follows.

1. S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, T. Ikemura. Analysis of codon usage diversity for bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome., *Gene*, 276, 89-99 (2001)
2. T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, T. Ikemura, Informatics for unvailing hidden genome signature., *Genome Res.*, 13, 693-702 (2003).
3. M. Hirai, M. Yano, D. Goodenowe, S. Kanaya, T. Kimura, M.Awazuhara, M. Arita, T. Fujiwara, K. Saito, Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*, *Proc. Natl. Acad. Sci., USA*, 101, 10205-10210 (2004).
4. M. Hirai, M. Klein, Y. Fujikawa, M. Yano, D.B. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, J. Papenbrock, K. Saito, Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *J. Biol. Chem*., 280, 25590-5 (2005).

2.15 Supervised Data Maker

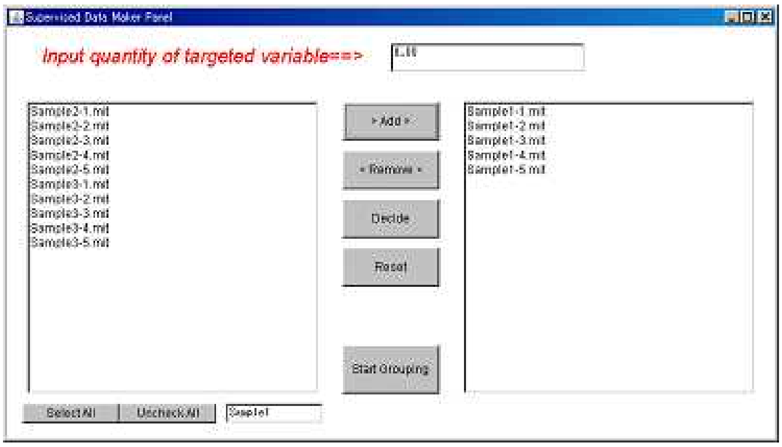
Construction of multivariate data for PLS.

Input file	**Red**
Output file	**SRed**
Execution	**[1] Click Supervised Data Maker button**, then ‘Select files to group’ panel is displayed.  **[2] Select a filename**, then ‘Supervised Data Maker Panel’ is displayed.  Supervised Data Maker Panel Input quantity of targeted variable==> Sample1-1.rmd Sample1-2.rmd Sample1-3.rmd Sample1-4.rmd Sample1-5.rmd Sample2-1.rmd Sample2-2.rmd Sample2-3.rmd Sample2-4.rmd Sample2-5.rmd Sample3-1.rmd Sample3-2.rmd Sample3-3.rmd Sample3-4.rmd Sample3-5.rmd < Add > < Remove > Divide Reset Start Grouping Select All Uncheck All **[3]** Select desired data by **clicking any number of data** or **‘Select All’ button.** If ‘Select All’ button is clicked then all data whose name starts with the text in the textbox are automatically selected.


Supervised Data Maker Panel
Input quantity of targeted variable==>
> Add >
< Remove <
Decide
Reset
Start Grouping
Select All
Uncheck All

Sample1-1.mit
Sample1-2.mit
Sample1-3.mit
Sample1-4.mit
Sample1-5.mit
Sample2-1.mit
Sample2-2.mit
Sample2-3.mit
Sample2-4.mit
Sample2-5.mit
Sample3-1.mit
Sample3-2.mit
Sample3-3.mit
Sample3-4.mit
Sample3-5.mit

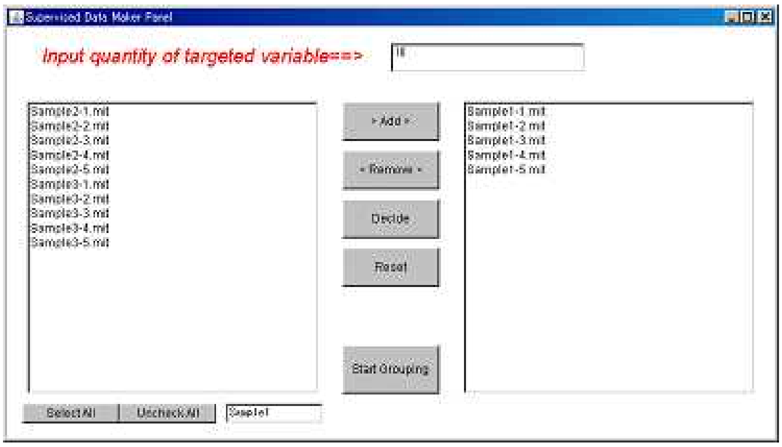
[4] Click '> Add >' button, then the selected data are moved from left to right side.


Supervised Data Maker Panel
Input quantity of targeted variable==>
> Add >
< Remove <
Decide
Reset
Start Grouping
Select All
Uncheck All

Sample2-1.mit
Sample2-2.mit
Sample2-3.mit
Sample2-4.mit
Sample2-5.mit
Sample3-1.mit
Sample3-2.mit
Sample3-3.mit
Sample3-4.mit
Sample3-5.mit

Sample1-1.mit
Sample1-2.mit
Sample1-3.mit
Sample1-4.mit
Sample1-5.mit

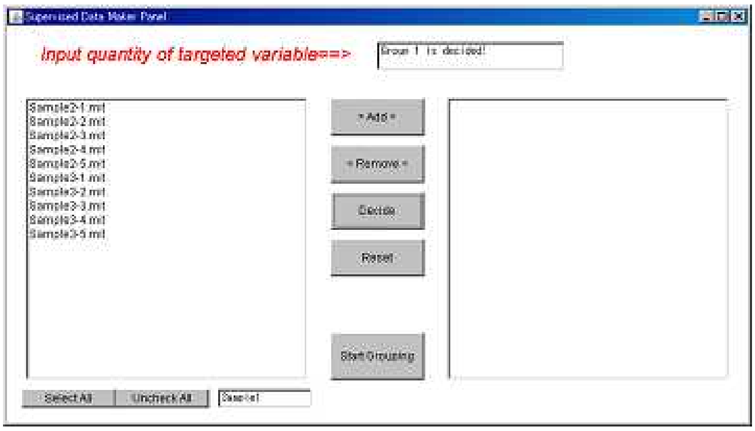
[5] Input quantity of targeted variable, in the corresponding textbox.


Supervised Data Maker Panel
Input quantity of targeted variable==>
> Add >
< Remove <
Decide
Reset
Start Grouping
Select All
Uncheck All

Sample2-1.mit
Sample2-2.mit
Sample2-3.mit
Sample2-4.mit
Sample2-5.mit
Sample3-1.mit
Sample3-2.mit
Sample3-3.mit
Sample3-4.mit
Sample3-5.mit

Sample1-1.mit
Sample1-2.mit
Sample1-3.mit
Sample1-4.mit
Sample1-5.mit

[6] Click 'Decide' button. So selected data is removed and their group is assigned to the output file started with 'S'.

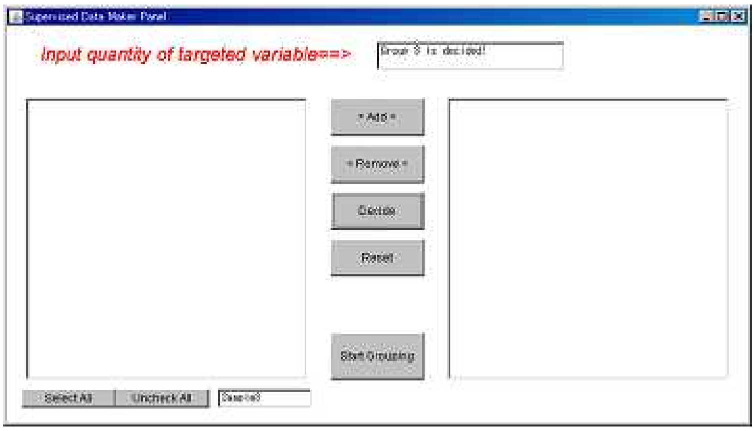

Supervised Data Maker Panel
Input quantity of targeted variable==>
* Add *
* Remove *
Decide
Reset
Start Grouping
Select All
Uncheck All

Group 1 is decided!

Sample2-1.mit
Sample2-2.mit
Sample2-3.mit
Sample2-4.mit
Sample2-5.mit
Sample3-1.mit
Sample3-2.mit
Sample3-3.mit
Sample3-4.mit
Sample3-5.mit

Sample1

[7] Continue the manipulation of [3] [4] [5] and [6] till groups for all data are assigned.


Supervised Data Maker Panel
Input quantity of targeted variable==>
* Add *
* Remove *
Decide
Reset
Start Grouping
Select All
Uncheck All

Group 3 is decided!

Sample3

[8] Click 'Start Grouping' button, then grouping process is started.

Output file format

Lines started with '>group' correspond to the numbers of individual groups. The line with '>no' corresponds to group ID, the line with 'filename' corresponds to the labels for individual measurements, and after that the *m/z* values and their intensities are arranged.

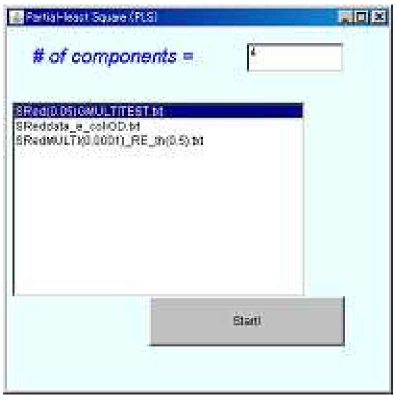
```
>group1              5
>group2              5
>group3              5
>no.                 1             …        1              2             …
>filename            Sample1-1.mit …        Sample1-5.mit  Sample2-1.mit …
>Target              13.5          …        13.5           15.7          …
63.67552147001616    0.0           …        0.0            0.0476        …
72.99098211088538    0.0832        …        0.0836         0.0635        …
….
….
```

2.16 PLS and its Viewer

Calculation of regression equation by using Partial Least Squares Regression and verifying calculations.

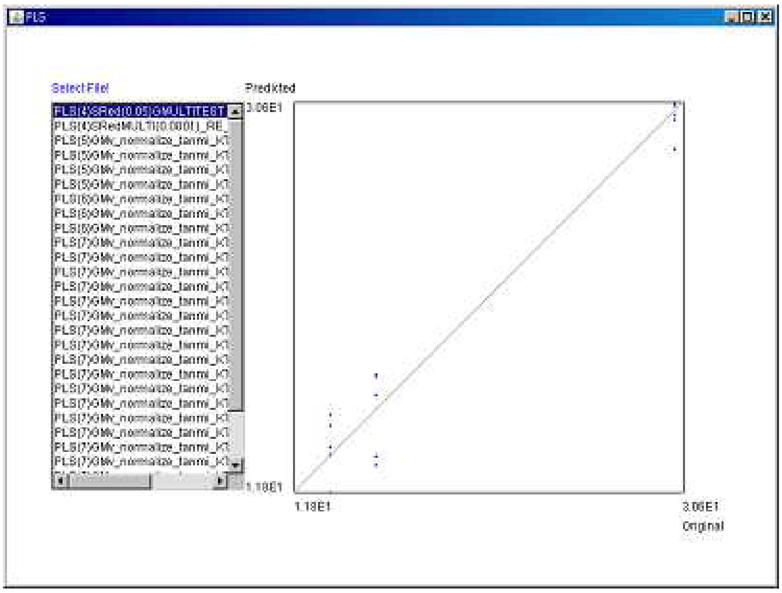
<table>
<tr><td><b>Input file</b></td><td><b>SRED</b></td></tr>
<tr><td><b>Output file</b></td><td><b>CoefPLS(#), PLS(#)</b></td></tr>
<tr><td><b>Execution</b></td><td>

[1] Click filename and **input number of components.**



[2] After finishing execution of Partial Least Squares Regression, the results of PLS can be verified by **clicking 'viewer'** button on right side of 'PLS' button.

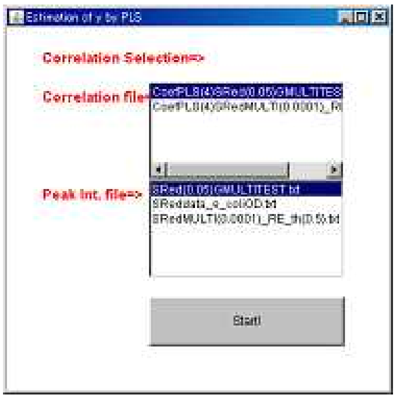
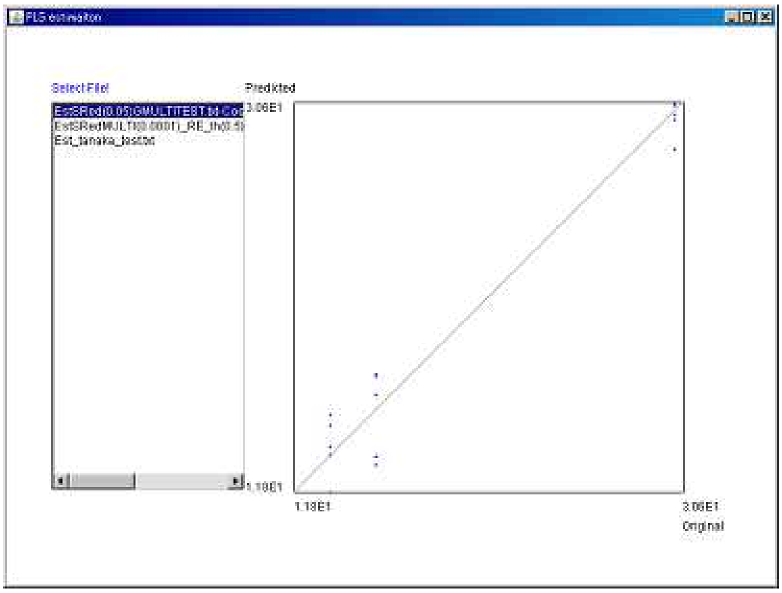
[3] By selecting file, we can verify comparison result between original data and predicted data obtain from regression equation corresponding to the selected file.


PLS
Select File!
FLS(4)SRed(0.05)GMULTITEST
PLS(4)SRedMULTI(0.0001)_RE_
PLS(5)GMv_normalize_tanmi_KT
PLS(5)GMv_normalize_tanmi_KT
PLS(5)GMv_normalize_tanmi_KT
PLS(5)GMv_normalize_tanmi_KT
PLS(6)GMv_normalize_tanmi_KT
PLS(5)GMv_normalize_tanmi_KT
PLS(6)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
PLS(7)GMv_normalize_tanmi_KT
Predicted
3.06E1
1.18E1
1.18E1
3.06E1
Original

</td></tr>
</table>

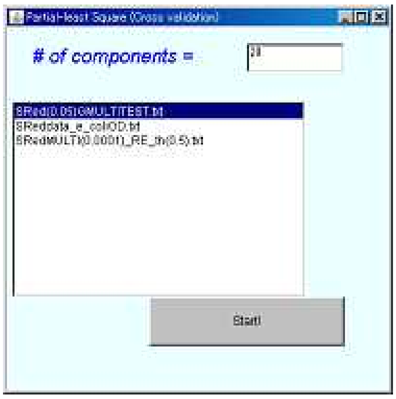
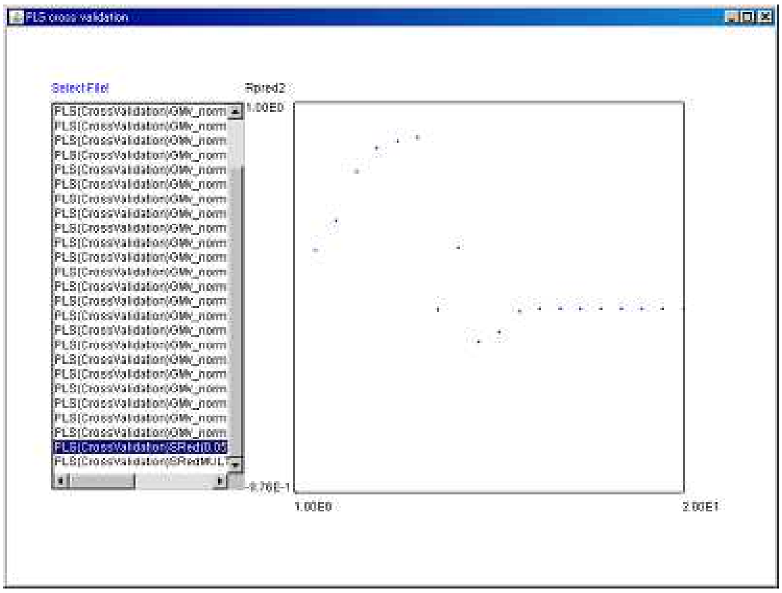
2.17 Estimation by PLS model and its Viewer

Estimation by PLS model and verifying calculations.

Input file	**CoefPLS(#), RED**
Output file	**Est**
Execution	**[1] Click filename.**  Estimation of y by PLS Correlation Selection=> Correlation file=> CoefPLS(4)SRed(0.05)GMULTITES CoefPLS(4)SRedMULTI(0.0001)_R Peak Int. file=> SRed(0.05)GMULTITEST.txt SReddata_e_coliOD.txt SRedMULTI(0.0001)_RE_th(0.5).txt Start! **[2]** After finishing execution of estimation by Partial Least Squares Regression, we can verify calculations by **clicking ‘viewer’** button on right side of ‘Estimation by PLS model’ button. **[3] By selecting file,** we can verify comparison result between original data and predicted data obtain from regression equation corresponding to the selected file.  PLS estimation Select File! EstSRed(0.05)GMULTITEST.txt-Coe EstSRedMULTI(0.0001)_RE_th(0.5) Est_tanaka_test.txt Predicted 3.06E1 1.18E1 1.18E1 3.06E1 Original

2.18 PLS (cross-validation)

Calculate the optimum number of latent parameter by using cross validation.

Input file	**SRed**
Output file	**PLS(CrossValidation)**
Execution	**[1] Click filename** and **input number of components.**  Partial-least Square (Cross validation) # of components = 20 SRed(0.05)GMULTITEST.txt SReddata_e_coliOD.txt SRedMULTI(0.0001)_RE_th(0.5).txt Start! **[2]** After finishing execution of cross validation, the following window can be displayed by **clicking 'viewer'** button on right side of 'PLS (cross-validation)' button. **[3] By selecting file,** we can determine optimum number of components corresponding to maximum Rpred2.  PLS cross validation Select File! FLS(CrossValidation)GMr_norm FLS(CrossValidation)SRed(0.05 FLS(CrossValidation)SRedMULT Rpred2 1.00E0 9.76E-1 1.00E0 2.00E1