COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 4, Issue: 5, January 2013, e201301010, http://dx.doi.org/10.5936/csbj.201301010

CSBJ

Data Mining Methods for Omics and Knowledge of Crude Medicinal Plants toward Big Data Biology

Farit M. Afendi [a,b], Naoaki Ono [a], Yukiko Nakamura [a], Kensuke Nakamura [d], Latifah K. Darusman [c], Nelson Kibinge [a], Aki Hirai Morita [a], Ken Tanaka [e], Hisayuki Horai [f], Md. Altaf-Ul-Amin [a], Shigehiko Kanaya [a,*]

Abstract: Molecular biological data has rapidly increased with the recent progress of the Omics fields, e.g., genomics, transcriptomics, proteomics and metabolomics that necessitates the development of databases and methods for efficient storage, retrieval, integration and analysis of massive data. The present study reviews the usage of KNApSAcK Family DB in metabolomics and related area, discusses several statistical methods for handling multivariate data and shows their application on Indonesian blended herbal medicines (Jamu) as a case study. Exploration using Biplot reveals many plants are rarely utilized while some plants are highly utilized toward specific efficacy. Furthermore, the ingredients of Jamu formulas are modeled using Partial Least Squares Discriminant Analysis (PLS-DA) in order to predict their efficacy. The plants used in each Jamu medicine served as the predictors, whereas the efficacy of each Jamu provided the responses. This model produces 71.6% correct classification in predicting efficacy. Permutation test then is used to determine plants that serve as main ingredients in Jamu formula by evaluating the significance of the PLS-DA coefficients. Next, in order to explain the role of plants that serve as main ingredients in Jamu medicines, information of pharmacological activity of the plants is added to the predictor block. Then N-PLS-DA model, multiway version of PLS-DA, is utilized to handle the three-dimensional array of the predictor block. The resulting N-PLS-DA model reveals that the effects of some pharmacological activities are specific for certain efficacy and the other activities are diverse toward many efficacies. Mathematical modeling introduced in the present study can be utilized in global analysis of big data targeting to reveal the underlying biology.

Mini Review Article

1. Introduction

Data-intensive sciences have progressed in modern astronomy [1], biology [2-8], computational materials science [9], ecology [10-11] and social science [12] because open-access data has increased drastically. Data-intensive or -driven discovery in biology requires a large open pool of data across the full breadth of the life sciences and the access to the pool will invite "New" logic, strategies and tools to discover new trends, associations, discontinuities, and exceptions that reveal aspects of the underlying biology [2, 5, 6]. Big data biology, which is a discipline of data-intensive science, was proposed based on the rapid increasing of omics data produced by genomics, transcriptomics, proteomics and metabolomics [2-8]. This situation is also a feature of the ethnomedicinal survey and the number of medicinal plants is estimated to be 40,000 to 70,000 around the world [13] and many countries utilize these plants as blended herbal medicines, e.g., China (traditional Chinese medicine), Japan (Kampo medicine), India (Ayruveda, Siddha and Unani) and Indonesia (Jamu). Blended herbal medicines as well as single herb medicines include a large number of constituent substances which exert effects on human physiology through a variety of biological pathways. To comprehensively understand the medicinal usage of plants based upon traditional and modern knowledge, we add to KNApSAcK Family database systems the selected herbal ingredients i.e., the formulas of Kampo and Jamu, omics information in plants and humans, and physiological activities in humans [14-16]. These information need to be connected in a way that enables scientists to make predictions based on general principles.

In this mini-review, we discuss the usage of KNApSAcK Family DB in metabolomics, explain mining techniques such as principal component analysis (PCA), partial least square regression (PLSR) and multiway model, and show their application on Indonesian blended herbal medicines (Jamu) as a case study.

2. KNApSAcK Family Database

Omics biology, like most scientific disciplines, is in an era of accelerated increase of data, so called big data biology [2-8]. Large-scale sequencing centers, high-throughput analytical facilities and

[a]*Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0101, Ikoma, Japan*
[b]*Department of Statistics, Bogor Agricultural University, Jln. Meranti, Kampus IPB Darmaga, Bogor 16680, Indonesia*
[c]*Biopharmaca Research Center, Bogor Agricultural University, Kampas IPB Taman Kencana, Jln. Taman Kencana No. 3 Bogor 16151, Indonesia*
[d]*Maebashi Institute of technology, 450-1 Kamisadori, Maebashi-shi, Gunma, 371-0816 Japan*
[e]*Department of Medicinal Resources, Institute of Natural Medicine, University of Toyama, 2630 Toyama, 930-0194, Japan*
[f]*Department of Electronic and Computer Engineering, Ibaraki National College of Technology, 866 Nakane, Hitachinaka, Ibaraki 312-8508, Japan*

* Corresponding author.
E-mail address: skanaya@gtc.naist.jp (Shigehiko Kanaya)

individual laboratories produce vast amounts of data such as nucleotide and protein sequences, gene expression measurements, protein and genetic interactions, mass spectra of metabolites and phenotype studies. The goal of investigating the interactions between medicinal/edible plants and humans is to comprehensively understand the molecular mechanism of medicinal plants on human physiology based on current and traditional knowledge. Optimization of blended herbal formulas should be developing using information derived from plant and human omics. To reach this goal we need to develop databases based on the platform shown in Fig. 1A. KNApSAcK family DBs have been developed for this purpose [14-16]. Relations among individual DBs are illustrated in Fig. 1A and main page of KNApSAcK Family DB is shown in Fig. 1B.

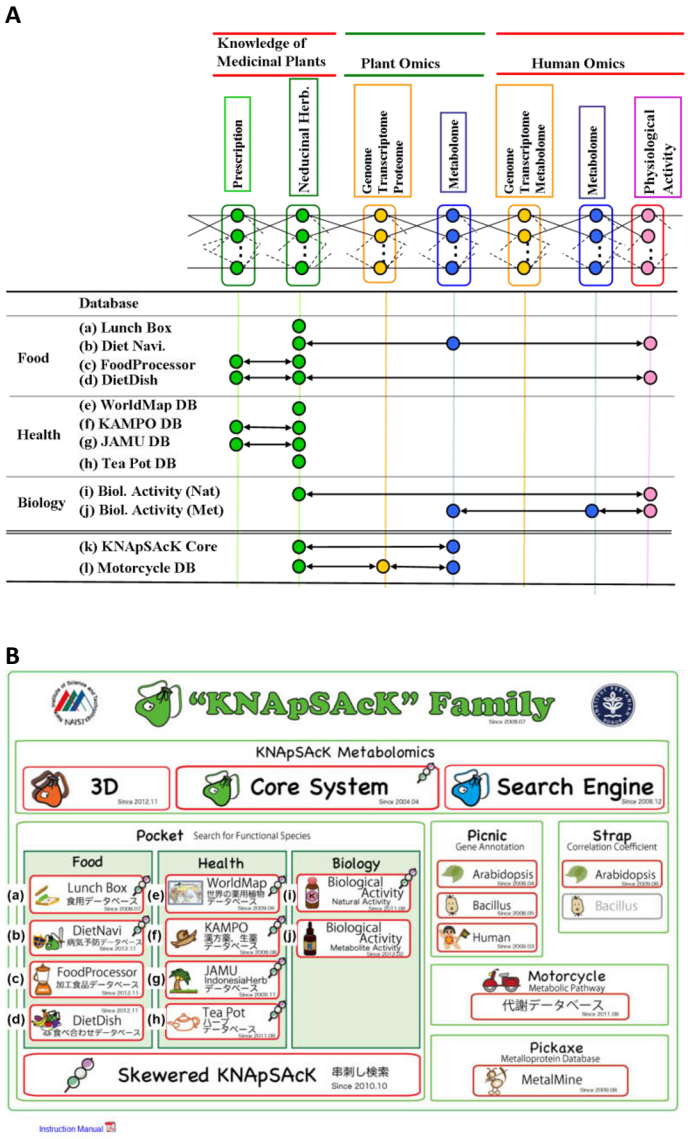

Figure 1. Integrated platform of knowledge of medicinal plants and plant and human –omics and KNApSAcK Family databases. **(A)** The relations of attributes among individual DBs. **(B)** Main window of KNApSAcK Family DB, indexes from a to i in panel A correspond to those in panel B.

Four DBs (Lunch Box DB, DietNavi DB, Food Processor DB and DietDish DB, a-d in Fig. 1) are about Food & Health related with Japanese foods and ingredients explained in Japanese language because initially we developed them targeting the Japanese people, but we are planning to translate them into English as early as possible. Lunch Box DB comprises information on 800 edible species which include the species introduced to Japan from outside or originally grown in Japan, general information of the crops and the effect of them on human health.

Noncommunicable diseases such as heart disease, metabolic disease, cancer and respiratory disease, which superseded the infectious diseases because of the development and widespread distribution of vaccines and antimicrobial drugs, account for 60% of all deaths worldwide and 80% of deaths in low- and middle-income countries [17]. Food and ingredients in sanative diet and more effective combination of foods beneficial against those noncommunicable diseases are accumulated in DietNavi and DietDish DBs, respectively (b and d in Fig. 1). FoodProcessor DB comprises 309 retortable pouch foods encompassed by 261 food ingredients produced in Japan, and connected with DietNavi and KNApSAcK core by species names of foods.

To systematize crude drugs by multifaceted view points, we have developed four DBs (WorldMap, KAMPO, JAMU and TeaPot DBs as shown in e-h of Fig. 1). The KNApSAcK WorldMap DB comprises 46,256 geographic zone-plant pair entries in 217 geographical zones except mini-states such as the Principalities of Liechtenstein, Monaco and Andorra, and the Vatican City. Prescriptions corresponding to Japanese and Indonesian herbal medicines have been accumulated in KAMPO and JAMU DBs, respectively. KAMPO DB is comprised of 1,581 primary formulas classified in to 336 formula names encompassed by 278 medicinal plants which are approved by the National health insurance authority in Japan. JAMU DB is comprised of 5,310 formulas encompassed by 550 medicinal plants and 12 anatomical regions which are approved by the National Agency of Drug and Food Control (NA-DFC) of Indonesia. Medicinal/edible plants reported in the scientific literature have been classified into geographic zones using the International Organization for Standardization (ISO3166), which defines geographic zones based on the borders between nations and small islands. Herbs are defined as any plants with leaves, seeds, and flowers used for flavoring, food, medicine, perfume and parts of such a plant as used in cooking. Those are accumulated in TeaPot DB.

Two types of biological activities, that is, activities of natural resources and metabolites to other species including human, i.e., antibiotic, anticancer and so on are accumulated in Natural Activity and Metabolite Activity DBs (Fig. 1B), respectively. The former and the latter comprised 33,703 and 6,677 entries, respectively. For extension of species-metabolite relationship DB to metabolic pathways, it is needed to design secondary metabolic pathway DB for detection of metabolic pathways based on enzyme reactions and prediction of reactions by peptide sequences. So we have developed Motoercycle DB containing 2,421 entries. The metabolomics of plants is developing rapidly [18-20 and references in Table 1], and it will be an important topic in the systems-biological studies of interactions between plants and humans, which is included in the topics of big data biology [2-8], with the goal of achieving a holistic understanding of plant function and healthcare, including the activity of medicinal plants as well as interaction between plants and their environment [14-16, 21, 22].

To facilitate access to metabolite information obtained from analytical techniques, we have developed species-metabolite relationship DB (KNApSAcK Core DB) which contains 106,418 species-metabolite relationships encompassing 21,705 species and 50,897 metabolites. Nine databases of KNApSAcK family (except DietDish) are connected with KNApSAcK Core DB to easily obtain candidates of secondary metabolites in species utilized in several purposes [23]. The KNApSAcK Core DB was utilized in very

Table 1. Studies that cite KNApSAcK Core DB.

Article type	The purpose of study [References]
< 2006-2008 >	
Review	Bridge between Chemistry and Biology [24], GC-MS DB [29], Metabolomics technologies [31], Functional genomics research strategy of combining transcriptome and metabolome [32], The role of MS in metabolomics [34], Mass spectrometry platforms [38], Metabolomics technologies and functional genomics platform [42], Technology and informatics [49], Atmospheric pressure ionization mass spectrometry [52]
Exp	Metabolite accumulation caused by herbicidal enzyme inhibitors [30], Assignment of UGT89C1 to a flavonol 7-O-rhamnosyltransferase [33], Light/dark regulation of metabolite activities [35], Characterization of mutants in flavonoid and phenylpropanoid biosynthetic pathways [37], Metabolism of dietary phytochemicals [39], Metabolic networks in primary and secondary pathways for achene and receptacle [40], High-resolution mass spectrometry and 13C-isotope labeling of entire metabolomes [41], Phenolic biosynthesis pathway [43], Metabolic profiling in strawberry receptacle development [44], Regulation of glucosinolate biosynthesis [46], Integrated analysis of metabolome and transcriptome [48], Protocol in metabolite fingerprints [50]
Bioinfo	Metabolome platform DrDMASS in FT-ICR-MS [25], Taxonomic diversity of flavonoids [26], MS Peak storage and processing [28], Metabolite annotation based on MS and MS2 [45], Identification of metabolites by MS and MS-tagged MS2 data [47], Metabolome platform DrDMASS in FT-ICR-MS [51]
DB	Chemical biology [27], Metabolome tools and databases [36]
< 2009 >	
Review	Integrated omics [58], MS-based technologies [59], Web-resources in MS-based metabolomics [75], Functional genomics [78]
Exp	Metabolic profiling in cold-temperature [56], Antioxidant compounds in white cabbage during winter storage [60], Hydroxylation of fatty acids by P450 proteins [62], Dietary phytochemicals and human [63], Classification of Ephedra sp. [67], Selection of metabolites [68], Matrix-assisted laser desorption/ionization mass spectrometry [69], Determination of gene function [70], Quality assessment [73, 74], Diarylheptanoid biosynthesis [77]
Bioinfo	Annotation of metabolite information to MS [53], Tools for the annotation of High Resolution MS metabolomics data [57], Comparison of metabolite DB using rice metabolites [61], Assessment of annotation of metabolites using FDR [64], Graph representation of multiple databases [65], Peak detection based on MS/MS patterns [66], Complexity of relation between plants and metabolites [71], Metabolic pathway prediction [72], Metabolite Complexity of relation between plants and metabolites [71], Metabolic pathway prediction [72], Metabolite annotation [76]
DB	Embedded string-search commands on MediaWiki [54]
< 2010 >	
Review	MS data processing [84], Metabolomics in plant ecology and genetics [85], Identification of metabolites [87], FT-ICR-MS, Reaction representation based on van Krevelen diagram [89], Relationship among individual omics data based on multivariate analysis and DB [16], Dietary intake [90], Functional Genomics [92], Annotation of gene function based on co-response gene and identification of metabolites [95]
Exp	Metabolite composition [79], QTL of barley, against Fusarium head blight [80], Changing color of flower from dark purple to white [81], Metabolic profiling of different tissues [86], Quality assessment [94]
Bioinfo	Chemical similarity search and substructure matching of compounds [82], Multiple metabolomics platforms for different types of MS [91], MS data processing [96], Network analysis of species-metabolite relations [97]
DB	MassBank, MS DB [83], Polyphenol contents in foods [88], Binzylsioquinone alkaloids [93]
< 2011 >	
Review	Pesticide research [100], Metabolome DB [108], Traditional medicinal plants [111], Pesticide research [113]
Exp	Hepatotoxicity [55], Subcellular distribution of metabolites [99], Assessment of metabolites of barley against Fusarium head blight [102], Metabolic responses of ultraviolet-B light [103], Transport of 12-Oxo-phytodienoic acid-glutathione into vacuole [104], Demetylation of oligogalacturonides by FAPE1 leads to defense against fungus Botytis cinerea [105], Cytochrome P450, CYP81F4 [109], Imaging mass spectrometry [112]
Bioinfo	QTL informatics [98], Metabolomics in medical purpose with systems chemical biology and chemoinformatics [101], Molecular formula annotation of polar and lipophilic metabolites [107], Metabolic profiling [114]
DB	Food phytochemicals [106], Medicinal plants in Indonesia [110]
< 2012-13 >	
Review	Plant responses to abiotic stress [115], Phytoalexins [118], Plant biotechnology [119], Integrative system biology [121], Systems biology in Japanese traditional Kampo medicine [15]
Exp	Camptothecin biosynthesis [117], Herbivore (*Spodoptera littoralis*)-induced metabolites [120], Natural distance [122], Molecular marker [123], Metabolic changes during fruit maturation [124], Metabolites in seed kernels [125], mQLT [126], Salt and drought stress [127], Mass spectrometric imaging [130], Defence against pathogens (Penicillium digitatum) [131]
Bioinfo	Repository for metabolomics studies [128], Visualization of metabolome data [129]
DB	Metabolite annotation [116]

diverged purposes of metabolomics studies including identification of metabolites ('Exp' in Table 1), construction of integrated databases ('DB'), bioinformatics and systems biology ('Bioinfo') , and cited in at least 110 papers listed in Table 1, that is, in 29 papers in the period of 2006-2008, 25 papers in the period of 2009, 20 papers in 2010, 18 papers in 2011, 18 papers in 2012-2013. In addition, it was applied in diverged species from bacteria to plants and animals, in total 28 species, that is, *Angelica acutiloba* [74], *Arabidopsis lyrata* ssp. *petraea* [56], *Arabidopsis thaliana*[25, 30, 33, 35, 37, 46, 47, 62, 70, 86, 99, 103, 104, 108, 109, 121, 122], *Atriplex halimus* [127], *Bacillus subtilis* [113], *Brassica oleraceae* var *capitata* [60], *Brufelsia calycina* [81], *Capsicum sp.* [123], *Citrus sinensis* [131], *Curcuma longa* [77], *Ephedra* sp. [67], *Escherichia coli* [51], *Fragaria* x *ananassa* [40, 43, 44], *Fragaria vesca* [105], *Glycine max* [53], *Glycyrrhiza uralensis* [94], *Hordeum vulgare* [80, 102], *Homo sapiens* [63, 101], *Jatropha curcas* [124, 125], *Malx* x *domestica* [126], *Ophiorrhiza pumila* [117], *Oryza sativa* [49, 61], *Papaver somniferum* [42], *Rattus norvegicus* [39, 97], *Rizotania solani* [79], *Solanum lycopersicum* [45, 48], *Solanum tuberosum* [98] and *Zea mays* [120].

In the period of 2006-2008, many review papers ['Review' in Table 1] focused on metabolomics platforms integrated by mass-spectrometry and metabolite databases including KNApSAcK Core [29, 31, 34, 38, 42, 49, 52] and on linking chemistry with biology [24], and on metabolome researches targeting the model plant *Arabidopsis thaliana* [30, 33, 35, 37]. In 2009, metabolome studies were extended to diverged species such as crops and medicinal plants [53, 60, 61, 67, 68, 73, 74, 78] and to engineering studies such as quality assessment based on metabolomics [73, 74]. Thus metabolomics was applied from model species to crops and medicinal herbs. In the period of 2010-2013, metabolomics was further extended to genetics such as QTL [80, 98, 126], and to explanation of species by metabolites, i.e., ecological subjects [85] phytoalexins [119], herbivore-induced metabolites [120] and defense against pathogens [131], and to stress responses [115, 116, 127]. In addition, metabolomics has also been tried in imaging studies [112, 129]. Species-metabolite relation database KNApSAcK Core has been utilized in the extended fields of metabolomics researches and the horizon of metabolomics researches could be recognized by reviewing the works that utilized and/or cited the KNApSAcK DB.

Methodologies for multivariate analysis to statistically process the massive amount of metabolome data were reviewed in [16] and to systematize blended herbal medicines in Kampo [15]. In the following section, we focus on the mining studies of blended herbal medicines for systematically understanding the composition of medicinal herbs to efficacies on humans, that is, principal component analysis (PCA) that makes it possible to systematize the ingredient in individual blending systems, partial least squares (PLS) that can relate the ingredients of medicinal herbs to the efficacies and N-PLS that can connect multi-factors to the efficacies. We initially explain individual techniques in Section 3 and then discuss their application in data-mining of blended types of herbal medicines in Section 4.

3. Mathematical Methods of Data Mining

3.1 Principal Component Analysis (PCA)

PCA is a linear transformation of a large number of interrelated variables into a new set of variables, called as the principal components (PCs), which are uncorrelated and ordered so that the first few retain most of the variation present in all the original variables [132].

Consider a data matrix $\mathbf{A} = (\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_p)$ with n observations and let $\mathbf{V}$ (p x p) be the variance-covariance matrix of $\mathbf{A}$. The principal components of $\mathbf{A}$, $\mathbf{Z} = (\mathbf{z}_1\ \mathbf{z}_2\ \ldots\ \mathbf{z}_p)$, are calculated as

$$\mathbf{z}_j = \mathbf{A}\mathbf{c}_j \quad (j = 1, 2, \ldots, p) \tag{1}$$

where $\mathbf{c}_j$ is the j-th eigenvector of $\mathbf{V}$ which correspond to the j-th eigenvalue of $\mathbf{V}$ (λ_j). The properties of PCs are: (1) $\mathrm{Var}(\mathbf{z}_j) = \lambda_j$; (2) $\mathrm{Cov}(\mathbf{z}_j,\mathbf{z}_{j'}) = 0$, $j \neq j'$; (3) $\mathrm{Var}(\mathbf{z}_1) \geq \mathrm{Var}(\mathbf{z}_2) \geq \ldots \mathrm{Var}(\mathbf{z}_p)$. The cumulative proportion of variance of the original variables explained by the first J principal components can be obtained as

$$\Pr(z_j) = \frac{\sum_{j=1}^{J} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{2}$$

3.2 Partial Least Squares

PLSR is a regression method, which assumes underlying factors among the predictors account for most of the response variation [133, 134]. These underlying factors of X-variate

$$\mathbf{T} = \mathbf{X}\mathbf{W} \tag{3}$$

are obtained by maximizing their covariance with the corresponding underlying factors of Y-variate where $\mathbf{X}$ is an $n \times m$ matrix of predictors, $\mathbf{Y}$ is an $n \times p$ matrix of responses, $\mathbf{T}$ is an $n \times c$ matrix of X-score factors, and $\mathbf{W}$ is $m \times c$ matrix of weight. Note that n is the number of observations, m is the number of predictors, p is the number of responses, and c is the number of components.

The X-score factors, i.e. matrix $\mathbf{T}$, have the following properties [133].

a. When multiplied by loadings $\mathbf{P}$, they are good summaries of $\mathbf{X}$, i.e. the X-residuals $\mathbf{E}$ are small

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E} \tag{4}$$

b. The X-score factors are good predictors of $\mathbf{Y}$, i.e.

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^t + \mathbf{F} \tag{5}$$

The Y-residuals $\mathbf{F}$ express the deviations between the observed and modeled responses.

Based on Eq. (3), Eq. (5) can be rewritten as a multiple regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{W}\mathbf{Q}^t + \mathbf{F} = \mathbf{X}\mathbf{B} + \mathbf{F} \tag{6}$$

Thus, PLSR coefficients $\mathbf{B}$ can be written as

$$\mathbf{B} = \mathbf{W}\mathbf{Q}^t \tag{7}$$

whereas prediction of the responses can be obtained from

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}\mathbf{Q}^t \tag{8}$$

Although PLSR is not specifically designed to discriminate among groups, Barker and Rayens [135] have demonstrated that PLSR can be used for such purposes by connecting PLSR and Linear

Discriminant Analysis (LDA); this combined method is called as Partial Least Square Discriminant Analysis (PLS-DA). In PLS-DA, group membership is transformed into a dummy matrix, and this dummy matrix provides the response variables for PLSR.

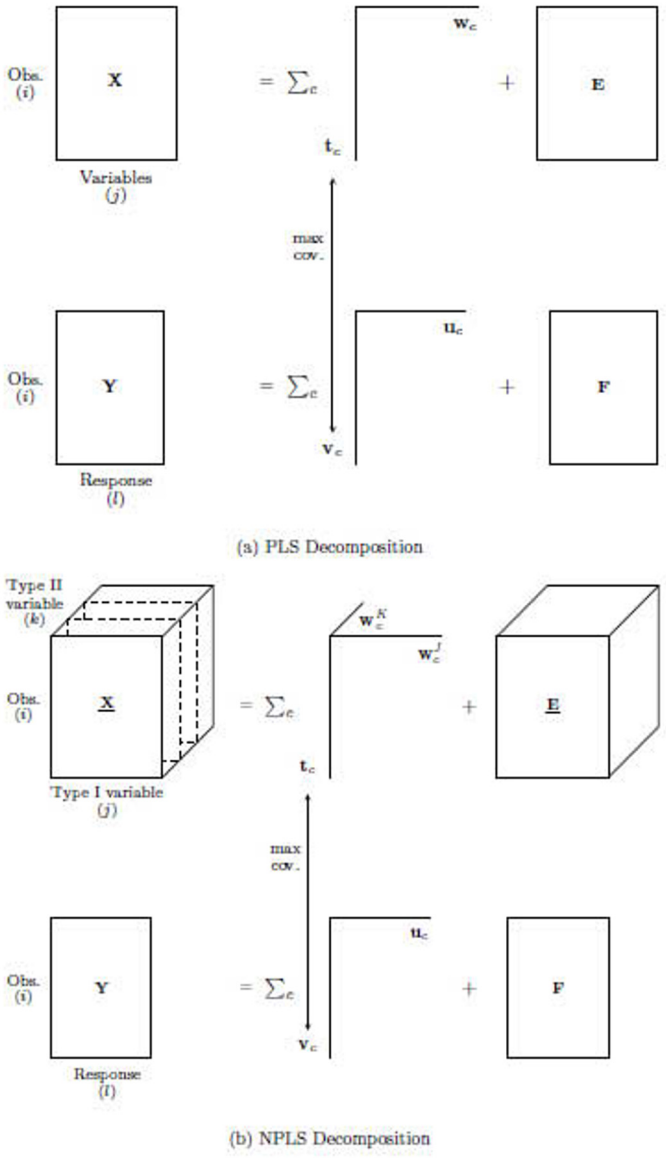

Figure 2. Schematic diagram of the decomposition of both predictor and response blocks for: (a) PLS and (b) N-PLS model.

3.3 Multiway model

An extension of PLSR to deal with multidimensional data known as Multiway Partial Least Squares has been developed by Bro [136] and is called as N-PLS. In this model, the same principle of PLSR for two dimensional data is utilized, that is, both predictor and response blocks are decomposed successively into multi-linear model such that the pairwise scores have maximal covariance. The score of the predictor is then regressed to the response variable. Fig. 2 illustrates the decomposition of N-PLS model. Moreover, N-PLS model can also be used for discrimination purpose, which is called as N-PLS-DA, that is the multiway version of PLS-DA, by utilizing the dummy matrix of group membership as the response variable.

Consider the three-dimensional array $\underline{\mathbf{X}}$ indexed by observation (i = 1, 2, …, I), type I variable (j = 1, 2, …, J) and type II variable (k = 1, 2, …, K). The decomposition of both the predictor and the response block based on N-PLS model are as follows

$$X_{ijk} = \sum_{c=1}^{C} T_{ic} W_{jc}^{J} W_{kc}^{K} + E_{ijk} \tag{9}$$

$$Y_{il} = \sum_{c=1}^{C} V_{ic} V_{lc} + F_{il} \tag{10}$$

The array $\underline{\mathbf{X}}$ is decomposed into a tri-linear model consisting of one score vector for observation called $\mathbf{t}_c$ (I x 1), and two weight vectors, one for type I variable called w_c^J (J x 1) and one for type II variable called w_c^K (K x 1). Similarly, a bi-linear model is used in decomposing the matrix $\mathbf{Y}$ into one score vector $\mathbf{v}_c$ (I x 1) and one weight vector $\mathbf{u}_c$ (L x 1). The decomposition is conducted such that the covariance among the score of predictor $\mathbf{t}$ and the corresponding score of the response $\mathbf{v}$ is maximized. All scores and weights are indexed with c showing that they correspond to cth multiway component, while C represents the total number of multiway components used in N-PLS model. Moreover, $\underline{\mathbf{E}}$ and $\mathbf{F}$ are the residuals of the decomposition of the three-dimensional array $\underline{\mathbf{X}}$ and matrix $\mathbf{Y}$, respectively.

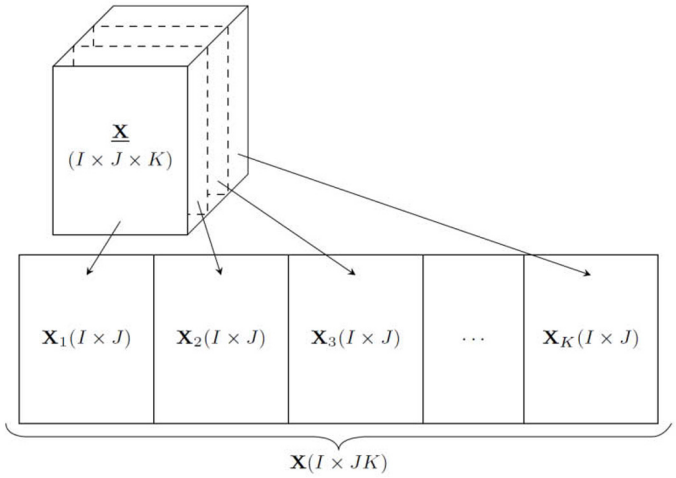

Figure 3. Illustration of matricizing three-dimensional array $\underline{\mathbf{X}}$ (I x J x K) into matrix $\mathbf{X}$ (I x JK).

Furthermore, let $\mathbf{X}_k$ (I x J) be the kth slice of $\underline{\mathbf{X}}$ (I x J x K) for the corresponding kth of type II variable, then matricizing three-dimensional array $\underline{\mathbf{X}}$ into matrix $\mathbf{X}$ (I x JK) is performed as follows [137]

$$\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2 \mid \ldots \mid \mathbf{X}_K] \tag{11}$$

Fig. 3 depicts this unfolding process of array $\underline{\mathbf{X}}$ into matrix $\mathbf{X}$. Using this notation, the score $\mathbf{t}_c$ of the cth component can be calculated as [138]

$$t_c = \mathbf{X}(\mathbf{w}_c^K \otimes \mathbf{w}_c^J)$$

or

$$t_{ic} = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk} w_{jc}^{J} w_{kc}^{K} \quad (12)$$

From Eq. (12), the weight corresponding to cth component, $\mathbf{w}_c$ (JK x 1), can be defined as

$$\mathbf{w}_c = (\mathbf{w}_c^K \otimes \mathbf{w}_c^J) \quad (13)$$

Smilde [140] also described that, due to the deflation in $\mathbf{X}$ during the decomposition, the weight matrix $\mathbf{W}$ (JK x C) can be applied directly to the original unfolded matrix $\mathbf{X}$ is defined as

$$\mathbf{W} = [\mathbf{w}_2 | (\mathbf{I}_{JK} - \mathbf{w}_1\mathbf{w}_1^t)\mathbf{w}_2 | ... | (\mathbf{I}_{JK} - \mathbf{w}_1\mathbf{w}_1^t)(\mathbf{I}_{JK} - \mathbf{w}_2\mathbf{w}_2^t)...(\mathbf{I}_{JK} - \mathbf{w}_{Q-1}\mathbf{w}_{Q-1}^t)\mathbf{w}_Q] \quad (14)$$

Hence, the scores in $\mathbf{T}$ (I x C) expressed directly in terms of the X-columns is

$$\mathbf{T} = \mathbf{XW} \quad (15)$$

After the decomposition procedure, the next step is to regress $\mathbf{Y}$ on the component scores $\mathbf{T}$

$$\hat{\mathbf{Y}} = \mathbf{TB} \quad (16)$$

with

$$\mathbf{B} = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{Y} \quad (17)$$

From Eq. (15) and (16) we have

$$\hat{\mathbf{Y}} = \mathbf{XWB} \quad (18)$$

Therefore, the regression coefficients $\mathbf{B}_{NPLS}$ (JK x L) needed to predict $\mathbf{Y}$ from $\mathbf{X}$ are obtained as

$$\mathbf{B}_{NPLS} = \mathbf{WB} \quad (19)$$

4. Illustration of Data Mining Techniques

Indonesia, the mega-biodiversity center like Brazil, has at least 9,600 species of plants with pharmacological activity [110] and has developed blended herbal medicines called Jamu taking modern and traditional knowledge of herbs into consideration. To prepare Jamu, several plants are selected and mixed such that the concoction has the desired efficacy. Traditionally, plants are chosen based on prior experience which is passed down from generation to generation. In curing a particular disease, each ethnic group in Indonesia may have its own formulas, whose specific nature depends strongly on the local plant resources in the region where a given population lives and the efficacies of Jamu medicines have been empirically demonstrated [139-142]. Data mining techniques with the blended herbal medicine databases such as KAMPO and JAMU (Fig. 1) makes it possible to comprehensively and mathematically understand those blended herbal systems. Fig. 4 illustrates a network connecting efficacy, herbal medicine, plant, and pharmacological activity of plant. The network showing that crude medicines M_1, which is useful for efficacy E_1, use three plants in its ingredients: plant P_1, P_3, and P_4. Plant P_1 has two pharmacological activities: A_2 and A_4. Plant P_2 also has two pharmacological activities: A_1 and A_2, while plant P_3 has three activities: A_3, A_4, and A_K. The other connections can be described similarly.

From the concept of integrated platform of knowledge of medicinal plants and plant and human-omics depicted in Fig. 1, the efficacy layer in Fig. 4 represents the physiological activity layer in human-omics attribute, the herbal medicine and plant layer represent the prescription and medicinal herb layer, respectively, in knowledge of medicinal plants attribute, while the pharmacological activity layer represents the metabolomics layer in plant-omics attribute. On the following section we will illustrate the data mining techniques on herbal medicine database analyzing relationship among entities for two, and more than two attributes.

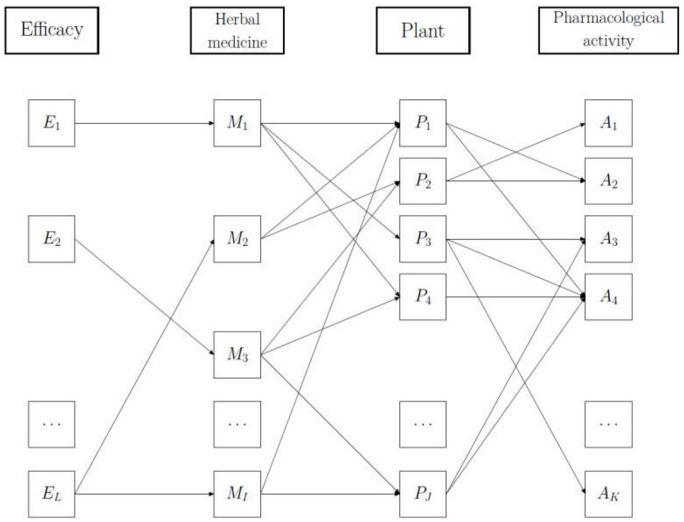

Figure 4. A typical network illustrating connections between efficacy, herbal medicine, plant, and pharmacological activity of plant.

4.1 Two attributes

As an illustration for data mining of herbal medicine database which rely on relationship between two attributes, the relationship between the efficacy of Jamu and medicinal plants used in Jamu is explored using PCA [143-145]. The efficacies of 3,138 Jamu are classified into one of nine categories, namely: (1) disorders of appetite (DOA), (2) disorders of mood and behavior (DMB), (3) female reproductive organ problems (FML), (4) gastrointestinal disorders (GST), (5) musculoskeletal and connective tissue disorders (MSC), (6) pain/inflammation (PIN), (7) respiratory disease (RSP), (8) urinary related problems (URI), and (9) wounds and skin infections (WND). In total, those 3,138 Jamu use 465 plants in their ingredients. The distribution of Jamu and plant utilized in Jamu for each efficacy is shown in Table 2.

Note that, one plant may be used in many Jamu with varying efficacies. Hence, it is interesting to find out the most significant effects of specific plants by analyzing their usage in Jamu, and considering that the more useful a given plant in having certain effect, the more frequently the plant will be used in Jamu when that effect is desired. Biplot, a multivariate exploration tool, is suitable for this purpose because it provides simultaneous plot of principal component scores and loadings, as representation of observations and variables, respectively [145]. Considering plants as observations and efficacy groups as variables, the relationship between them can be explored using a biplot.

Table 2. Distribution of Jamu and plant utilized in Jamu for each efficacy.

Efficacy	Number of Jamu	Number of plants utilized in Jamu formulas
Urinary-related problems (URI)	72	80
Disorders of appetite (DOA)	249	148
Disorders of mood and behavior (DMB)	22	47
Gastrointestinal disorders (GST)	980	290
Female reproductive organ problems (FML)	398	182
Musculoskeletal and connective tissue disorders (MSC)	840	270
Pain and inflammation (PIN)	311	183
Respiratory diseases (RSP)	107	105
Wounds and skin infection (WND)	159	120

Following the explanation of PCA in previous section, the data matrix **A** as an input for PCA is generated by putting plant as observation and efficacy as variables. So, **A** consists of 465 rows and 9 columns. Each cell a_{ij} shows the number of Jamu that use plant i and useful for efficacy j.

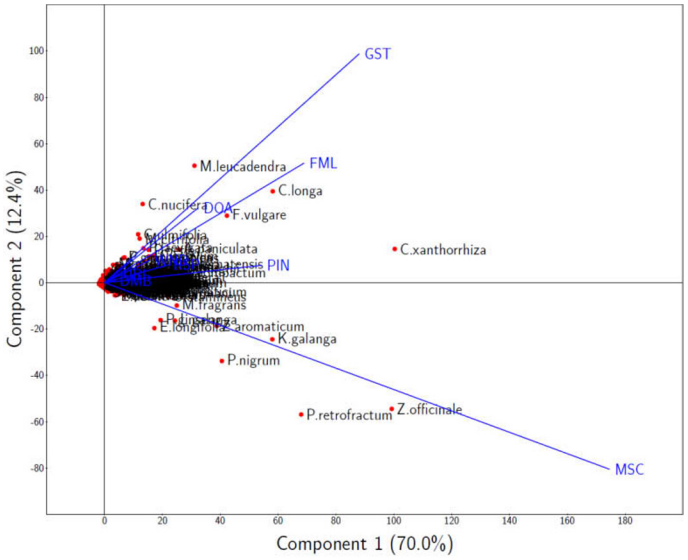

Figure 5. Biplot configuration based on PCA analysis of Jamu data. Plants and Jamu efficacies are represented as red points and blue lines, respectively.

Biplot configuration using the first two components is shown in Fig. 5. In the figure, plants are represented as red points while Jamu efficacies as blue lines, i.e. vectors based on loadings. The length of a given efficacy line showing the variability of plant usage for the corresponding efficacy, that is, the longer the efficacy line the larger the variability of plant usage for that efficacy. From Fig. 6, it is obvious that efficacy MSC has the largest variability of plant usage, followed by efficacy GST and FML. On the other hand, efficacy DMB has the smallest variability of plant usage, followed by efficacy URI and RSP. This finding can be addressed due to two factors, that is, the number of Jamu as well as the number of plant utilized in the corresponding efficacy (see Table 2). Efficacies with large variability of plants usage (MSC, GST, and FML) have large values for both factors; in contrast, efficacies with small variability of plants usage (efficacy DMB, URI, and RSP) have small values for both factors.

In the configurations, many plants are clustered in the center. Note that, the projection value of plants' point on a given efficacy line is the prediction of the frequency of plants usage on that efficacy. So, these clustered plants are basically plants whose frequencies of usage in Jamu are very low. In contrast to the clustered plants, some plants are spread out and located near the efficacy for which the plants are highly utilized. For example, Ginger (*Zingiber officinale*) is located near the efficacy MSC. Ginger is well known for its function of refreshing body, and for this reason many Jamu use Ginger for efficacy MSC which can easily be identified from biplot configuration. Another example is Turmeric (*Curcuma longa*) which located near the efficacy FML. Due to its analgesic and antimicrobial activity, this plant is well known and highly utilized in Indonesia as ingredient of Jamu formula for women during menstruation, which is a problem that classified into efficacy FML. Thus, the biplot configuration exhibits useful information in exploring the relationship between plants and the efficacy of Jamu.

Another illustration for relationship between two attributes on data mining of herbal medicine database is the modeling of Jamu ingredients (representation of knowledge of medicinal plants) to predict the efficacy (representation of human omics). This analysis is performed because of the fact that Jamu is prepared from a mixture of several plants. The plants are chosen so that the Jamu has the desired efficacy. As a result, the composition of the plants used in Jamu formula determines the efficacy. Thus, it is interesting to model the ingredients of Jamu, i.e. the constituent plants, and use this model to predict efficacy. PLS-DA, a statistical model for classification and discrimination based on Partial Least Square Regression (PLSR), is suitable for this analysis because a large number of plants are used in Jamu, whereas Jamu efficacies can be grouped into a few categories or classes. In this method, the plants used in each Jamu medicine served as the predictors, whereas the efficacy of each Jamu provided the responses.

The data structure used for PLS-DA is as follows. The data matrix **X** in *X*-block contains plant usage status. The dimension of matrix **X** is (I x J), where I is the number of Jamu (in this case, 3,138), and J is the number of plants (in this case, 465). Because of the availability of information about Jamu products, which generally do not state in detail the mixing ratio of the plants used, the predictors **X** is constructed only in binary data. Each cell x_{ij} (i = 1, 2, …, I; j = 1, 2, …, J) is set to 1 if Jamu i uses plant j, and is set to 0 otherwise. In the present study, nine indicator variables, which correspond to the 9 efficacies listed in Table 2 perform as the *Y*-block in PLS-DA modeling. Thus, the dimension of data matrix **Y** is (I x 9). Each cell y_{il} (l = 1, 2, …, 9) is set to 1 if Jamu i is classified into efficacy group l, and is set to 0 otherwise. Note that $\sum_{l=1}^{9} y_{il} = 1$ because each Jamu is classified to one efficacy only.

Using the derived PLS-DA model, we can then use it to predict the efficacy of Jamu given information of the ingredients. In this analysis, among the 3,138 Jamu medicines, the efficacies of 2,248 Jamu medicines (71.6%) can be assigned to an individual efficacy reported. Hence, the efficacy in most Jamu medicines can be predicted on the basis of medicinal plants used. The percentages of correct prediction for each efficacy (see Table 3) vary from 22.7% for efficacy DMB to 89.8% for efficacy GST. The low percentage of correct prediction for efficacy DMB can be addressed due to the small number of Jamu for this efficacy, which is only 22 out of 3,138 Jamu (see Table 2).

Table 3. Confusion matrix of the prediction of Jamu efficacy using the PLS-DA model.

Observed efficacy	Predicted efficacy									Total	% Correct
	URI	DOA	DMB	GST	FML	MSC	PIN	RSP	WND		
URI	39	0	0	21	2	10	0	0	0	72	54.2
DOA	0	164	0	29	36	18	0	0	2	249	65.9
DMB	0	1	5	10	0	3	1	2	0	22	22.7
GST	3	17	0	880	12	46	9	6	7	980	89.8
FML	0	13	0	61	266	50	5	1	2	398	66.8
MSC	6	6	1	127	41	638	16	0	5	840	76
PIN	1	0	0	90	4	77	133	4	2	311	42.8
RSP	3	0	0	21	4	23	3	52	1	107	48.6
WND	2	3	0	57	11	11	4	0	71	159	44.7
Total	54	204	6	1296	376	876	171	65	90	3138	71.6

Furthermore, plants in the ingredients of Jamu are used as main ingredients, which contribute primarily to the medicines' efficacies; other plants are used as supporting ingredients [146, 147]. Investigating which plants are main ingredients and which are supporting is important in order to comprehensively understand the mechanisms by which specific plants achieve desired efficacies. The regression coefficients of previous PLS-DA model, which relates plants usage in Jamu as predictors and Jamu efficacy as response, can be helpful in this attempt because they summarize the effect of plant on efficacy. Plants that act as main ingredients will have significant effect on the model developed. Furthermore, due to the absence of parametric testing for the PLS-DA coefficients, the evaluation for significance is performed using permutation testing, in which the distribution of coefficients under the null hypothesis is generated via resampling of the existing data [149].

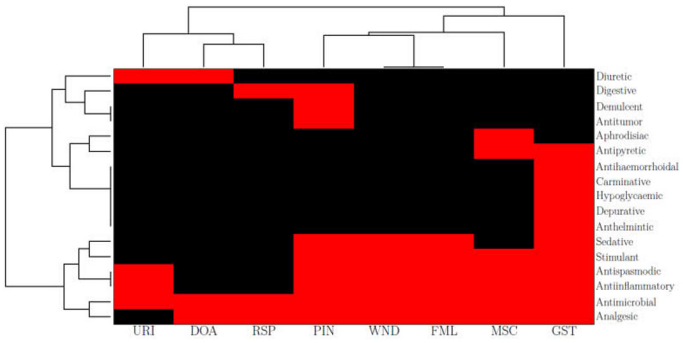

Figure 6. Clustergram of pharmacological activity against Jamu efficacy. The red and black cells indicate that the pharmacological activity is significant or non-significant, respectively, to the corresponding efficacy.

The resampling is performed by permuting the order of the responses (in this case, Jamu efficacies) while maintaining the order of the predictors (in this case, plant utilization as Jamu ingredients) so that the existing relationship between the predictors and the response is destroyed and a new data set is generated under the null hypothesis, i.e., plant utilization in Jamu does not affect Jamu efficacy. If we perform such resampling many times and apply the PLS-DA model on the new data generated from the resampling, the accumulation of the PLS-DA coefficients obtained from this process generates a distribution, against which a *p*-value can be calculated and subsequently evaluated for significance [150].

The results of the significance testing of all plants used in each 9 efficacies are shown in Table 4. Note that one plant may be used for more than one efficacy. From the testing, we observed 234 plants (50.3% among all 465 plants) showing no significant status for all 9 efficacies; whereas the other 231 plants have significant status which comprise of 189 plants (40.6%) are significant only for 1 efficacy, 38 plants (8.2%) are significant for 2 efficacies, and the other 4 plants (0.9%) are significant for 3 efficacies. Besides testing the plants usage statistically, furthermore, we also checked from scientific papers the usage of significant plants in their corresponding efficacy. Many of the results we obtained by our analysis are supported by scientific papers.

Note that in predicting Jamu efficacy based on the information of its ingredients we can also use other methods such as discrimination analysis, nominal logistic regression, and support vector machine. However, in the present study we focus on PLS-DA in classifying Jamu efficacy by taking into consideration that we also intend to evaluate the significance of plant usage in Jamu to achieve specific efficacy as well as extending the analysis into three-way model by adding the plant pharmacological activity into predictors' block.

Table 4. Number of significant plants for each efficacy.

Efficacy	Total	Support from scientific paper	
URI	20	15	-75.00%
DOA	21	20	-95.20%
DMB	12	6	-50.00%
GST	26	23	-88.50%
FML	40	30	-75.00%
MSC	40	39	-97.50%
PIN	39	37	-94.90%
RSP	36	33	-91.70%
WND	43	38	-88.40%

4.2 More than two attributes

During the modeling process of PLS-DA in the previous section, the ingredients of Jamu provide the predictor while the Jamu efficacy serves as the response. In order to identify the function of the plants in Jamu to achieve specific efficacy, the reported pharmacological activities of the plants are added to the predictors block. Thus, the predictors block can be represented as a three-dimensional array $\underline{\mathbf{X}}$ (I x J x K) indexed by Jamu medicine (i), plant (j), and pharmacological activity (k) as depicted in Fig. 2 with Jamu medicine, plant, and pharmacological activity serve as observation, type I and type II variables, respectively. Furthermore, the response block is represented as matrix $\mathbf{Y}$ (I x 9). This analysis then connects three attributes: (1) knowledge of medicinal plants (represented by Jamu and plants corresponding to JAMU DB in Fig 1); (2) plant omics (represented by pharmacological activity corresponding to Biological activity (Nat) in Fig 1); and (3) human omics (represented by efficacy).

The detail about the elements of array $\underline{\mathbf{X}}$ and matrix $\mathbf{Y}$ is as the following. Let x_{ijk} (k = 1, 2, …, K; K = 46 where K is the number of reported pharmacological activity; see previous section on definition of i, j, I, and J) denotes the usage status of plant j with pharmacological activity k in Jamu i, where x_{ijk} = 1 if the plant j with pharmacological activity k is used in Jamu i, and x_{ijk} = 0 otherwise. On the other hand, let y_{il} represents the status of Jamu i on efficacy l, where y_{il} = 1 if Jamu i is classified into efficacy l, and y_{il} = 0 otherwise.

In order to identify the pharmacological activity that is significantly related with the efficacy, we adopt the guidelines from Hair et al. [150] that all weights $\mathbf{w}^K$ (in absolute values) of 0.3 or above are significant for sample sizes of 350 or greater. Figure 6 depicts the 2-dimensional dendrogram of Jamu efficacy and the pharmacological activity significantly related with the efficacy. The cluster of Jamu efficacy and the pharmacological activity was performed using Ward Linkage based on the Euclidean distance among the entities. The clustering of the pharmacological activity side clearly exhibits two groups. The first group consists of activities useful for one or two efficacies only. This group can be regarded as a group of specific activity because the effects of the activities are specific for certain efficacy. For example the diuretic activity is useful for efficacy URI and DOA. Diuretic is an agent that increases the secretion and elimination of urine from the body [151]. Obviously, this activity is beneficial for the efficacy URI. Diuretic also help the body eliminate waste and support the whole process of inner cleansing, which is an action that is useful for efficacy DOA especially related with a slimming purpose. The five activities (antihaemorrhoidal, carminative, hypoglycaemic, depurative, and anthelmintic) are specifically related with efficacy GST. Antihaemorrhoidal means an activity that treats haemorrhoids (piles), while the carminative is defined as an activity that eases discomfort caused by flatulence. Hypoglycaemic activity helps reduce the levels of sugar in the blood, whereas the depurative eliminates toxins and purifies the system especially the blood, and the anthelmintic helpful in expelling parasites from the gut. Thus, all of these activities are helpful for the problem related with the digestive system, i.e. the efficacy GST.

Furthermore, the second group of activity revealed by the dendrogram consists of activities useful for at least four efficacies. In contrast to the first group, this group can be regarded as the general activities because of the diverse efficacies related to this group. Among all activities clustered to this group, antimicrobial activity is significantly related with all 8 efficacies. We can interpret this result as follows. Due to the environmental conditions, hygiene, and its location as a tropical country which led to many microbes that are harmful to health, then it is reasonable that antimicrobial activity is important and should be available in many Jamu formulas in Indonesia. It should be noted that many popular medicinal plants in Indonesia such as Temulawak (*Curcuma xanthorriza*), Ginger (*Zingiber officinale*), Turmeric (*Curcuma longa*) or Kencur (*Kaempferia galanga*) have content of this activity [152].

Anti-inflammation, antispasmodic, analgesic, sedative, and stimulant are also clustered into this general activity group. Since many health problems or diseases are often accompanied with inflammation or spasm, then the plants with anti-inflammation and/or antispasmodic activity are chosen in many Jamu formulas. Those health problems/diseases often cause pain or other discomforts, thus plants with certain activities such as analgesic or sedative effects are chosen in many Jamu medicines. Finally, stimulant activity, which excites or quickens activity of the physiological processes, is important for the recovery reason after one experiencing those health problems or diseases.

From the previous explanation regarding the grouping of pharmacological activity, it can be concluded that in formulating Jamu the plants are selected so that, beside curing the targeted diseases or health problems as indicated by the specific activities, the plants also should overcome the other discomforts caused by the targeted diseases or health problems as indicated by the general activities. It is in accordance with the process of making the Jamu medicines that involving whole part of plant and not only the specific active components. Hence specific or general pharmacological activities of components are involved during the curing process of Jamu medicines towards targeted diseases or health problems.

5. Concluding Remarks

Biology, like most scientific disciplines, is in an era of accelerated information gathering and scientists increasingly depend on the availability of amounts of data such as nucleotide and protein sequences, protein and gene expression, dynamics of metabolites etc. The nature of current systematic understanding of big data biology towards health, nutrition, and other societal issues have recently become the focus of scholar in societal studies of science and information studies. The rise of community databases, i.e., KNApSAcK family DB introduced in the present review, has been strongly associated with the current emphasis on data-intensive science. The central question is whether scientists can deduce how systems and whole organisms work from this torrent of molecular data. To progress this situation, data-intensive approach is needed for understanding intra- and inter-relations in individual layers represented in Fig. 1. The former can be solved based on a type of multivariate analyses such as cluster analysis and principal component analysis. Though the latter is more complicated, several approaches including PLS and N-PLS make it possible to clarify and understand those relations. The big data biology has become an inevitable part of biology, and the laws of nature could be clarified based on global analysis of big data biology the era of which has appeared. For centuries biological research mainly depended on experiments and for a decade or two computational analysis has usually followed experimentation but future it might be the opposite i.e., computational analysis is done first to guide the experimental design facilitated by versatile and freely available omics data at various databases.

Acknowledgements

This work was partially supported by the National Bioscience Database Center in Japan, the Ministry of Education, Culture, Sports, Science and Technology of Japan (Grant-in-Aid for Scientific Research on Innovation Areas "Biosynthetic Machinery. Deciphering and Regulating the System for Creating Structural Diversity of Bioactive Metabolites (2007)") and CREST.

Citation
Afendi FM, Ono N, Nakamura Y, Nakamura K, Darusman LK, Kibinge N, Morita AH, Tanaka K, Horai H, Altaf-Ul-Amin M, Kanaya S (2013) Data Mining Methods for Omics and Knowledge of Crude Medicinal Plants toward Big Data Biology. Computational and Structural Biotechnology Journal. 4 (5): e201301010. doi: http://dx.doi.org/10.5936/csbj.201301010

References

1. Raddick MJ & Szalay AS (2010) The universe online, Sci. 329: 1028-1029
2. Callebaut W (2012) Scientific perspectivism: A philosopher of science's response to the challenge of big data biology, Studies History Philosophy Biol. Biomed. Sci. 43: 69-80
3. Aronova E, Baker KS & Oreskes N (2010) Big science and big data in biology: From the international geophysical year through the international biological program to the long term ecological research (LTER) network, 1957-present, Historical Studies Natural Sci., 40: 183-224
4. Liu CH, Wu, DY & Pollock, JD (2012) Bioinformatic challenges of big data in non-coding RNA research, General Commentary 3: 1-3
5. Thessen AE & patterson DJ (2011) Data issues in the life sciences, Zookeys 150: 15-51
6. Pennisi E (2005) How will big pictures emerge from a sea of biological data, Sci. 309: 94
7. Ranganathan S, Schonbach C, Kelso J, Rost B, Nathan S & Tan TW (2011) Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference, BMC Bioinf., 12: 51.1-4
8. Birney E (2012) Lessons for big-data project, Nature 489: 49
9. Service RF (2012) Materials scientists look to a data-intensive future, Sci., 335: 1434-1435
10. Michener WK & Jones MB (2012) Ecoinformatics: supporting ecology as a data-intensive science, Trends Ecol. Evol. 27: 85-93
11. Hochachka WM, Fink D, Hutchinson RA, Sheldon D, Wong WK & Kelling S (2011) Data-intensive science applied to broad-scale citizen science, Trend Ecol. Evol. 27: 130-137
12. Schadt EE (2012) The changing privacy landscape in the era of big data, Mol. Sys. Biol. 8: 1-3
13. Verpoorte R, Kim HK & Choi YH (2006) Plants as source of medicines, Medicinal and Aromatic Plants Chapter 19, Edited by Boger RJ, Craker LE & Lange D
14. Afendi FM, Okada T, Yamazaki M, Hirai-Morita, A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Amin MAF, Daruman LK, Saito K & Kanaya S (2011) KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research, Plant Cell Physiol. 53:e1.1-12
15. Afendi FM, Katsuragi T, Kato A, Nishihara N, Nakamura K, Nakamura Y, Tanaka K, Hirai-Morita A, Amin MAU, Takahashi H & Kanaya S (2012) Systems biology approaches and metabolomics for understanding Japanese traditional Kampo medicine, Curr Pharm Personalized Med. 10: 111-124
16. Okada T, Afendi FM, Amin MAU, Takahashi H, Nakamura K & Kanaya S (2010) Metabolomics of medicinal plants: The importance of multivariate analysis of Analytical chemistry data, Curr Computer-Aided Drug Design 6: 179-196
17. Ash C, Kiberstis P, Marshall E & Travis J (2012) It takes more than an apple a day, Sci. 337: 1467
18. Binoa RJ, Hall RD,Fiehn O, Kopka J,Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale M, Trethewey RN, Lange BM, Wurtele ES & Sumner LW (2004) Potential of metabolomics as a functional genomics tool, Trends Plant Sci. 9: 418-425
19. Macel M, van Dam NM & Keurentjes JB (2010) Metabolomics: the chemistry between ecology and genetics, Mol. Ecol. Resources 10: 583-593
20. Saito K & Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology, Annu. Rev. Plant. Biol., 61: 463-489
21. Verpoorte R, Choi YH & Kim HK (2005) Ethnopharmacology and systems biology: a perfect holistic match, J. Ethnopharmacol. 100: 53-56
22. Shyur LF & Yang NS (2008) Metabolomics for phytomedicine research and drug development, Curr. Opin. Chem. Biol. 12: 66-71
23. Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, Arita M, Saito K, Ohta D, Shibata D & Kanaya S (2006) KNApSAcK: a comprehensive species-metabolite relationship database. In: Saito K, Dixon RA, Willmitzer L, editors. Biotechnology in agriculture and forestry 57. Plant metabolomics. Berlin: Springer. pp. 165-181.
24. Kikuchi K. & Kakeya H (2006) A bridge between chemistry and biology, Nature Chem. Biol. 2: 392-394
25. Oikawa A, Nakamura Y, Ogura T, Kimura A, Suzuki H, Sakurai N, Shinbo Y, Shibata D, Kanaya S & Ohta D (2006) Clarification of pathway-specific inhibition by Fourier transform ion cyclotron resonance/mass spectrometry-based metabolic phenotyping studies, Plant Physiol. 142: 398-413
26. Shinbo Y, Sakaguchi S, Nakamura Y, Altaf-Ul-Amin M, Kurokawa K, Funatsu K & Kanaya S (2006) Species-metabolite Database (KNApSAcK): Elucidating Diversity of Flavonoids, Comput. Aided Chem. 7: 94-101
27. Tomiki T, Saito T, Ueki M, Konno M, Asaoka T, Suzuki R,Uramoto M, Kakeya H & Osada H (2006) RIKEN Natural Products Encyclopedia (RIKEN NPEdia), a Chemical Database of RIKEN Natural Products Depository (RIKEN NPDepo), Comput. Aided Chem. 7: 157-162
28. Gaida A. & Neumann SJ (2007) MetHouse: Raw and Preprocessed Mass Spectrometry Data, J. Integrative Bioinformatics 4:1-8
29. Hummel J, Selbig J, Walther D & Kopka J (2007) The Golm metabolome database: a database for GC-MS based metabolite profiling. Topics Curr Genet. 18: 75-95
30. Ohta D, Shibata D & Kanaya S (2007) Metabolic profiling using Fourier-transform ion-cyclotron-resonance mass spectrometry, Anal. Bioanal. Chem. 389: 1469-1475
31. Moco, S, Vervoort J, de Vos, CHR & Bino RJ. (2007) Metabolomics technologies and metabolite identification. Trends Anal. Chem. 26: 855-866
32. Saito K, Hirai MY and Yonekura-Sakakibara K. (2008) Decoding genes with coexpression networks and metabolomics - 'majority report by precogs', Trends Plant Sci. 13: 36-43

33. Yonekura-Sakakibara K, Tohge T, Niida R & Saito K (2007) Identification of a Flavonol 7-O-Rhamnosyltransferase Gene Determining Flavonoid Pattern in Arabidopsis by Transcriptome Coexpression Analysis and Reverse Genetics, J. Biol. Chem. 282: 14932-14941
34. Want EJ, Nordstrom A, Morita H & Siuzdak G (2007) From Exogenous to Endogenous: The Inevitable Imprint of Mass Spectrometry in Metabolomics, J. Proteome Res. 6: 459-468
35. Nakamura Y, Kimura A, Saga H, Oikawa A, Shinbo Y, Kai K, Sakurai N, Suzuki H, Kitayama M, Shibata D, Kanaya S and Ohta D (2007) Differential metabolomics unraveling light/dark regulation of metabolic activities in Arabidopsis cell culture, Planta 227: 57-66
36. Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T, Shinozaki K, Hirai MY., Sakurai T, Kikuchi J. & Saito K (2008) PRIMe: a Web site that assembles tools for metabolomics and transcriptomics, In Silico Biol. 8: 339-345
37. Bottcher C, von Roepenack-Lahaye E, Schmidt J, Schmotz C, Neumann S, Scheel D & Clemens S (2008) Metabolome Analysis of Biosynthetic Mutants Reveals a Diversity of Metabolic Changes and Allows Identification of a Large Number of New Compounds in Arabidopsis, Plant Physiol. 147: 2107-2120
38. Dunn WB (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes, Phys. Biol. 5: 1-24
39. Fardet A, Llorach R, Orsoni A, Martin JF, Pujos-Guyot E, Lapierre, C & Scalbert A (2008) Metabolomics provide new insights on the metabolism of dietary phytochemicals in rats, J. Nutr. 138: 1282-1287
40. Fait A, Hanhineva K, Beleggia R, Dai N, Rogachev I, Nikiforova VJ, Fernie AR & Aharoni A (2008) Reconfiguration of the achene and receptacle metabolic networks during strawberry fruit development, Plant Physiol. 148: 730-750
41. Giavalisco P, Hummel J, Lisec J, Inostroza AC, Catchpole G & Willmitzer L (2008) High-resolution direct infusion-based mass spectrometry in combination with whole 13C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas, Anal. Chem. 80: 9417-9425
42. Hagel JM & Facchini P (2008) Plant metabolomics: analytical platforms and integration with functional genomics, Phytochem. Rev. 7: 479-497
43. Hanhineva K, Rogachev I, Kokko H, Mintz-Oron S, Venger I, Karenlampi S & Aharoni A (2008) Non-targeted analysis of spatial metabolite composition in strawberry (Fragaria x ananassa) flowers, Phytochemistry 69: 2469-2481,
44. Hanhineva K (2008) Metabolic Engineering of Phenolic Biosynthesis Pathway and Metabolite Profiling of Strawberry (Fragaria x ananassa), Doctoral dissertation, Univ. of Kuopio
45. Iijima Y, Nakamura Y, Ogata Y, Tanaka K, Sakurai N, Suda K, Suzuki T, Suzuki H, Okazaki K, Kitayama M, Kanaya S, Aoki K & Shibata D (2008) Metabolite annotations based on the integration of mass spectral information, Plant J 54: 949-962
46. Malitsky S, Blum E, Less H, Venger I, Elbaz M & Morin S, Eshed Y & Aharoni A (2008), Plant Physiol. The transcript and metabolite networks affected by the two clades of Arabidopsis glucosinolate biosynthesis regulator, 148: 2021-2049,
47. Matsuda F, Yonekura-Sakakibara K, Niida R, Kuromori T, Shinozaki K & Saito K (2008) MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites, Plant J 57: 555-577
48. Mintz-Oron S, Mandel T, Rogachev I, Feldberg L, Lotan O, Yativ M, Wang Z, Jetter R, Venger I, Adato A & Aharoni A (2008) Gene expression and metabolism in tomato fruit surface tissues, Plant Physiol. 147: 823-851
49. Oikawa A, Matsuda F, Kusano M, Okazaki Y & Saito K (2008) Rice metabolomics, Rice 1: 63-71
50. Overy DP, Enot DP, Tailliart K, Jenkins H, Parker D, Beckmann & M, Draper J (2008) Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints., Nature Protochols 3: 471-485
51. Takahashi H, Kai K, Shinbo Y, Tanaka K, Ohta D, Oshima T, Altaf-Ul-Amin M, Kurokawa K, Ogasawara N & Kanaya S (2008) Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry, Anal. Bioanal. Chem. 391: 2769-2782
52. Werner E, Heilier JF, Ducruix C, Ezan E, Junot C & Tabet JC (2008) Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends., J. Chromatogr. B. 871: 143-163
53. Ara T, Sakurai N, Tange Y, Morishita Y, Suzuki H, Aoki K, Saito K & Shibata D (2009) Improvement of the quantitative differential metabolome pipeline for gas chromatography-mass spectrometry data by automated reliable peak selection, Plant Biotechnol. 26: 445-449
54. Arita M & Suwa K, (2009) Search extension transforms Wiki into a relational system: a case for flavonoid metabolite database, BMC BioData Mining 1: 7.1-8
55. Bando K, Kunimatsu T, Sakai J, Kimura J, Funabashi H, Seki T, Bamba T & Fukusaki E (2011) GC-MS-based metabolomics reveals mechanism of action for hydrazine induced hepatotoxicity in rats, Appl. Toxicol. 31: 524-535
56. Davey MP, Woodward FI & Quick WP (2009) Intraspecfic variation in cold-temperature metabolic phenotypes of Arabidopsis lyrata ssp petraea, Metabolomics 5: 138-149
57. Draper J, Enot DP, Parker D, Beckmann M, Snowdon S, Lin W & Zubair H (2009) Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules', BMC Bioinformatics 10: 277.1-16
58. Fukushima A, Kusano M, Redestig H, Arita M & Saito K (2009) Integrated omics approaches in plant systems biology, Curr. Opin. Chem. Biol. 13: 532-538
59. Han J, Datla R, Chan S & Borchers CH (2009) Mass spectrometry-based technologies for high-throughput metabolomics, Bioanalysis 1: 1665-1684
60. Hounsome N, Hounsome B, Tomos D & Edward-Jones G (2009) Changes in antioxidant compounds in white cabbage during winter storage, Postharvest Biol. Tech. 52: 173-179
61. Kind T, Scholz M &Fiehn O (2009) How large is the metabolome? A critical analysis of data exchange practices in chemistry, PLoS One 4: e5440.1-10
62. Kai K, Hashidzume H, Yoshimura K, Suzuki H, Sakurai N, Shibata D & Ohta D (2009) P450-dependent fatty acid hydroxylation reactions in Arabidopsis, Plant Biotechnol. 26: 175-182
63. Manach C, Hubert J, Llorach R & Scalbert A (2009) The complex links between dietary phytochemicals and human health deciphered by metabolomics, Mol. Nutr. Food Res. 53: 1303-1315
64. Matsuda F, Shinbo Y, Oikawa A, Hirai MY, Fiehn O, Kanaya S & Saito K (2009) Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches, PLoS One 4: e7490.1-10

65. Matsuda F, Redestig H, Sawada Y, Shinbo Y, Hirai MY, Kanaya S & Saito K (2009) Visualization of metabolite identifier information, Plant Biotechnol. 26: 479-483
66. Oishi R, Tanaka K, Hashimoto T, Shinbo Y, Jumtee K, Bamba T, Fukusaki E, Suzuki H, Shibata D, Takahashi H, Asahi H, Kurokawa K, Nakamura Y, Hirai A, Nakamura K, Altaf-Ul-Amin M & Kanaya S (2009) An approach to peak detection in GC-MS chromatograms and application of KNApSAcK database in prediction of candidate metabolites, Plant Biotechnol. 26: 167-174
67. Okada T, Nakamura Y, Kanaya S, Takano A, Malla KJ, Nakane T, Kitayama M & Sekita S (2009) Metabolome analysis of Ephedra plants with different contents of ephedrine alkaloids by using UPLC-Q-TOF-MS, Planta Med., 75: 1356-1352
68. Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, Sakurai T, Saito K & Hirai MY (2009) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants., Plant Cell Physiol. 50: 37-47
69. Shroff R, Rulisek L, Doubsky J & Svatos A (2009) Acid-base-driven matrix-assisted mass spectrometry for targeted metabolomics, Proc. Natl. Acad. Sci. USA 106: 10092-10096
70. Stracke R, De Vos RC, Bartelniewoehner L, Ishihara H, Sagasser M, Martens S & Weisshaar B (2009) Metabolomic and genetic analyses of flavonol synthesis in Arabidopsis thaliana support the in vivo involvement of leucoanthocyanidin dioxygenase, Planta 229: 427-445
71. Takemoto K & Arita M (2009) Heterogeneous distribution of metabolites across plant species, Physica A 388: 2771-2780
72. Tanaka K., Nakamura K., Saito T., Osada H, Hirai A, Takahashi H, Kanaya S &, Altaf-Ul-Amin M (2009) Metabolic pathway prediction based on inclusive relation between cyclic substructures, Plant Biotechnol. 26: 459-468
73. Tanaka K, Ina A & Ohta Y (2009) Comparative study of chemical constituents of the traditional medicine hochuekkito by LC-MS with multivariate statistical analysis, J. Trad. Med. 26: 179-186
74. Tianniam S, Bamba T & Fukusaki E (2009) Non-targeted metabolite fingerprinting of Oriental folk medicine Angelica acutiloba roots by ultra performance liquid chromatography time-of-flight mass spectrometry, J. Sep. Sci. 32: 2233-2244
75. Tohge T & Fernie A (2009) Web-based resources for mass-spectrometry-based metabolomics: a user's guide., Phytochemistry 70: 450-456
76. Wishart DS, (2009) Computational strategies for metabolite identification in metabolomics, Bioanalysis 1: 1579-1596
77. Xie Z, Ma X & Gang DR. (2009) Modules of co-regulated metabolites in turmeric (Curcuma longa) rhizome suggest the existence of biosynthetic modules in plant specialized metabolism, J. Exp. Botany 60: 87-97
78. Yonekura-Sakakibara K & Saito K (2009) Functional genomics for plant natural product biosynthesis, Nat. Prod. Rep. 26: 1466-1487
79. Aliferis KA, & Jabaji S (2010) Metabolite composition and bioactivity of Rhizoctonia solani sclerotial exudates, J. Agric. Food Chem., 58: 7604-7615
80. Bollina V, Kumaraswamy GK, Kushalappa AC, Choo TM, Dion Y, Rioux S, Faubert D & Hamzehzarghani H (2010) Mass spectrometry based metabolomics application to identify quantitative resistance related metabolites in barley against Fusarium head blight, Mol. Plant Pathol. 11: 769-782
81. Bar-Akiva A, Ovadia R, Rogachev I, Bar-Or C, Bar E, Freiman Z, Nissim-Levi A, Gollop N, Lewinsohn E, Aharoni A, Weiss D, Koltai H & Oren-Shamir M (2010) Metabolic networking in Brunfelsia calycina petals after flower opening, J. Exp. Botany 61: 1393-1403
82. Hattori M, Tanaka N, Kanehisa M & Goto S (2010) SIMCOMP/SUBCOMP: chemical structure search servers for network analyses, Nucl. Acids Res. 38: W652-656
83. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Yokota Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K & Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrometry 45: 702-714
84. Kind T & Fiehn O (2010) Advances in structure elucidation of small molecules using mass spectrometry, Bioanal. Rev. 2: 23-60
85. Macel M, Van Dam NM. & Keurentjes, JJB (2010) Metabolomics: the chemistry between ecology and genetics, Mol. Ecol. Resources 10: 583-593
86. Matsuda F, Hirai MY, Sasaki E, Akiyama K, Yonekura-Sakakibara K, Provart NJ, Sakurai T, Shimada Y & Saito K (2010) AtMetExpress Development: a phytochemical atlas of Arabidopsis development, Plant Physiol. 152: 566-578
87. Neumann S & Bocker S, (2010) Computational mass spectrometry for metabolomics - a review, Anal. Biol. Chem. 398: 2779-2788
88. Neveu V, Perez-Jimenez J, Vos F, Crespy V, du Chaffaut L, Mennen L, Knox C, E isner R, Cruz J, Wishart D & Scalbert A (2010), Database, (2010) Phenol-Explorer: an online comprehensive database on polyphenol contents in foods, doi:10.1093/database/bap024
89. Ohta D, Kanaya S & Suzuki H (2010) Application of Fourier-transform ion cyclotron resonance mass spectrometry to metabolic profiling and metabolite identification, Curr. Opinion in Biotechnol. 21: 35-44
90. Penn L, Boeing H, Boushey CJ, Dragsted LO, Kaput J & Scalbert A (2010) Welch AA, Mathers JC. Assessment of dietary intake: NuGO symposium report, Genes Nutr., 5: 205-213
91. Redestig H, Kusano M, Fukushima A, Matsuda F, Saito K & Arita M (2010) Consolidating metabolite identifiers to enable contextual and multi-platform metabolomics data analysis, BMC Bioinformatics, 11: 214.1-11
92. Saito K & Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology, Annu. Rev. Plant Biol. 61: 463-89
93. Singla D, Sharma A, Kaur J, Panwar B & Raghava GP (2010) BIAdb: a curated database of benzylisoquinoline alkaloids, BMC Pharmacol. 10: 4.1-8
94. Tanaka K, Ina A, Hayashi K & Komatsu K (2010) Comparison of Glycyrrhizae Radix from various sources using a multivariate statistical approach, J. Trad. Med., 27: 210-216
95. Tohge T & Fernie AR (2010) Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function, Nature Protocols, 5: 1210-1227
96. Weber RJM. & Viant MR (2010) MI-Pack:Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways, Chemometrics Intel. Lab. Sys. 104: 75-82
97. Takemoto K (2010) Global architecture of metabolite distributions across species and its formation mechanisms, BioSystems, 100: 8-13
98. Acharjee A, Kloosterman B & Maliepaard C (2011) Data integration and network reconstruction with omics data using Random Forest regression in potato, Anal. Chim. Acta 705: 8

99. Krueger S, Giavalisco P, Krall L, Steinhauser MC, Bussis D, Usadel B, Flugge UI., Fernie AR, Willmitzer L &teinhauser D (2011) A topological map of the compartmentalized Arabidopsis thaliana leaf metabolome, PLoS One 6: e17806.1-16
100. Aliferis KA & Chrysayi-Tokousbalides M (2011) Metabolomics in pesticide research and development: review and future perspectives, Metabolomics 7: 35-53
101. Kouskoumvekaki I & Panagiotou G (2011) Navigating the Human Metabolome for Biomarker Identification and Design of Pharmaceutical Molecules, J, Biomedicine and Biotechnol (doi:10.1155/2011/525497)
102. Kumaraswamy GK, Bollina V, Kushalappa AC, Choo TM, Dion Y, Rioux S, Mamer O & Faubert D (2011) Metabolomics technology to phenotype resistance in barley against Gibberella zeae, Eur. J. Plant Pathol. 130: 29-43
103. Kusano M, Tabuchi M, Fukushima A, Funayama K, Diaz C, Kobayashi M, Hayashi N, Tsuchiya YN, Takahashi H, Kamata A, Yamaya T & Saito K (2011) Metabolomics data reveal a crucial role of cytosolic glutamine synthetase 1;1 in coordinating metabolic balance in rice, Plant J. 66:456-66
104. Ohkama-Ohtsu N, Sasaki-Sekimoto Y, Oikawa A, Jikumaru Y, Shinoda S, Inoue E, Kamide Y, Yokoyama T, Hirai MY, Shirasu K, Kamiya Y, Oliver DJ & Saito K (2011) 12-oxo-phytodienoic acid-glutathione conjugate is transported into the vacuole in Arabidopsis, Plant Cell Physiol. 52: 205-209
105. Osorio S, Bombarely A, Giavalisco P, Usadel B, Stephens C, Araguez I, Medina-Escobar N, Botella M, Fernie AR & Valpuesta V (2011) Demethylation of oligogalacturonides by FaPE1 in the fruits of the wild strawberry Fragaria vesca triggers metabolic and transcriptional changes associated with defence and development of the fruit, J. Exp. Botany doi:10.1093/jxb-erq465
106. Scalbert A, Andres-Lacueva C, Arita M, Kroon P, Manach C, Urpi-Sarda M & Wishart D (2011) Databases on food phytochemicals and their health-promoting effects, J. Agric. Food Chem. 59: 4331-4348
107. Giavalisco P, Li Y, Matthes A, Eckhardt A, Hubberten HM, Hesse H, Segu S, Hummel J, Kohl K & Willmitzer L (2011) Elemental formula annotation of polar and lipophilic metabolites using 13C, 15N and 34S isotope labelling, in combination with high-resolution mass spectrometry, Plant J. 68: 364-376
108. Fiehn O, Barupal DK & Kind T (2011) Extending biochemical databases by metabolomic surveys, J. Biol. Chem. 286: 23637-23643
109. Kai K, Takahashi H, Saga H, Ogawa T, Kanaya S & Ohta D (2011) Metabolomic characterization of the possible involvement of a Cytochrome P450, CYP81F4, in the biosynthesis of indolic glucosinolate in Arabidopsis, Plant Biotechnol. 28: 379-385
110. Yanuar A, Mun'im A, ApLagho ABA, Syahdi RR, Rahmat M & Suhartanto H (2011) Medicinal plants database and three dimensional structure of the chemical compounds from medicinal plants in Indonesia, Int. J. Computer Sci. Issue 8: 180-183
111. Katoh A, Fukuda S, Fukusaki E, Hashimoto T, Hayasaki T, Kanaya S, Komura H, Nomoto K, Shojo M & Takeno KJ (2011) Systems biology in a commercial quality study of the Japanese Angelica radix: toward an understanding of traditional medicinal plants, Am. J. Chinese Med. 39: 757-777
112. Kaneko Y, Obata Y, Nishino T, Kakeya H, Miyazaki Y, Hayasaka T, Setou M, Furusu A & Kohno S (2011) Imaging mass spectrometry analysis reveals an altered lipid distribution pattern in the tubular areas of hyper-IgA murine kidneys, Exp. Mol. Pathol. 91: 614-621
113. Aliferis KA & Chrysayi-Tokousbalides M (2011) Metabolomics in pesticide research and development: review and future perspectives, Metabolomics 7: 35-53
114. Takahashi H, Morimoto T, Ogasawara N & Kanaya S (2011) AMDORAP: non-targeted metabolic profiling based on high-resolution LC-MS, BMC Bioinformatics 12: 259.1-8
115. Obata T & Fernie AR (2012) The use of metabolomics to dissect plant responses to abiotic stress, Cell. Mol. Life Sci. 69: 3225-3243
116. Sartor MA, Ade A, Wright Z, States D, Omenn GS, Athey B & Karnovsky A (2012) Metab2MeSH: annotating compounds with medical subject headings, Bioinformatics 28: 1408-1410
117. Asano T, Kobayashi K, Kashihara E, Sudo H, Sasaki R, Iijima Y, Aoki K, Shibata D, Saito K, Yamazaki M (2012) Suppression of camptothecin biosynthetic genes results in metabolic modification of secondary products in hairy roots of Ophiorrhiza pumila, Phytochemistry (in press)
118. Ahuja I, Kissen R & Bones AM (2012) Phytoalexins in defense against pathogens, Trends Sci. 17: 73-90
119. Okazaki Y & Saito K (2012) Recent advances of metabolomics in plant biotechnolgy, Plant Biotechnol. Rep. 6: 1-15
120. Marti G, Erb M, Boccard J, Glauser G, Doyen GR,Villard N, Robert CA, Turlings TC, Rudaz S & Wolfender JL (2012) Metabolomics reveals herbivore-induced metabolites of resistance and susceptibility in maize leaves and roots. Plant, Cell & Env. 36: 621-639
121. Liberman LM, Sozzani R & Benfey PN (2012) Integrative systems biology: an attempt to describe a simple weed, Curr. Opin. Plant Biol. 15: 162-167
122. Houshyani B, Kabouw P, Muth D, de Vos RCH, Bino RJ & Bouwmeester HJ (2012) Characterization of the natural variation in Arabidopsis thaiana metabolome by the analysis of metabolic distance, Metabolomics 8: S131-S145
123. Wahyuni Y, Ballester AR, Tikunov Y, de Vos RCH, Pelgrom KTB, Maharijaya A, Sudarmonowati E, Bino RJ & Bovy AG (2012) Metabolomics and molecular marker analysis to explore pepper (Capsicum sp.) biodiversity, Metabolomics doi 10.1007/s11306-0432-6
124. Sano R, Ara R, Akimoto N, Sakurai N, Suzuki H, Fukuzawa Y, Kawamitsu Y, Ueno M & Shibata D (2012) Dynamic metabolic changes during fruit maturation in Jatropha curcas L., Plant Biotechnol. 29: 175-178
125. Ohtani M, Nakano Y, Usami T & Demura T (2012) Comparative metabolome analysis of seed kernels in phorbol ester-containing and phorbol ester-free accessions of Jatropha curcas L., Plant Biotechnol. 29: 171-174
126. Khan SA, Chibon PY, de Vos RC, Schipper BA, Walraven E, Beekwilder J, van Dijk T, Finkers R, Visser RG, van de Weg EW, Bovy A, Cestaro A, Velasco R, Jacobsen E & Schouten HJ. (2012) Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16, J. Exp. Botany, 63: 2895-2908
127. Alla MMN, Khedr A, Serag MM, Abu-Alnaga AZ & Nada RM (2012) Regulation of metabolomics in Atriplex halimus growth under salt and drought stress, Plant Growth Regul. 67: 281-304
128. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, Gonzalez-Beltran A, Sansone SA, Griffin JL & Steinbeck C (2012) MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data, Nucleic Acid Res. doi:101093/nar/gks1004

129. Wagele B, Witting M, Schmitt-Kopplin P & Suhre K (2012) MassTRIX reloaded: combined analysis and visualization of transcriptome and metabolome data, PlosOne 7: e39860
130. Peukert M, Matros A, Lattanzio G, Kaspar S, Abadia J & Mock HP (2012) Spatially resolved analysis of small molecules by matrix-assisted laser desorption/ionization mass spectrometric imaging (MALDI-MSI)., New Phytol. 193: 806-815
131. Ballester AR, Lafuente MT, de Vos RC, Bovy AG & Gonzalez-Candelas L (2012) Citrus phenylpropanoids and defence against pathogens. Part I: Metabolic profiling in elicited fruits, Food Chemistry 136: 178-185
132. Jolliffe IT (2002) Principal Component Analysis 2nd Ed. Springer-Verlag New York
133. Wold S, Sjostrom M & Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58: 109-130
134. Boulesteix A & Strimmer K (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8: 32-44
135. Barker M & Rayens W (2003) Partial least squares for discrimination. J. Chemometrics 17: 166-173
136. Bro R (1996) Multi-way calibration. multi-linear pls. Journal of Chemometrics 10: 47-61
137. Smilde AK, Bro R, & Geladi P (2004) Multi-way analysis with application in the chemical sciences. West Sussex, England: John Wiley & Sons
138. Smilde AK (1997) Comments on multilinear pls. Journal of Chemometrics 11: 367-377
139. Beers SJ, (2001) Jamu: The ancient Indonesian art of healing, Periplus Editions (HK) Ltd. Vermont
140. Pramono S (2007) Jamu in Indonesia daily life and industry, Institute of Natural Medicine University of Toyama
141. Adnyana LK & Soemardji AA (2007) Evaluation of Pharmacological efficacy of Jamu medicine, Institute of Natural Medicine University of Toyama
142. Sangat, HM, Zuhud EAM (2000) Damayanti EK, Komus penyakit dan tumbuhan obat Indonesia [Ethnofitomedika], Yayasan Obat Indonesia, Jakarta
143. Afendi FM, Darusman LK, Hirai A, Altaf-Ul-Amin Md, Takahashi H, Nakamura K & Kanaya S (2010) System Biology Approach for Elucidating the Relationship between Indonesian Herbal Plants and the Efficacy of Jamu. IEEE International Conference on Data Mining Workshops, ICDM 2010 (December 14-17 2010, University of Technology Sydney, Sydney, Australia)
144. Afendi FM, Sulistiyani, Hirai A, Altaf-Ul-Amin Md, Takahashi H, Nakamura K & Kanaya S (2011) Bootstrapping Jamu Dataset to Examine Assignment Consistency of Plants to Jamu Efficacy. The 2nd International Symposium on Temulawak (May 25-27 2011, Bogor Agricultural University, Bogor, Indonesia)
145. Gabriel KR (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika 58: 453-467
146. Adnyana IK & Soemardji AA (2007) Evaluation of pharmacological efficacy of Jamu medicine. Toyama: University of Toyama Institute of Natural Medicine
147. Pramono S (2007) Jamu in Indonesian daily life and industry. Toyama: University of Toyama Institute of Natural Medicine
148. Good PI (2005) Permutation, Parametric and Bootstrap Tests of Hypotheses 3rd edition. New York: Springer
149. Afendi FM, Altaf-Ul-Amin Md & Kanaya S (2011) Permutation test in evaluating the significance of plants in PLS-DA model of Jamu ingredients. The 7th Asian Crop Science Association Conference, ACSAC 2011 (September 27-30 2011, Bogor Agricultural University, Bogor, Indonesia)
150. Hair JF, Black WC, Babin BJ & Anderson RW (2010) Multivariate Data Analysis, 7th Ed. Prentice Hall
151. Hoffmann D (2003) Medical herbalism: the science and practice of herbal medicine. Rochester Vermont: Healing Arts Press
152. Duke JA, Bogenschutz-Godwin MJ, duCellier J & Duke PK (2002) Handbook of medicinal herbs, 2nd ed. CRC Press

Competing Interests:
The authors have declared that no competing interests exist.