Subset-binding: A novel algorithm to detect paired itemsets from heterogeneous data including biological datasets

Yayoi Natsume-Kitatani (natsume@nibiohn.go.jp)
National Institutes of Biomedical Innovation, Health and Nutrition https://orcid.org/0000-0003-1749-3318
Kenji Mizuguchi
National Institutes of Biomedical Innovation, Health and Nutrition https://orcid.org/0000-0003-3021-7078
Naonori Ueda (naonori.ueda@riken.jp)
RIKEN Center for Advanced Intelligence Project

Research Article

Keywords: heterogeneous data integration, biomarker detection, patient stratification, fuzzy association rule mining

Posted Date: April 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-405195/v1

Abstract

The integration of heterogeneous data to infer latent relationships across them and find the factors in the relationship is a challenging task. In this regard, various machine learning techniques have provided novel insights through data integration. However, concerns remain regarding their application to biological datasets because the latent consensus information across all views is often limited to partial components that do not have a significant impact on the mutual agreement across views. Advocating the idea of “subset-binding,” which focuses on finding inter-related attributes in heterogeneous data according to their co-occurrence, this study developed a novel algorithm to perform subset-binding by extending fuzzy association rule mining techniques. Our method could detect genes related to liver toxicity caused by acetaminophen in a data-driven manner; the results are consistent with those reported in the literature. This technology paves the way for a wide range of applications, including biomarker detection and patient stratification.

Background

The integration of heterogeneous data to infer latent relations across them and find factors responsible for the relations can be a challenging task. In recent years, technological advances have accelerated attempts to collect heterogeneous data comprehensively, and methods to find factors responsible for events of interest by utilizing such heterogeneous data have been drawing considerable attention. Multi-omics analysis is a good example of such a trend in life sciences, and various analytical methods have been developed to date to extract informative knowledge from the data obtained through different techniques.

Multi-view learning is an important topic in machine learning that deals with multi-view data (such as biological data collected using multiple “omics” technologies, i.e., multi-omics data), as opposed to single-view data (e.g., single-omics data). A variety of clustering-based methods can be used to investigate complementary and/or consensus information across multi-view data [1,2]. Multi-view clustering approaches include co-training [3], multi-kernel learning [4], and multi-view graph clustering [5]; however, they assume that global consensus information exists across the views and all the views must be used for better clustering performance. The co-training approach attempts to maximize the mutual agreement across all the views, whereas multi-kernel learning attempts to combine predefined kernels to improve the clustering performance, and multi-view graph clustering attempts to represent views as graphs, fuse them, and apply graph-cut algorithms for clustering [1].

Crucially, there remains a concern in the application of these approaches to biological datasets because the latent consensus information across all the views is often limited to partial components that do not have a significant impact on the mutual agreement across the views. For example, typically, only a small number of genes are related to specific disease features and such a relationship may not be detected through conventional methods, because the mutual agreement across the gene expression profile and the clinical features will be dominated by the expression patterns of the majority “unrelated” genes. In addition, there is a growing demand to handle multi-view data that include both quantitative and categorical data. One example is the combination of electronic health records (EHR) and omics data from patients; biomolecules whose abundances are related to attributes in EHR data can be of considerable interest in terms of precision medicine. However, the analysis of quantitative data (e.g., omics data) and categorical data (e.g., EHR data) together is a challenging task, and the definition of metrics among views in multi-view clustering is not trivial. Moreover, all the items in a single view are partitioned into clusters, and it may not be possible to easily obtain the relationships between the items across the views. Weighted correlation network analysis (WGCNA) [6] is a popular method in computational biology that can be used for relating attributes to external traits. For example, this method can be used to find highly correlated genes and relate them to physiological properties. However, this method can suffer from the same problem as above because it depends on hierarchical clustering and is designed to relate a group of attributes in one view (e.g., genes in a gene expression profile) to only a single attribute (not multiple attributes) in another view (e.g., measurements in clinical information).

Herein, we propose for the first time a solution based on association rule mining (ARM) that addresses these challenges. ARM is a popular technique in data mining that aims to find co-occurring attributes and rules for the frequency of their appearance in a dataset. ARM has a long history, and various surveys are available on this topic [7,8]. The Apriori algorithm was first introduced by Agrawal et al. [9,10] and is one of the most popular and well-known ARM algorithms that efficiently generates significant association rules across items in a dataset. It may be classified as a candidate generation approach, and extensive studies have been conducted on its improvements or extensions [11]. An important extension of the Apriori algorithm involves the handling of datasets that contain both quantitative and categorical attributes [12]. The initially proposed method in the quantitative ARM converts quantitative attributes into binary ones by applying Boolean logic to handle them similar to categorical attributes. However, one concern regarding this method is the sharp boundary problem, that is, loss of information by the sharp partitioning of the data.

To solve this issue, fuzzy logic was introduced to ARM, and the approach was named fuzzy association rule mining (FARM) [13]. In contrast to conventional ARM, in which quantitative attributes are converted into binary attributes with Boolean logic, FARM converts quantitative attributes into real numbers between 0 and 1, both inclusive, to handle uncertainty. FARM is well summarized in various survey papers [14,15,16]. In this study, we used FARM techniques to address the challenges mentioned earlier in finding paired itemsets across views. Our proposed algorithm

finds attribute responses for relations across two views, which we call *subset-binding*. In the text that follows, we demonstrate its efficacy and interpretability based on experiments conducted with artificial data and real/actual biological data.

Results

We performed experiments with two artificial datasets and one real biological dataset to confirm the performance of our algorithm in detecting paired frequent itemsets for subset-binding. Our algorithm can also be applied to arbitrary paired datasets. For simplicity, this study assumed that the dataset consists of gene expression profile data (data1) and clinical measurement data (data2).

- **Artificial data (small)**

Artificial data were generated, as shown in Fig. 1a. The gene expression profile data had 100 rows (for 100 patients) and 200 columns (for 200 genes), and random values were generated following a standard normal distribution (Supplementary Data 1). Clinical measurement data were generated using the same procedure (Supplementary Data 2). We added some irregular patterns to these two matrices, randomly generated according to the normal distribution, with different mean and standard deviation (SD) values, as frequent itemsets to be detected. We then evaluated the performance of the algorithm by confirming whether the irregular patterns we added were successfully detected. The patterns to detect are summarized in Table 1. With this dataset, we compared five membership functions: min-max scaling, conversion with sigmoid function, rank-based conversion, histogram-based conversion, and z-score-based conversion (details have been provided in the Methods section); Table 2 summarizes the results. With the parameter settings we used, only the histogram-based function and the z-score-based function could generate paired frequent itemsets, and all the three patterns, which we added, were included in the generated rules (Supplementary information, Supplementary Table 1-10). The other three methods (min-max scaling and sigmoid and rank-based conversions) could not detect any of the patterns, when the same parameters were used.

- **Artificial data (large)**

Next, we performed an experiment using a large dataset. Paired matrices with 1000 rows and 2000 columns were generated using the same procedure as in the case of artificial data (small), and we added three patterns to be detected (Supplementary Data 3,4), as shown in Fig. 1b and Table 1. The histogram-based function and z-score-based function successfully detected all the three patterns that were included in the inputs (Table 2, Supplementary information, Supplementary Table 11-14).

- **Real data (Liver toxicity)**

Finally, we performed an experiment using a real biological dataset [17]. Acetaminophen (50, 150, 1500, and 2000 mg/kg body-weight), which is known to cause liver toxicity at high doses, was administered to 64 rats (4 male rats per dose group) during a light period. These rats were sacrificed after 6, 18, 24, or 48 h. Hepatic gene expression profiling on the left liver lobe was performed using Agilent-011868 (G4130) rat oligonucleotide microarrays (Agilent Technologies, Palo Alto, CA), and 3116 genes were selected as being significantly differentially expressed in the acetaminophen-treated group by comparing them with a control group (Fig. 2a, data1). In addition, 48 histopathological observations (qualitative data) and 10 clinical measurements (quantitative data) were obtained from these rats (Fig. 2a, data2). The experimental conditions (dose and time point) were added to data2, the details of which are presented in the Methods section (Supplementary Data 5).

With the parameter settings we tested (min support for data1 and data2: 0.02, min items for data1 and data2: 10, lift: 4.8), 3986 paired-association rules were generated (Supplementary Table 15); the one with the highest conviction is shown in Fig. 2b. The results demonstrated that high values of alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), and total bile acid (TBA) with a low value of cholesterol co-occurred with histopathological observations, as can be seen under the “histopathological observation” column in data2 of Fig. 2b (“LLL_Centrilob_Necrosis,” “LLL_Hepato_Hypertrophy,” “LML_Centrilob_Necrosis,” and “LML_Sinusoid_Congestion”), and were further related to the toxic dose and time point of acetaminophen treatment. These attributes were then paired with 10 probes (“A_43_P10003” (gene name: Hsph1), “A_42_P717602” (gene name: Mat2a), “A_42_P484423” (gene name: Pgs1), “A_43_P17455” (gene name: Dnah9), “A_43_P14864” (gene name: Dynll1), “A_43_P16523” (gene name: Nomo1), “A_43_P19279” (gene name: Lyzl4), “A_43_P12811” (gene name: Srm), “A_42_P655825” (gene name: Smg9), and “A_42_P804499” (n.d.)) as the antecedent. The result was consistent with regard to the fact that high values of ALP, ALT, AST, and TBA and low levels of cholesterol are regarded as markers of liver toxicity, and a high dose of acetaminophen causes liver toxicity.

Discussion

Herein, we present a novel approach (subset-binding approach) to find attributes that are related to each other in paired data. Rather than combining multi-view data to maximize the mutual agreement, this approach focuses on finding attributes of interest according to their co-occurrence. The advantage of this approach is that the statistics of co-occurrence are easily computable, which makes the output easily

interpretable. In addition, this approach can relate heterogeneous data in a data-driven way without prior knowledge and can be used for a variety of objectives such as the exploration of biomarkers, inference of the molecular basis of events, or patient stratification, depending on input data.

This advantage is closely reflected in the experiment using real biological data (Fig. 2b). An overdose of acetaminophen has been reported to cause sinusoidal congestion and centrilobular necrosis [18] owing to the formation of a hepatotoxic intermediate metabolite, and it can even be fatal. In addition, acetaminophen has been reported to cause hepatocyte hypertrophy in rats [19]. Furthermore, some of the genes detected in the antecedent were indicated to be involved in liver injury. Methionine adenosyltransferase (Mat) is responsible for the biosynthesis of S-adenosylmethionine (AdoMet) and exists in two isoforms in mammals, namely Mat1a and Mat2a. Mat2a is induced in response to a liver injury, which in turn accelerates cell division and hepatocyte growth [20]. AdoMet functions as a precursor of antioxidative glutathione (GSH) and polyamines, which are involved in cell growth and apoptosis. GSH depletion by acetaminophen treatment and its influence on hepatic necrosis have also been studied extensively [21]. In addition, spermidine synthase (Srm) plays a key role in polyamine synthesis. Under the conditions of a liver injury, downregulation of Mat1a and upregulation of Mat2a are observed, which resulting in a low level of AdoMet that has a protective effect against liver injury [22]. A database search demonstrated that Hsph1 interacts with Mat2a [23]. Pgs1 is responsible for the biosynthesis of phosphatidylglycerol and cardiolipin, located in the inner mitochondrial membrane, and the reactive oxygen species (ROS)-induced oxidation of these compounds is associated with mitochondrial dysfunction [24]. AdoMet is reported to prevent mitochondrial dysfunction triggered by chronic alcohol treatment [25].

Overall, the results indicate that the dysregulation of methionine metabolism and decrease in AdoMet are associated with hepatic necrosis, mitochondrial dysregulation, and cell growth caused by acetaminophen treatment. These reports are summarized in Fig. 3. The liver toxicity dataset we used for the case study was also used to evaluate and prove the importance of the mixOmics R package [26], which offers a variety of multivariate methods for the analysis of multi-omics data. The authors of mixOmics have published case studies on their well-organized website (http://mixomics.org/) for users and have illustrated independent principal component analysis (IPCA, [27 28]) and sparse Partial Least Squares (sPLS, [29,30]) using the same liver toxicity dataset. They showed that IPCA can be used for separating dose- and treatment time-dependent patterns in the gene expression and sparse IPCA for selecting important genes (http://mixomics.org/case-studies/ipca-liver-toxicity/). In addition, sPLS could show the association between some probes/genes (A_42_P698740, A_42_P681650, A_43_P10006, A_43_P22616, A_43_P23376, A_42_P705413, A_43_10003, A_43_10606, A_43_P17415, and A_42_P620915) and clinical measurements (high levels of BUN, TBA, AST, and ALT and a low level of cholesterol; http://mixomics.org/case-studies/spls-liver-toxicity/). However, the striking advantage of our method over these methods (IPCA and sPLS) is that our method accepts categorical data or a mixture of continuous values and discrete values (one-hot representation for categorical data) as well and provides the frequency of the occurrence of the detected association rules as scores such as *support*. Our results showed that some genes and clinical measurements were also associated with histopathological observations and experimental conditions (dose and treatment time), which clearly demonstrated that the detected genes and their associated attributes were related to liver toxicity caused by acetaminophen.

Owing to the interpretability of the output, our approach is expected to be applied to any paired data in a broad area of interest. Future directions to extend or improve our algorithm include: i) selection of its underlying method, ii) selection of the t-norm in FARM, iii) selection of the membership function in FARM, iv) detection of rare yet interesting patterns, and v) pruning redundant patterns. Although the fuzzy Apriori algorithm is one of the most popular methods in the field of FARM, its space inefficiency is a drawback. Mangalampalli and Pudi previously reported that while numerous studies in the field of FARM have been conducted on the selection of t-norms or implicators [31-33], only a few have been conducted on methods for preparing fuzzy datasets [34]. They had proposed the application of fuzzy *c*-means clustering for this process; however, we tested five membership functions that were simpler than fuzzy *c*-means because biological datasets (such as omics data) tend to have n << p. ARM has been extensively investigated and several other extensions of basic methods have been developed, including rare association rule mining that focuses on rules containing infrequent items [35] and non-redundant association rule mining that suppresses the production of many redundant rules using closed frequent itemsets [36]. These methods can be incorporated into our approach for further improvement of performance.

Methods

Algorithm

ARM aims to search for frequent itemsets (items that tend to co-occur) and find association rules (patterns of co-occurrence between items within a frequent itemset). Our study aimed to develop a method that would enable i) the search of frequent itemsets in a given dataset containing continuous values and ii) finding association rules between frequent itemsets that are detected in different given datasets to associate two heterogeneous datasets focusing on a limited number of items. The workflow of this algorithm is shown in Fig. 4a.

- **Apriori algorithm**

Let with *n* binary attributes be a set of "items" called the "itemset". Let be a set of *m* observations in which each *t* has *I*. When an observation has *k* items (*I* has *k* ones and (*n* - *k*) zeros), the itemset is called a *k*-length itemset. An association rule is defined next, where *X* (called antecedent) and *Y* (called consequent) co-occur, with , and :

$X \rightarrow Y$. (1)

First, the Apriori algorithm evaluates the frequencies of 1-length itemsets (itemsets containing only 1 item), and those with a low frequency that do not satisfy the user-specified minimum support are pruned. Support is a score that represents the frequency of *X*, *Y* or co-occurrence of *X* and *Y*. Given $X \rightarrow Y$, support can be expressed as follows:

$$support(X \rightarrow Y) = (|X \cup Y|)/m. \quad (2)$$

Next, all possible (*k* + 1)-length itemsets are generated from *k*-length itemsets, and the ones containing *k*-length itemsets, whose support is smaller than user-specified minimum support, are pruned. These processes are iterated until convergence is achieved. Using this procedure, frequent itemsets (those with support higher than the user-specified minimum support) are detected.

Thereafter, association rules are generated by searching for an antecedent and a consequent within each frequent itemset. Several scores are commonly used, such as lift, which can be expressed as follows:

lift (X → Y) = *support* (X → Y)/{*support*(X) * *support*(Y)}. (3)

- Fuzzy association rule mining

Conventional ARM approaches assume that input data contain categorical attributes. However, the data that we handle can be quantitative or a mixture of qualitative and quantitative data. Therefore, quantitative attributes are converted into categorical attributes by putting thresholds for quantization (e.g., when threshold1 and threshold2 are given, and threshold1 < threshold2, quantitative attributes can be assigned to one of these categories: i) ≤ threshold1, ii) between threshold1 and threshold2, and iii) ≥ threshold2).

This forms the crisp data; however, the procedure results in loss of information. To solve this problem, fuzzy logic is introduced in the Apriori algorithm.

Fuzzy logic is defined as "a class of objects with a continuum of grades of membership," and quantitative attributes are converted into several categories with "membership values" ranging from 0 to 1 (e.g., quantitative attribute at threshold1 is converted into i) 0.5 category and ii) 0.5 category). The notions of union, intersection, and complement, which are used to calculate several important scores in ARM, such as *support*, can also be extended to fuzzy sets.

- Functions for calculating membership values

When converting quantitative attributes into fuzzy categorical sets, the main problem lies in defining membership functions, which are used to calculate membership values. Because membership values range from 0 to 1, *min-max* scaling, sigmoid transformation, or rank-based conversion is typically used. However, these methods reduce the differences in membership values between the most applicable category and the least applicable category and obtain too-fuzzy data for the Apriori algorithm. With this preliminary observation, we designed novel membership functions, as described next.

Histogram-based conversion

The frequency distribution of the quantitative data was converted into a histogram with a user-specified number of bins (Fig. 4b). The quantitative attributes are converted into three categories: "low," "average," and "high." Their membership values for quantitative attribute *v* can be expressed as given subsequently, where the frequency of the bin that includes the quantitative attribute *v* is F_v, frequency of the highest bin is F_H, lower boundary of the highest bin is b_L, and upper boundary of the highest bin is b_H.

Membership value for category "low" = $1 - F_v/F_H$ if $v < b_L$, (4)

- otherwise; (5)

Membership value for category "high" = $1 - F_v/F_H$ if $v > b_H$, (6)

0 otherwise. (7)

The sum of membership values for categories “low,” “average,” and “high” is supposed to be 1. However, the information in category “average” will not be used for ARM because it will be handled as frequently occurring items and will fail to detect interesting association rules that include category “low” or category “high.”

Z-score-based conversion

The frequency distribution of the quantitative data was converted into a standard normal distribution to obtain the z-scores (Fig. 4c). Because approximately 95% of the data is known to range from -2 to 2, the quantitative attributes are converted into three categories: “low,” “average,” or “high.” Their membership values for quantitative attribute v can be expressed as follows:

Membership value for category “low” (v) = -(z-score(v)/2) if $-1 \leq$ z-score (v)/2 < 0 (8)

1 if -(z-score(v)/2) > 1 (9)

0 otherwise (10)

Membership value for category “high” (v) = z-score(v)/2 if 0 < z-score (v)/2 ≤ 1 (11)

1 if z-score(v)/2 > 1 (12)

0 otherwise (13)

By analogy with the histogram-based conversion, the sum of membership values for categories “low,” “average,” and “high” is supposed to be 1. However, the information in category “average” will not be used for ARM because it will be handled as frequently occurring items and will fail to detect interesting association rules that include category “low” or category “high.”

- **Other membership functions used for comparison**

The formula of min-max scaling for the quantitative attribute v is as follows:

Membership value for category “Low” = $1 - (v - v_{min}) / (v_{max} - v_{min})$, (14)

Membership value for category “High” = $(v - v_{min}) / (v_{max} - v_{min})$. (15)

The formula of sigmoid function for the quantitative attribute v is as follows:

Membership value for category “Low” = $1 - 1/(1 + e^{-v})$, (16)

Membership value for category “High” = $1/(1 + e^{-v})$. (17)

The formula of rank-based conversion for the quantitative attribute v is as follows:

Membership value for category “Low” = $1 -$ r(v)/n, (18)

Membership value for category “High” = r(v)/n, (19)

where r is the rank, and n is the number of observations.

- **Association of heterogeneous datasets**

In the conventional ARM approach, association rules are generated “within” each frequent itemset. Our method aims to generate association rules by which their antecedents are derived from one dataset and consequents from the other, so that the detected association rules represent itemsets derived from different datasets related to each other. With this purpose in mind, we developed a novel algorithm, as described subsequently.

Let $I_1 = \{i_{1,1}, i_{1,2}, \ldots, i_{1,p}\}$ of p attributes and $I_2 = \{i_{2,1}, i_{2,2}, \ldots, i_{2,q}\}$ of q attributes be sets of “items” and I_1 and I_2 be called “itemsets”. Let $T_1 = \{t_{1,1}, t_{1,2}, \ldots, t_{1,m}\}$ and $T_2 = \{t_{2,1}, t_{2,2}, \ldots, t_{2,m}\}$ be sets of m observations in which each of t_1 and t_2 has I_1 and I_2, respectively. T_1 and T_2 are assumed to have the same number of observations, and $t_{1,a}$ and $t_{2,a}$ ($a \in \{1, 2, \ldots, m\}$) are associated with each other (e.g., $t_{1,a}$: medical record of patient ID a, $t_{2,a}$: gene expression profile of patient ID a). If T_1 and/or T_2 contain(s) quantitative attributes, the calculation of membership values for category “Low” and category “High” would be required as a pre-processing step.

First, the fuzzy Apriori algorithm detects frequent itemsets in T_1 and T_2 independently with user-specified minimum support. Support can be expressed as follows:

$$support\,(X \rightarrow Y) = \Sigma_{(a \in \{1, 2, \ldots, m\})} \min(X(a), Y(a))/m, \quad (20)$$

where

$X(a)$: a membership value of X in observation a, and

$Y(a)$: a membership value of Y in observation a.

Second, association rules are generated, so that antecedents are selected from the frequent itemsets detected in T_1, and consequents are selected from those detected in T_2, or vice versa. Several scores can be used as thresholds to limit the number of rules to the output; for example, lift can be expressed as follows:

$$lift\,(X \rightarrow Y) = support\,(X \rightarrow Y)/\{support(X) * support(Y)\}, \quad (21)$$

where

X: a frequent itemset constituted of I_1,

Y: a frequent itemset constituted of I_2,

$$support(X) = \Sigma_{(a \in \{1, 2, \ldots, m\})} X(a)\,/\,m, \quad (22)$$

and $$support(Y) = \Sigma_{(a \in \{1, 2, \ldots, m\})} Y(a)\,/\,m. \quad (23)$$

This novel algorithm would enable the identification of relevant items in heterogeneous datasets that associate with each other.

- **Pre- and post-processing of the liver toxicity dataset for the experiments**

In the original biological data that we used for the experiment (liver toxicity data), histopathological observations were described as “minimal,” “mild,” “moderate,” or “marked.” For this experiment, they were converted into “0,” “1,” “2,” or “3,” respectively. Authors of the original paper had reported 50 and 150 mg/kg body-weight ratios of acetaminophen as subtoxic, and 1500 and 2000 mg/kg body-weight ratios as severely toxic [17]. Hence, the column of dose levels in data2 was converted into binary attributes to represent their toxicity level (50 and 150 mg/kg body-weight: 0; 1500 and 2000 mg/kg body-weight: 1). In addition, the values in the column containing the time points in data2 were converted into binary attributes to represent their toxicity level (6, 18, 48 h: 0; 24 h: 1) because 24 h was reported to be the time of peak toxicity, and rats were in a recovery phase after 48 h of acetaminophen treatment [17]. In data1 (gene expression profile), the genes were indicated by Agilent probe IDs. DAVID [37] was used to convert Agilent probe IDs into Entrez gene IDs and gene names.

Experiments

The AI Bridging Cloud Infrastructure (ABCI) operated at the National Institute of Advanced Industrial Science and Technology (AIST), Japan, was used for the experiments.

Implementation

Our method is implemented in Python 3.0 and is dependent on the *pandas*, *joblib*, and *os* modules. The detection of frequent itemsets was conducted by a modified *apriori* function in the *mlxtend* Python module, to introduce fuzzy logic.

Declarations

Acknowledgments

This study was supported by the Ministry of Health, Labour and Welfare [grant number JPMH20AC5001 to Y.N.K. and grant number 19AC5001 to K.M.] and Cabinet Office of Japan Government for the Public/Private R&D Investment Strategic Expansion PrograM (PRISM). We thank Yuji Kosugi (Lifematics Ltd., Tokyo, Japan) for his technical support.

Author contributions

Y.N.K., K.M., and N.U. conceived the study. Y.N.K. and N.U. developed the algorithm. Y.N.K. conducted the analyses. All authors contributed to the writing of the manuscript.

Patents

Y.N.K. and N.U. have a patent application on the subset-binding algorithm.

Conflicts of Interest: none declared.

References

1 Yang, Y. & Wang, H. Multi-view clustering: A survey. *Big Data Min. Anal.* **1**, 83-107 (2018).

2 Ye, F. *et al.* New approaches in multi-view clustering. *Recent Appl. Data Clustering*, 195 (2018).

3 Kumar, A. & Daumé, H. A co-training approach for multi-view spectral clustering. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 393-400 (2011).

4 Tzortzis, G. & Likas, A. Kernel-based weighted multi-view clustering. *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, 675-684 (2012).

5 Li, J.-H., Wang, C.-D., Li, P.-Z. & Lai, J.-H. Discriminative metric learning for multi-view graph partitioning. *Pattern Recognit.* **75**, 199-213 (2018).

6 Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinf.* **9**, 559 (2008).

7 Hipp, J., Güntzer, U. & Nakhaeizadeh, G. Algorithms for association rule mining—A general survey and comparison. *SIGKDD Explor.* **2**, 58-64 (2000).

8 Zhao, Q. & Bhowmick, S. S. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 135 (2003).

9 Agrawal, R., Imieliński, T. & Swami, A. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* **22**, 207-216 (1993).

10 Agarwal, R. & Srikant, R. Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, 487-499 (1994).

11 Pazhanikumar, K. & Arumugaperumal, S. Association rule mining and medical application: A detailed survey. *Int. J. of Comput. Appl.* **80** (2013).

12 Srikant, R. & Agrawal, R. Mining quantitative association rules in large relational tables. *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, 1-12 (1996).

13 Au, W.-H. & Chan, K. C. FARM: A data mining system for discovering fuzzy association rules. *Proceedings of the 1999 IEEE International Fuzzy Systems. FUZZ-IEEE'99* **3**, 1217-1222 (1999).

14 Gyenesei, A. A fuzzy approach for mining quantitative association rules. *Acta Cybern.* **15**, 305-320 (2001).

15 Delgado, M., Manín, N., Martin-Bautista, M., Sanchez, D. & Vila, M.-A. Mining fuzzy association rules: An overview. *Soft Comput. Inf. Proc. Anal.*, 351-373 (2005).

16 Solanki, S. K. & Patel, J. T. A survey on association rule mining. *Proceedings of 2015 Fifth International Conference on Advanced Computing & Communication Technologies*, 212-216 (2015).

17 Bushel, P. R., Wolfinger, R. D. & Gibson, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Syst. Biol.* **1**, 15 (2007).

18 Nassar, I., Pasupati, T., Judson, J. P. & Segarra, I. Histopathological study of the hepatic and renal toxicity associated with the co-administration of imatinib and acetaminophen in a preclinical mouse model. *Malays. J. Pathol.* **32**, 1-11 (2010).

19 Kishi, S. *et al.* Preexisting diabetes mellitus had no effect on the no-observed-adverse-effect-level of acetaminophen in rats. *J. Toxicol. Sci.* **45**, 151-162 (2020).

20 Martínez-Chantar, M. L. *et al.* Importance of a deficiency in S-adenosyl-L-methionine synthesis in the pathogenesis of liver injury. *Am. J. Clin. Nutr.* **76**, 1177S-1182S (2002).

21 James, L. P., Mayeux, P. R. & Hinson, J. A. Acetaminophen-induced hepatotoxicity. *Drug Metab. Dispos.* **31**, 1499-1506 (2003).

22 Lu, S. C. & Mato, J. M. S-adenosylmethionine in liver health, injury, and cancer. *Physiol. Rev.* **92**, 1515-1542 (2012).

23 Rouillard, A. D. *et al.* The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016** (2016).

24 Paradies, G., Paradies, V., Ruggiero, F. M. & Petrosillo, G. Oxidative stress, cardiolipin and mitochondrial dysfunction in nonalcoholic fatty liver disease. *World J. Gastroenterol.* **20**, 14205 (2014).

25 Bailey, S. M. *et al.* S-adenosylmethionine prevents chronic alcohol-induced mitochondrial dysfunction in the rat liver. *Am. J. Physiol. Gastrointest. Liver Physiol.* **291**, G857-G867 (2006).

26 Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comp. Biol.* **13**, e1005752 (2017).

27 Hyvärinen, A. & Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **13**, 411-430 (2000).

28 Yao, F., Coquery, J. & Lê Cao, K.-A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinf.* **13**, 1-15 (2012).

29 Shen, H. & Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.* **99**, 1015-1034 (2008).

30 Lê Cao, K.-A., Rossow, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* **7**, 35 (2008).

31 De Cock, M. *et al.* Fuzzy association rules: A two-sided approach. *Proceeedings of the International Conference on Fuzzy Information processing: Theories and Applications (FIP2003)*, 385-390 (2003).

32 Yan, P., Chen, G., Cornelis, C., De Cock, M. & Kerre, E. Mining positive and negative fuzzy association rules. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 270-276 (2004).

33 De Cock, M., Cornelis, C. & Kerre, E. E. Elicitation of fuzzy association rules from positive and negative examples. *Fuzzy Sets Syst.* **149**, 73-85 (2005).

34 Mangalampalli, A. & Pudi, V. Fuzzy logic-based pre-processing for fuzzy association rule mining. *IIIT Hyderabad, India* (2008).

35 Bhatt, U. Y. & Patel, P. A. A recent overview: Rare association rule mining. *Int. J. Comput. Appl.* **107**, 1-4 (2014).

36 Zaki, M. J. Generating non-redundant association rules. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 34-43 (2000).

37 Dennis, G. *et al.* DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, 1-11 (2003).

Tables

Table 1: The patterns to be detected in the artificial datasets

Artificial data (small)		
	Gene expression profile data (100 rows x 200 columns)	Clinical measurements data (100 rows x 200 columns)
Pattern 1	column1_High, column2_High, column3_High	column10_High, column20_High, column30_High
Pattern 2	column20_Low, column40_Low, column60_Low	column40_Low, column80_Low, column120_Low
Pattern 3	column50_High, column75_High, column100_High, column125_Low, column150_Low	column50_High, column100_High, column150_Low
Artificial data (large)		
	Gene expression profile data (1000 rows x 2000 columns)	Clinical measurements data (1000 rows x 2000 columns)
Pattern 1	column1_High, column2_High, column3_High	column10_High, column20_High, column30_High
Pattern 2	column20_Low, column40_Low, column60_Low	column40_Low, column80_Low, column120_Low
Pattern 3	column100_High, column200_High, column300_High, column400_Low, column500_Low	column50_High, column100_High, column150_Low

Table 2: The summary of the experiments with different membership functions

Artificial data (small)									
	Histogram (lift: 10)	Histogram (lift: 8)	Histogram (lift: 5)	Zscore (lift: 10)	Zscore (lift: 8)	Zscore (lift: 5)	Min-Max	Sigmoid	Rank
Min support = 0.08	0 (n.d.)	0 (n.d.)	0 (n.d.)	5 (n.d.)	16 (Pattern1,2)	16 (Pattern1,2)	n.d.	n.d.	n.d.
Min support = 0.07	1 (Pattern2)	1 (Pattern2)	1 (Pattern2)	7 (Pattern3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	n.d.	n.d.	n.d.
Min support = 0.06	18 (Pattern1,2,3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	7 (Pattern3)	18 (Pattern1,2,3)	19 (Pattern1,2,3)	n.d.	n.d.	n.d.
Min support = 0.05	18 (Pattern1,2,3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	155 (Pattern3)	569 (Pattern1,2,3)	14853 (Pattern1,2,3)	n.d.	n.d.	n.d.

Artificial data (large)							
	Histogram (lift: 10)	Histogram (lift: 8)	Histogram (lift: 5)	Zscore (lift: 10)	Zscore (lift: 8)	Zscore (lift: 5)	Number of rules generated (detected patterns)
Min support = 0.08	0 (n.d.)	0 (n.d.)	0 (n.d.)	5 (Pattern3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	
Min support = 0.07	17 (Pattern1)	17 (Pattern1)	17 (Pattern1)	5 (Pattern3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	
Min support = 0.06	18 (Pattern1,2,3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	n.d.	n.d.	n.d.	
Min support = 0.05	18 (Pattern1,2,3)	18 (Pattern1,2,3)	18 (Pattern1,2,3)	n.d.	n.d.	n.d.	

Figures

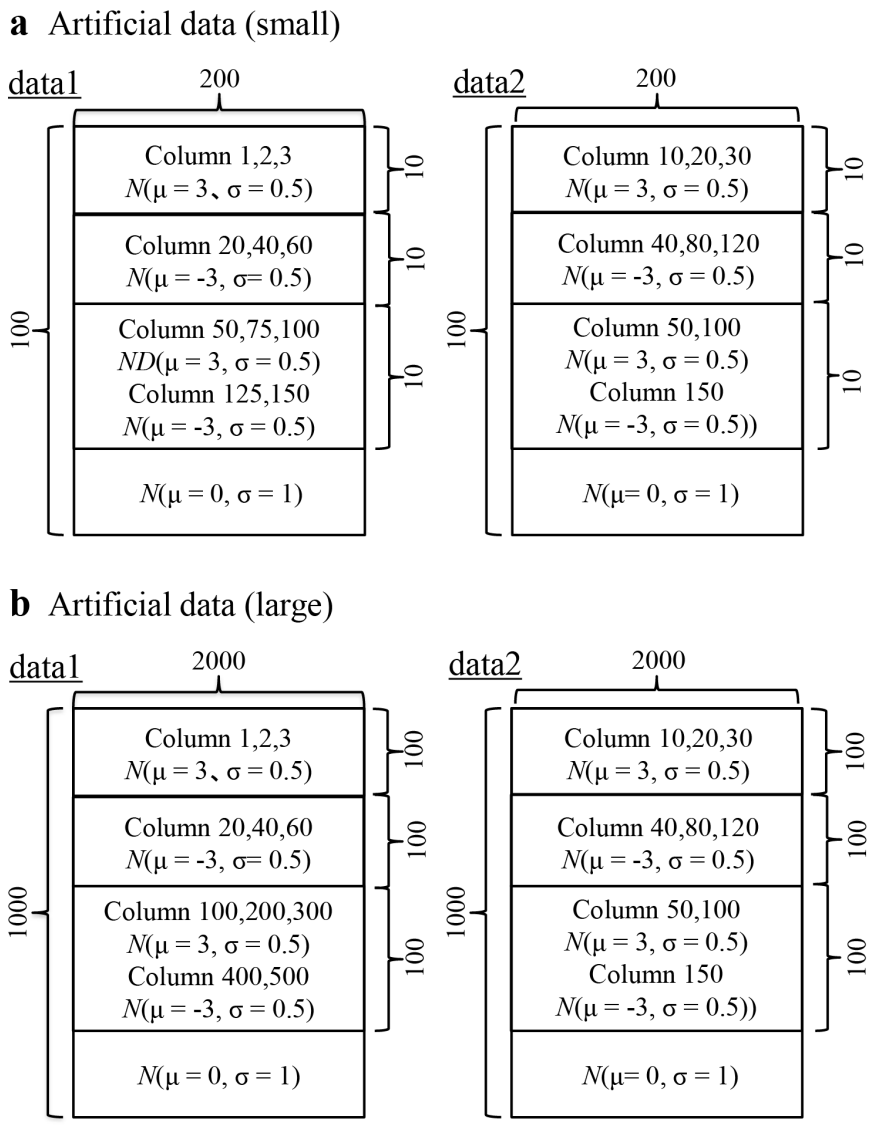
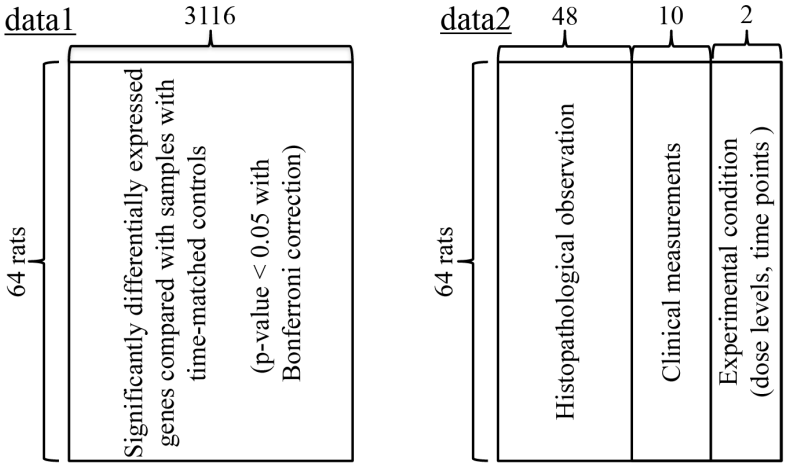


Figure 1

Generation of artificial data. a) Artificial data (small). Two matrices with 100 rows and 200 columns, whose values follow standard normal distribution, were generated assuming that each row represents an observation (e.g., patient) and each column represents an attribute (e.g., a gene or a clinical measurement). Let one matrix be data1 and the other be data2 in Fig. 1a. For the first 10 rows, values in columns 1, 2, and 3 of data1 and values in 10, 20, and 30 of data2 were replaced by values that followed N(3, 0.5). For the second 10 rows, values in columns 20, 40, and 60 of data1 and values in 40, 80, and 120 of data2 were replaced by values that followed N(-3, 0.5). For the third 10 rows, values in columns 50, 75, and 100 of data1 and values in columns 50 and 100 of data2 were replaced by values that followed N(3, 0.5), and values in columns 125 and 150 of data1 and a value in column 150 of data2 were replaced by values that followed N(-3, 0.5). b) Artificial data (large). Two matrices with 1000 rows and 2000 columns, whose values follow the standard normal distribution, were generated and certain values were replaced, as shown in Fig. 1b, using the same procedure as under “Artificial data (small)” in Fig. 1a.

a The liver toxicity data used for the experiment



b Paired-association rule detected by this experiment

Probe	Gene (Entrez)	Gene name	High / Low
A_43_P10003	288444	heat shock protein family H (Hsp110) member 1(Hsph1)	High
A_42_P717602	171347	methionine adenosyltransferase 2A(Mat2a)	High
A_42_P484423	303698	phosphatidylglycerophosphate synthase 1(Pgs1)	High
A_43_P17455	117251	dynein, axonemal, heavy chain 9(Dnah9)	Low
A_43_P14864	58945	dynein light chain LC8-type 1(Dynll1)	High
A_43_P16523	361578	nodal modulator 1(Nomo1)	High
A_43_P19279	363168	lysozyme-like 4(Lyzl4)	Low
A_43_P12811	84596	spermidine synthase(Srm)	High
A_42_P655825	365215	SMG9 nonsense mediated mRNA decay factor(Smg9)	High
A_42_P804499	n.d.	n.d.	High

Clinical measurement	Description	High / Low
ALP (alkaline phosphatase)	positive correlation with bile flow interruption	High
ALT (alanine aminotransferase)	positive correlation with liver injury	High
AST (aspartate aminotransferase)	positive correlation with liver injury	High
cholesterol	negative correlation with liver injury	Low
TBA (total bile acid)	positive correlation with bile flow interruption	High
Histopathological observations	**Description**	**High / Low**
LLL_Centrilob_Necrosis_High	Necrosis of the centrilobular tissue of the liver lobule	High
LLL_Hepat_Hypertrophy_High	Enlargement of the liver	High
LML_Centrilob_Necrosis_High	Necrosis of the centrilobular tissue of the liver lobule	High
LML_Sinusoid_Cogestion_High	Enlargement of the hepatic capillaries	High
Experimental conditions	**Description**	**High / Low**
Toxic_Dose	50, 150 mg/kg BW : subtoxic, 1500, 2000 mg/kg BW: severely toxic	High
Toxic_Time	Peak of the severe toxicity: 24hrs after exposure	High

Figure 2

Summary of results of experiment with liver toxicity dataset a) Liver toxicity data [17] used for the experiment. We applied our algorithm to this dataset to find genes (data1) that related to histopathological observations and/or clinical measurements (evaluation criteria for the degree of liver toxicity by administration of acetaminophen) and/or experimental conditions of administration of acetaminophen (data2). Data1 had 64 rows (rats) and 3116 columns (genes), and data2 had 64 rows (rats) and 60 (= 48+10+2) columns (48 histopathological observations, 10 clinical measurements, and 2 experimental conditions). b) Paired-association rule detected by this experiment. Among the 3986 paired-association rules detected (threshold: lift (frequent itemsets from data1 → frequent itemsets from data2) = 4.8), one rule with the highest conviction is shown in Fig. 2b. Ten genes were related to five clinical measurements, four histopathological observations, and two experimental conditions.

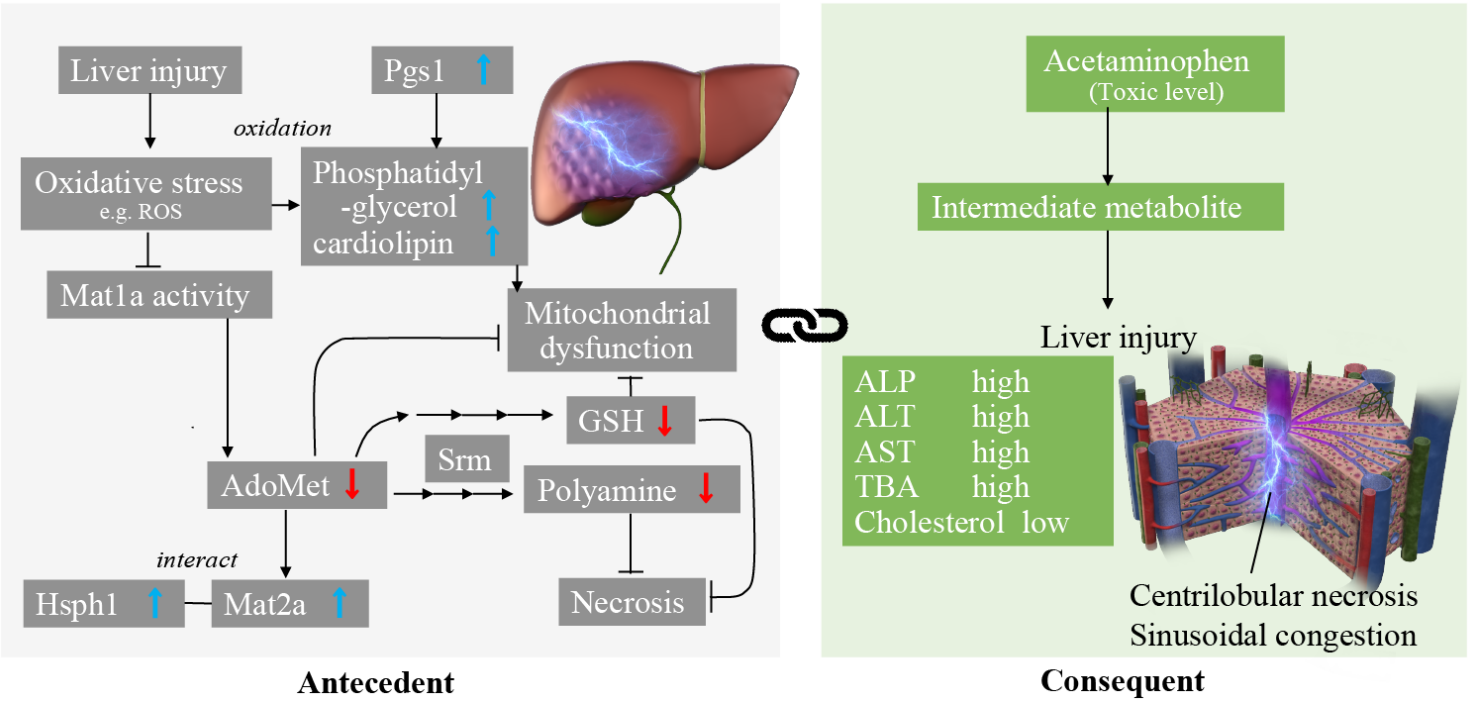


Figure 3

Inferred molecular basis of liver injury by acetaminophen Among the genes shown in Fig. 2b, 4 were found to be involved in liver injury. Their relationships and associated biological events are depicted. Hsph1: heat shock protein family H (Hsp110) member 1, Mat2a: methionine adenosyltransferase 2A, Pgs1: phosphatidylglycerophosphate synthase 1, Srm: spermidine synthase, GSH: glutathione.

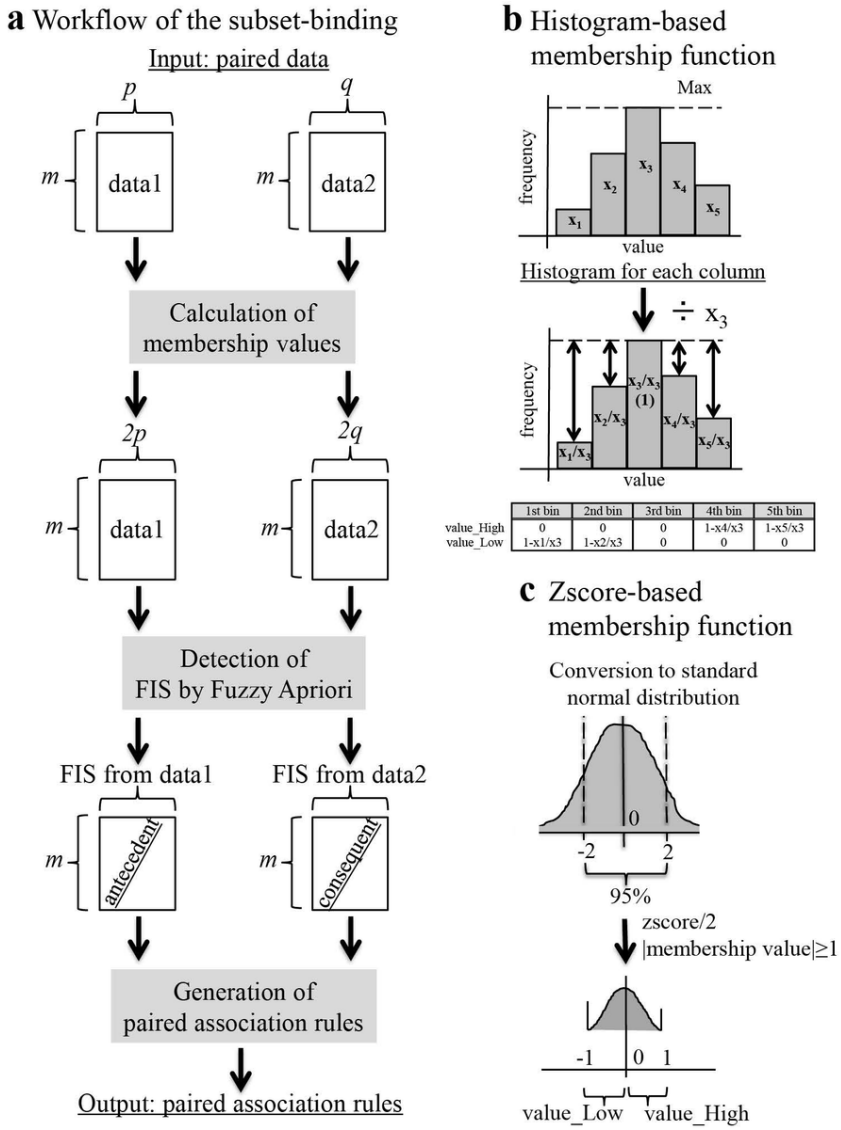


Figure 4

Workflow of the subset-binding algorithm and the membership functions utilized a) Workflow of the subset-binding algorithm. Input data: two paired matrices (quantitative and/or categorical). First, quantitative attributes in input are converted into fuzzy categorical attributes (“Low” and “High”) with membership functions. Because membership values for “Low” and “High” categories are obtained for each attribute, this process doubles the number of columns when all the attributes are quantitative. For example, let the column of gene A be quantitative. The membership function calculates a membership value for both gene A_High and gene A_Low, which results in obtaining two columns. Next, these matrices are used to detect FIS (frequent itemsets) independently. Thereafter, association rules are generated so that FIS derived from one matrix will be antecedent, and those from the other matrix will be consequent. User-specified threshold (e.g., lift) is used for pruning and paired (antecedent from data1 and consequent from data2) association rules are obtained as output. b) Histogram-based membership function. For each attribute (column of input data), a histogram is produced with a user-specified number of bins (e.g., 5). Let the frequency of each bin be x1, x2, x3, x4, and x5, respectively, and x3 be the largest. The formulas to calculate membership values of “Low” and “High” categories are shown at the bottom of Fig. 4b. c) Z-score-based membership function. For each attribute (column of input data), values are converted into z-scores. To scale them between -1 and 1, the z-scores are divided by 2, and values bigger than 1 and those smaller than -1 are converted to 1 or -1, respectively. The obtained values are used as membership values of “Low” and “High” categories.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata1.csv
- Supplementarydata2.csv

- Supplementarydata4.csv.zip
- Supplementarydata5.csv
- Supplementaryinformation.xlsx
- SupplementaryTable1.csv
- SupplementaryTable2.csv
- SupplementaryTable3.csv
- SupplementaryTable4.csv
- SupplementaryTable5.csv
- SupplementaryTable6.csv
- SupplementaryTable7.csv
- SupplementaryTable8.csv
- SupplementaryTable9.csv
- SupplementaryTable10.csv.zip
- SupplementaryTable11.csv
- SupplementaryTable12.csv
- SupplementaryTable13.csv
- SupplementaryTable14.csv
- SupplementaryTable15.csv.zip
- Supplementarydata3.csv.zip