Association rule mining to shortlist plant phenolic compounds likely to decrease methane emissions by ruminants

Sylvie Guillaume, Didier Macheboeuf

Association rule mining to shortlist plant phenolic compounds likely to decrease methane emissions by ruminants

Sylvie Guillaume and Didier Macheboeuf

Abstract The purpose of this work was to find phenolic compounds in plants that could act on ruminal fermentations to limit methane emissions by ruminants, in order to propose natural additives or food alternatives. We used a data mining method to extract class association rules that would identify compounds likely to have a significant effect. Such extraction usually generates a large number of rules. Our problem was to select the best rules, and thereby the most promising compounds. We carried out a new kind of extraction: mining for *strongly expressed* rules, that is to say rules that govern whether compounds are abundant in the plants. We propose two new interesting measures to evaluate the intensity of expression rules, and a new type rule visualization. Among the *1,075* phenolic compounds found in the 208 plants analysed, *7* promising compounds and *5* useful associations of compounds were shortlisted.

1 Introduction

Ruminants are important to mankind not only because they provide useful produce such as milk, wool or meat, but also because they have a crucial ability to digest plant fibres. This is due to the microbial ecosystem in the hosts rumen considered to be one of the most highly diverse ecosystems in terms of species diversity and functional richness [Mizrahi and Jami, 2018]. The fermentation process results in volatile fatty acids. These nutrients are absorbed and serve as the main energy source for the animal. Gases (CO_2 *and methane*) are natural by-products of the process and

Sylvie Guillaume
CNRS, UMR 6158, LIMOS, Université Clermont Auvergne, F-63173 Aubière, France, e-mail: sylvie.guillaume@uca.fr

Didier Macheboeuf
Université Clermont Auvergne, INRA, VetAgro Sup, UMR Herbivores, F-63122 Saint-Genès-Champanelle, France e-mail: Didier.Macheboeuf@inra.fr

are expelled into the atmosphere by eructation. Methane is produced by microorganisms from domain Archaea, which are strict anaerobes that use the metabolic pathways of methanogenesis [Garcia et al., 2000]. The main electron acceptor is CO_2 which is reduced to methane by the hydrogenotrophic pathway and less so methyl-group and acetate converted to methane by the methylotrophic or acetoclastic pathways [Patra et al., 2017, Garcia et al., 2000].

In terms of the gross energy intake from the feed, methane emissions can be considered as energy lost (*in the process*). This is estimated at 5 - 7% [Hristov et al., 2013] or 2 - 12 % [Huws et al., 2018] depending on the type of feed or feed quality. This energy inefficiency represents a lack of production and an economic loss for the farmer. In addition, methane is a powerful greenhouse gas (GHG) that has a 28- to 34-fold higher global warming potential than CO_2 on a 100-y horizon [Duin et al., 2016] but it depends on the metrics used [Lynch, 2019]. Enteric methane emissions mainly coming from ruminant livestock, comprise 17 and 3.3% of global methane and GHG emissions annually [Knapp et al., 2014]. In the European Union, there were about 150 million-metric-tonne of CO_2 equivalent (*data of 2011*). As livestock productions will continue to increase to contribute to the food security of the growing human population [Dangal et al., 2017], the mitigation of methane emissions has become a research priority in ruminant nutrition for the implementation of sustainable and environmentally friendly livestock systems.

Numerous methane mitigation strategies have been explored [Patra et al., 2017]. They can be classified into 3 broad groups: animal genetic, feed management and rumen modifiers [Knapp et al., 2014]. In the last group, since the ban on the use of antibiotics and all synthetic additives in animal feed in the European Union from 2006 (*regulation 1831/2003/EC*), there has been a great demand for natural plant extracts that are effective additives to manipulate ruminal fermentations and limit methanogenesis. Research has been directed towards secondary plant metabolites. These chemical compounds that are not essential for the constitution of plants, but are produced in response to stress (*e.g. water*) or aggression (*insects, micro-organisms, herbivores*) or during their reproduction. It would therefore be of interest to know the effect of these compounds on methane emissions by the ruminal ecosystem. However, the very wide diversity of their chemical structures, estimated at more than 200,000 [Patra and Saxena, 2010], makes testing the activity of all these compounds unfeasible. While many studies have focused on certain types of these compounds (*essential oils, saponins, tannins*), very little is known about the effect of small phenolic compounds on the methanogenesis. Here we focus on low-molecular-weight ($<$ 1000 *Dalton*) phenolic compounds.

We screened 208 plants by *in vitro* fermentation to identify bioactive plants against methane emission. Some of these plants were chosen for their medicinal properties such as anti-microbial properties that could be useful in handling Archaea. Others, on the contrary, were chosen due to the lack of information. The choice took into account the results of other research teams in order to avoid redundancies along with the need some common plants for comparison. At the same time, the profile of small phenolic compounds extracted from these plants was determined for each plant. At this stage, the compounds were not identified, because there were

on average more than a hundred compounds per plant (*detected as peaks by the analytical method*). It was imperative to first select a small number of compounds involved in the plants anti-methanogenic effect, because the subsequent steps, namely the identification of the compounds, and the *in vitro* validation of the expected effect with the pure compound, are costly and time-consuming.

Given the very large fluctuations in the relative abundance of compounds in the profiles, and their low frequencies in plants, it is difficult to link the presence of a compound or combination to the plants activity. We therefore used data mining, specifically the association rule mining, to select a few compounds that could have an effect. The large number of projects focusing on association rules in research papers is a good evidence of the importance of this data mining task.

Association rule mining is one of the most popular data mining methods and is a powerful method for discovering the relationship between variables in large databases. This method has the advantage of overcoming the primary limitations of the general linear models [Vougas et al., 2019]. Except in the domain of soil, association rules have rarely been used in agricultural research. However, we found this method interesting for three reasons. First, the method can quickly highlight synergic effect between 2 or more compounds. Secondly, the method is well adapted to the data particularities related to the acquisition of profiles of phenolic compounds. This analytical method is semi-quantitative. The data are quantitative when comparing the same compound which is found at different concentrations in several plants but the data is qualitative because the response factor of the phenolic structure, measured at 280 nanometers (nm) can be very different from one compound to another. As a result, the data has been binarized and the association rules are a simple and easy way to process this type of data. Thirdly, our search for compounds active against methane emissions is similar to the search for efficient drugs in the pharmacological field where associative methods have already been used [Vougas et al., 2019, Wu et al., 2018]. It is therefore interesting to know whether this method could also be used in relation to our problem.

Furthermore, association rule discovery is a well-defined, deterministic task i.e. any association algorithm discovers precisely the same rule set, the differences in the proposed algorithms are mainly their relative efficiency (*some algorithms are faster than others*). In contrast to association rule discovery, classification is an ill-defined, non-determinist task i.e. using only the training data, one cannot be sure that a discovered classification rule will have a high predictive accuracy on the test set (*set which contains examples unseen during training*) [Freitas, 2000]. A very well-known technique is decision trees (*see* [Loh, 2011] *for an overview*). In addition, decision trees automatically split numerical variables and sometimes repeat the same attribute several times [Ordonez, 2006] which is not suited to our problem because we are only searching for rules with high values as we explain in Sect. 6. Finally, a decision tree partitions the data set whereas association rules on the same target attribute may refer to overlapping subsets [Ordonez, 2006], which is well suited to our problem.

This paper is organized as follows. Sect. 2 gives a brief reminder about association rules. Sect. 3 presents the data and how it was acquired. The rest of the paper fo-

cuses on knowledge discovery using this data. Sect. 4 presents the process of knowledge discovery in databases that was used, and explains why two kinds of extraction were needed to find the compounds and the component associations that could have a positive effect on methane emissions (*i.e. reduction of methane emissions*). We name these promising compounds and promising associations. The first extraction, carried out on the binary data, discovered all the class association rules [1], i.e. all the potentially promising compounds and associations reducing methane emissions. In fact, the association rule extraction, introduced in [Agrawal et al., 1993], discover interesting relationships between binary attributes only, this is why we work on the binary data. The second extraction, carried out on the numerical compound data and explained in Sect. 6, discovered all the class strongly expressed rules (*rules that govern whether compounds are abundant in the plants*), i.e. all the most promising compounds and associations. We know wether components are abundant with the determination of their phenolic profiles, this is why we work on the numerical compound data. A comparison of these two extractions enabled us to select promising compounds and associations. We conclude by indicating the families of the promising compounds identified.

2 Recall Association Rules

The purpose of association rule extraction, introduced in [Agrawal et al., 1993], is to discover interesting relations between binary attributes (*or variables*) in large databases.

Formally, the association rule problem can be stated as follows: Let $\mathscr{I} = \{i_1, i_2, ..., i_p\}$ be the set of items (*or binary attributes*), and let $\mathscr{T} = \{t_1, t_2, ..., t_n\}$ be the set of transaction identifiers or *tids*. The input database *DB* is a binary relation $\delta \subseteq \mathscr{I} \times \mathscr{T}$. The database consists of a set of transactions (*or individuals, objects*), where each transaction $t \in \mathscr{T}$ contains a set of items, such that $t \subseteq \mathscr{I}$. A set $X \subseteq \mathscr{I}$ is called an *itemset*, and a set $T \subseteq \mathscr{T}$ is called a *tidset*. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq \mathscr{I}$ are two sets of items and $X \cap Y = \emptyset$. The intuitive implication of the association rule is that presence of the set of items X in a transaction set also indicates possible presence of the itemset Y.
An example of an association rule extracted from a database of supermarket sales is: $pancakes, butter \Rightarrow cider$. This rule states that the customers who buy pancakes and butter also tend to buy cider.
Two notions for establishing the strength of a rule are those of *minimum support* and *minimum confidence* defined by the user. The *support* $sup(X \Rightarrow Y)$ of a rule $X \Rightarrow Y$ defines the range of the rule, i.e. the fraction of transactions that contain both X and Y. The *confidence* $conf(X \Rightarrow Y)$ of a rule $X \Rightarrow Y$ defines the precision of the rule, i.e. the fraction of transactions containing X that also contain Y.
An association rule is considered relevant if it has support and confidence at least

[1] A class association rule is a special case of association rule where the right-hand is a target class while the left-hand may contain one or more attributes.

equal to some user-specified minimal support min_{sup} and confidence min_{conf} thresholds.
An association algorithm simply returns all the rules having support and confidence greater than user-specified thresholds. Among all these rules, the algorithm has no criterion (no bias) to select one rule over another.

Association rule mining can be decomposed into the following subproblems [Agrawal and Srikant, 1994]:
(1) **Find all frequent itemsets**: we generate all combinations of items that have support above a minimum support (i.e. $sup(X) \geq min_{sup}$).
(2) **Find all valid rules**: for each frequent itemset X, we generate rules of the kind $X - Y \Rightarrow Y$ for each $Y \subset X$. Once these rules have been generated, only those rules above minimum confidence need be retained (i.e. $conf(X - Y \Rightarrow Y) \geq min_{conf}$).

Two other popular interest measures used for rule mining:

1. `Lift` introduced by [Brin et al., 1997] is defined as:

$$lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{sup(Y)} \tag{1}$$

   Lift "measures how many times more often X and Y occur together than expected if they were statistically independent" [Hahsler, 2015].
2. `Leverage` introduced by [Piatetsky-Shapiro, 1991] is defined as:

$$leverage(X \Rightarrow Y) = sup(X \Rightarrow Y) - sup(X)sup(Y) \tag{2}$$

   Leverage "measures the difference of X and Y appearing together in the data set and what would be expected if X and Y were statistically dependent" [Hahsler, 2015].

3 Data Presentation

The substrates used for *in vitro* fermentation and for the determination of phenolic compound profiles were obtained from *208* plant species harvested in the French Massif Central. Samples were frozen in liquid nitrogen to fix secondary metabolites, freeze-dried and ground.

Fig. 1 shows the process by which the data was acquired, explained in Sect. 3.1 and 3.2.

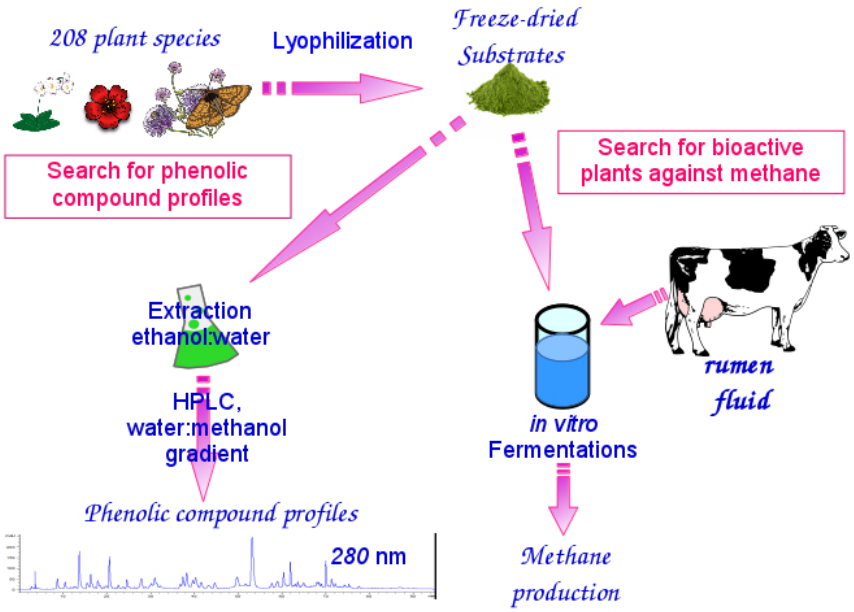

Fig. 1 Process by which the data was acquired.

3.1 Plant Phenolic Compounds (Descriptive Variables)

The phenolic compounds were extracted from the substrates with solution water:ethanol (20:80), and the profiles were analyzed by an High Pression Liquid Chromatographic system for *95* minutes on a *C18* column with a water:methanol gradient using a method adapted from [Sakakibara et al., 2003]. The chromatographic profile of each plant was recorded at the frequency of *280* nanometers (nm) with a diode array detector. A standard mixture of *21* known phenolic compounds is injected in all sequences of analysis and allows the alignment of chromatograms.

Among these standard compounds, flavone was used as a reference for calculating relative retention times (T_i) of peaks. The sequences were aligned by repositioning the T_i values of the standards. Since the retention time of the flavone was on average *84.52* min, the variation range of T_i lay in the interval [0, 1.124] for a separation over *95* min. Since the plant compounds were unknown, they were identified by their relative retention times T_i, $i \in [0, 1.124]$ (*see Fig. 2 for an example of chromatographic profile*).

A total of *1,075* different compounds were detected in the set of *208* plants. A mean of *106* compounds were found per plant, with a range from *29* to *161*. The number of occurrences of a compound in the plant set was highly variable, ranging from a single occurrence to a frequency of occurrence of nearly *58%*. The average frequency of occurrence was equal to *10%*. The data for compounds was peak area

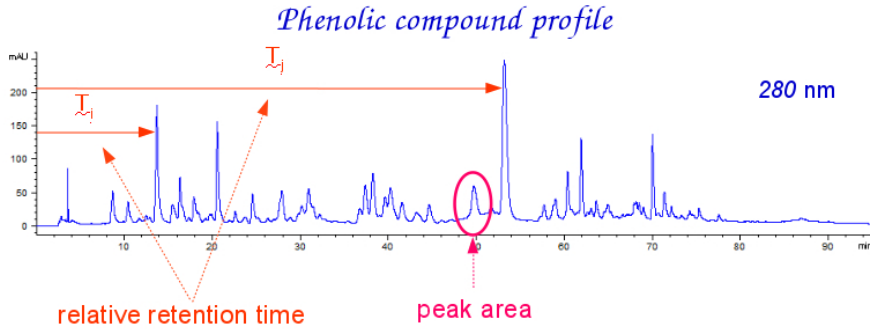

Fig. 2 Example of a chromatographic profile.

if the compound was present (*see Fig. 2*). Here again, the variations were extremely wide, ranging from *3* mAU (*detection threshold of the analysis system*) to *105,200* milli arbitrary units (*mAU*). The median peak area was equal to *100* mAU. The value of *10* times the area of the median peak was taken to sort minor peaks ($<$ *1,000 mAU*) from major peaks. The number of major peaks averaged *9* per plant, with a range from *0* to *43*.

The raw data was structured into a matrix of *208* (*plants*) x *1,075* (*compounds*) at *280* nm containing the numerical values of the peak areas. This matrix had a low filling rate of *10%*. Highly frequent compounds ($>$ *30% frequency*) were discarded to avoid false positives (*i.e. 28 compounds*). We therefore had a matrix composed of *208* plants described by *1,047* compounds: this formed database D_1 (*see Fig. 3*).

In the rest of this paper, T will have the following meaning: an itemset of compounds T_i.

3.2 The Anti-Methanogenic Index (Target Variable)

The particularity of this data mining process is that it involves only one target variable: the anti-methanogenic index (*AMI*). The *AMI* was built from *in vitro* rumen fermentation data. The fermentation profiles (*gas and volatile fatty acids (VFAs) produced*) were determined for all substrates after *24* hours of fermentation at *39* °C *in vitro* in rumen simulated systems. All incubations were repeated *3* times. Each run included perennial ryegrass (*PRG*) as control. Methane and *VFAs* productions were normalized and expressed as a ratio of mean *PRG* values for each period to remove inter-period drift. The normalized methane production was therefore a column vector of dimension *208* without missing data, whose values (*average of three repetitions*) were ratios between *0.10* and *1.33*. The index was calculated as [Macheboeuf et al., 2018]:

$AMI = (A_f - A_m)/A_{max}$, where A_f is the CH_4 value fitted to the $CH_4 = f(VFAs)$ linear regression minus *2.58* times the *PRG* standard deviation, A_m is the measured CH_4 value, and A_{max} is the maximum $(A_f - A_m)$ value observed among the *208* plant samples.

Concretely, the *AMI* represents the deviation between the measured methane production of a plant and the predicted value that would have been obtained under normal conditions (*without inhibition of methanogenesis*) at equivalent *VFAs* production. Any plant that has an index strictly higher than *0* has a significant anti-methanogenic effect ($p < 0.01$). The index ranged from *-0.74* to *1.00* with a mean of *-0.09*. The index was converted to binary data and named *indMeO*. The anti-methanogenic effect was observed ($AMI > 0$) in *64* plants, i.e. *30.77%* of the plant samples, for which *indMeO* took the value *1*. This formed database D_2 (*see Fig. 3*).

After showing how the data was acquired, we now describe the extraction process we used.

4 Process of Knowledge Discovery in Databases

The process of Knowledge Discovery in Databases (*KDD*) is defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [Fayyad, U.M. et al., 1996].

Many studies have used graph mining to analyze chemical/molecular data and thereby detect active effects [Berasaluce et al., 2004] [Fischer and Meinl, 2004]. This approach is not possible in our case because we do not know the chemical structures of the phenolic compounds. Obtaining this knowledge requires much time and technology (*mass spectrometer*). Furthermore, the plants were chosen deliberately because they were not well-studied, and their compositions were in no bibliography or database. Our HPLC-DAD set-up enables us to record UV spectra, and examination of these spectra by an expert makes it possible to classify unknown compounds into molecular families. However, the classification process is manual, laborious and clearly unfeasible when there are 1,047 compounds to classify. However, UV spectra can be studied over a small number of peaks, and according to the spectrum class identified, some compounds can be identified.

Our chosen strategy was therefore to use data mining with the hypothesis that all peaks were different although some might be similar. Association rule discovery was well-suited to solving our problem, since we were seeking relations between compounds and their anti-methanogenic index.

The KDD process we used is presented in Fig. 3 and explained below.

1. Selection and preprocessing steps
   The selection and preprocessing steps are set out in Sect. 3.1 and 3.2 where

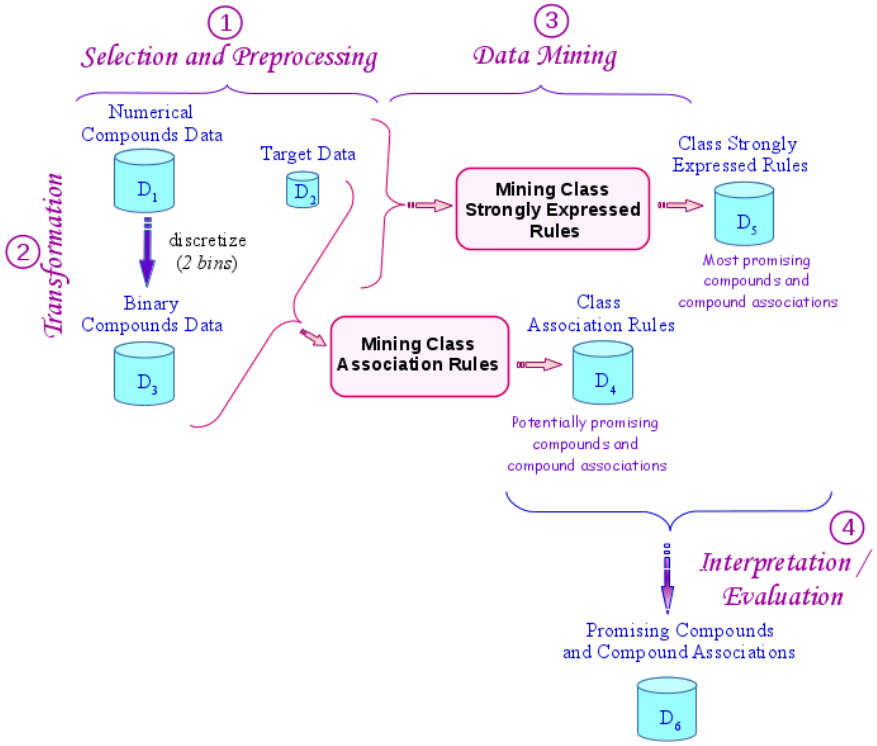

Fig. 3 Steps comprising the KDD process.

we describe how we obtained sets D_1 and D_2.

2. Data Transformation step

   Data transformation is a preprocessing task that includes any procedure that modifies the original form of the data. Mining association rules require that the data be in binary form, as mentioned in [Agrawal et al., 1993]. In the presence of numerical variables, discretization followed by complete disjunctive coding is necessary.

   We recall, as stated in Sect. 3.1 that the values taken by the peak areas differ widely from one compound to another, and the frequencies of occurrence of the compounds are also wide-ranging (*see Fig. 4*). Given these characteristics, we discretized numerical variables as follows: the value *1* was assigned to all the compounds expressed, i.e. for all the values strictly higher than *0*; and the value *0* to all those non-expressed. This formed database D_3 mentioned in Fig. 3. We recall that the target data D_2 is already binary.

   As stated in Sect. 3, one characteristic of the database D_1 is that it is sparse (*i.e. it has many null values*) since the plants only contain about one hundred com-

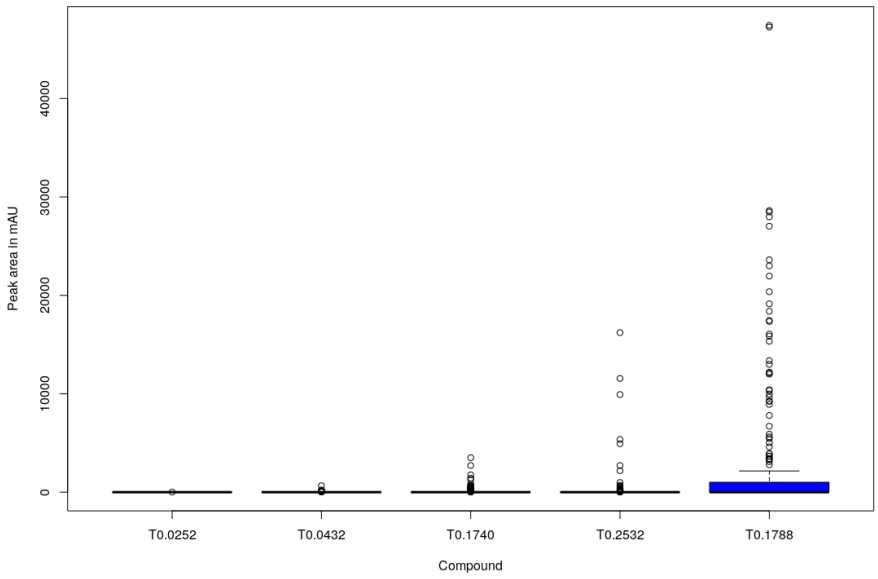

Fig. 4 Boxplots of some compounds.

pounds on average out of the *1,075* detected. The mean frequency of occurrence of compounds in our base was *10%*, a mean of *21* expressions per compound.

3. Data mining step

   We performed two kinds of extraction:

   The first one, carried out on the binary data (D_2 *and* D_3), enabled us to discover all the class association rules, whose conclusion is the binary anti-methanogenic index *indMeO*. Association rules where the conclusion is an item (*i.e. only one binary variable*) were named class association rules. This enabled us to detect all the compounds potentially promising for the desired effect (*anti-methanogenic action*), together with all potentially promising compound associations. At the end of this step, we obtained the set D_4 of class association rules as mentioned in Fig. 3.

   However, this extraction usually generates a large number of rules, and it is imperative to select the best ones.

   [Klemettinen et al., 1994] propose the use of rule templates to easily describe the structure of useful rules according to the user's intuitions, and have developed visualization tools (*such as the Rule Visualizer System*) to show them to the user. This method retrieves those match rules from the set of extracted rules and does not prune insignificant rules.

[Hahsler, 2016] cluster association rules into a small number of meaningful groups. These groups would be valuable for experts who need to manually inspect the rules, for visualization and as the input for other applications.
[Aggarwal and Yu, 1998] present association rules in a compact form, eliminating redundancy. This kind of redundancy arises when rules have more than one item in the consequent. [Bastide et al., 2000] extract minimal non-redundant association rules using frequent closed itemsets. These rules have minimal premises [2] and maximal conclusions [3] (*i.e. the most relevant association rules*). These authors [Pasquier et al., 1999a, Pasquier et al., 1999b] focused mainly on the discovery of frequent closed itemsets, and do not report any experiments on rule mining. [Zaki, 2000, Zaki, 2004] has been mainly interested in generating a smaller non-redundant rule set after mining the frequent closed itemsets.
[Djenouri et al., 2014, Djenouri et al., 2018] prune association rules using meta rules extraction. First, they cluster association rules and then, for each obtained cluster, different dependencies between rules are extracted using a meta-rules algorithm. At the end of the process, the meta rules discovered are used to find the set of representative rules, and the remaining rules are systematically pruned.
Multi-Criteria Decision Making methods have been used to rank the discovered association rules. [Choi et al., 2005] used ELECTREE II method to rank the association rules. [Shukla et al., 2019] proposed an approach based on DEA (*Data Envelopment Analysis*) to rank discovered association rules. DEA measures the efficiency of each discovered association rule based on multiple criteria. These efficiency scores for each rule help to rank these rules for implementation.

We know the rule template to extract and that is the following: what compounds or compound associations have a positive effect (*i.e. reduction of methane emissions*) ? These are rules where the conclusion is an item: the binary anti-methanogenic index. The best rules for biologists are those where compounds are abundant in the plants. The more abundant a compound is in a plant, the greater will be its peak area. Thus for our purposes, a class rule will be of greater utility the more strongly expressed the phenolic compounds are, i.e. the higher their values are. We named such rules "*class strongly expressed rules*". Thus a second extraction carried out on the numerical data D_1 enabled us to discover all the class "*strongly expressed*" rules. We proposed two new interest measures to evaluate the intensity of expression rules.

4. Interpretation/Evaluation step

   At the end of these two kinds of extraction, a comparison of the two sets of rules (D_4 *and* D_5) with their different measures enabled us to select the best rules. This is the set D_6 of rules.

[2] or left hand sides or antecedents.

[3] or right hand sides or consequents.

We now focus on the first data mining step in our KDD process: the discovery of class association rules.

5 Mining The Potentially Promising Compounds and Associations

The extraction was carried out using the package `arulesViz` [Hahsler, 2017] in the `R` statistical software [R Development Core Team, 2020]. This package proposes interfaces to the popular `C` implementations of `Apriori` and `Eclat` by Christian Borgelt [Borgelt and Kruse, 2002, Borgelt, 2003]. We retained the value of 0.025 for the minimum support, i.e. verified by at least 6 transactions (*substrates*). We considered a rule verified by fewer than 6 transactions as untrustworthy. For the minimum confidence, we chose a value strictly higher than *0.50* so that $conf(T \Rightarrow indMeO) > conf(T \Rightarrow \overline{indMeO})$. We recall that the database D_2 had *30%* of anti-methanogenic plants (*sup(indMe0)=0.30*). The minimum confidence therefore guarantees that extracted rules are necessarily in the attractive zone, i.e. the zone where $conf(T \Rightarrow indMeO) > sup(indMeO)$.

We found *676* class rules distributed as follows according to "*order*" (*see Tab. 1*), i.e., the number of items contained in the rule:

Table 1 Number of extracted class rules by order

Order	Number of rules
2	19
3	573
4	82
5	2
Total	676

The scatter plot of Fig. 5 shows us the 676 rules extracted and was generated by the `plot` function in the `R` statistical software. This is a special version of a scatter plot called `two-key plot` introduced by [Unwin et al., 2001]. Support and confidence are used for the `x` and `y`-axes respectively, and the color of the points is used to indicate order. From this plot, we verified that order and support had a very strong inverse relationship, known to be the case for association rules [Seno and Karypis, 2005].

We now study the rules of order *2* that reveal potentially promising compounds.

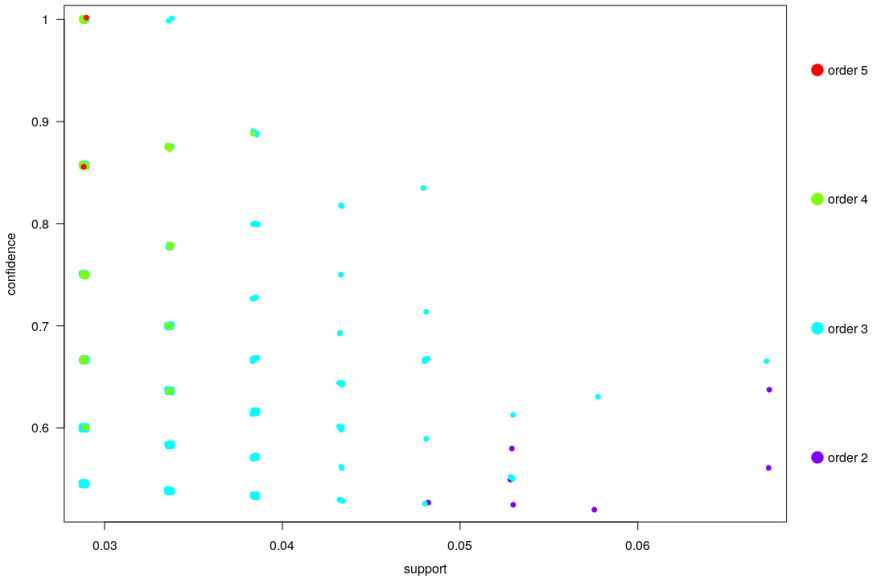

Fig. 5 The 676 rules extracted.

5.1 Mining The Potentially Promising Compounds

The plot in Fig. 6 displays the *19* extracted rules of order *2*. This graph-based visualization was generated with the `R`-extension package `arulesViz`. The size of the circle is proportional to the value of the confidence of the rule (*the larger the size, the higher the value of the confidence*) and the intensity of the color in the circle is proportional to the value of the support of the rule (*the darker the color, the higher the value of the support*).

The 5 most promising rules of order *2* are summarized in Tab. 2. These rules are ranked in decreasing order of confidence.

Table 2 The 5 most promising rules of order *2*.

Premise	Support	Confidence	Lift
T0.9792	0.03	0.70	2.31
T0.0492	0.029	0.67	2.20
T0.9756	0.029	0.67	2.20
T0.0816	0.03	0.64	2.10
T0.4404	0.067	0.64	2.10

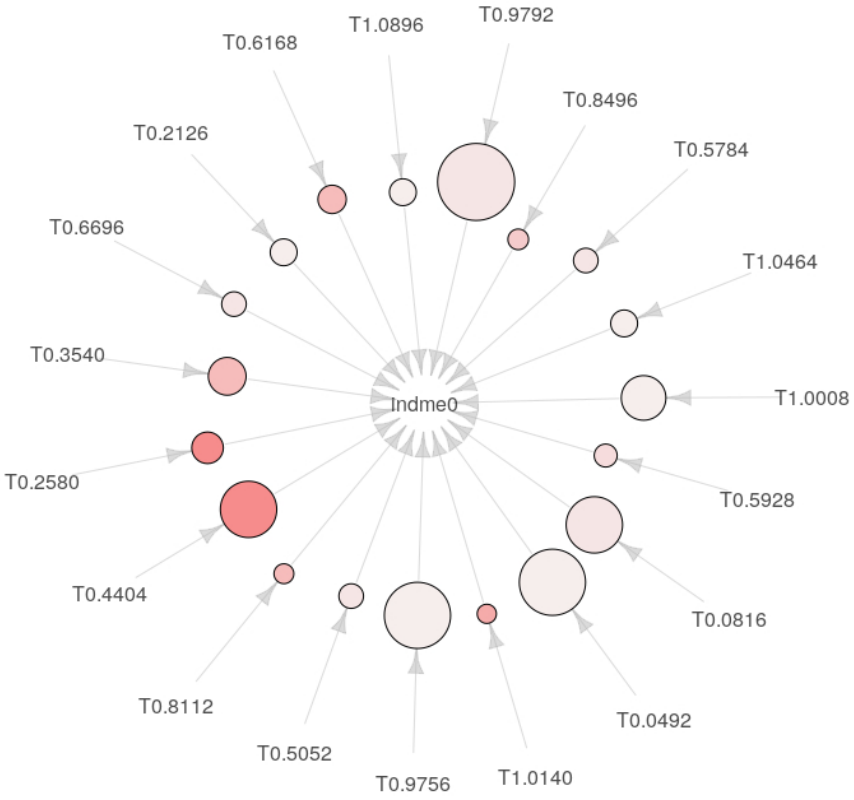

Fig. 6 General Rules $T_i \Rightarrow indMeO$ generated with the `R`-extension package `arulesViz`.

New visualization of rules

The graph in Fig. 6 lets us compare rules for the two main measures (*i.e. support and confidence*), which gives us a good overview of the quality of rules. However, we would like more information to select, as accurately as possible, a manageable number of promising compounds. It would therefore be useful to know how many of the substrates containing the compound have a positive effect (*information given by the value of the absolute support of the rule*) and how many have a negative effect (*methanogenic*). Confidence informs us about this proportion (*for a confidence equal to c and among substrates containing compound T_i, c% have a positive effect and* $(1-c)$*% have a negative effect*) but we would like to visualize these two numbers. For a confidence equal to $c\%$, to have $100 \times c$ substrates that have a positive effect or $1,000 \times c$, does not have the same value. Again, support informs us about this proportion, but we would like to visualize these two values side by side.

The first information corresponds to the number of examples of a rule (*i.e. the number of substrates verifying the premise and conclusion of the rule*) and the second to the number of counter-examples (*i.e. the number of substrates verifying the premise but not verifying the conclusion*). Knowing the proportion of one relative to the other can offer valuable information to select the most promising rules.

We propose the visualization represented by *Fig. 7*.

This graph gives us the following information:

1. The `number of substrates verifying compound` T_i or absolute support $sup_{abs}(T_i)$ of premise T_i, through the length of the line segment (*red and blue segments*).
2. The `number of examples` or absolute support $sup_{abs}(T_i indMeO)$ of rule $T_i \Rightarrow indMeO$ through the length of the line segment to the left of the line of equation $x = 0$ (*red segment*).
3. The `number of counter-examples` or absolute support $sup_{abs}(T_i \overline{indMeO})$ through the length of the line segment to the right of the line of equation $x = 0$ (*blue segment*).
   We recall that $sup_{abs}(T_i) = sup_{abs}(T_i indMeO) + sup_{abs}(T_i \overline{indMe0})$.
4. The `confidence of the rule` through the orientation of the line segment (*the more vertical the line segment, the higher the confidence of the rule*). A horizontal line segment corresponds to a value equal to *0.50* for the confidence, and a vertical line segment corresponds to a value equal to *1*.
   Supports $sup_{abs}(T_i)$ and $sup_{abs}(T_i indMeO)$ represented on the graph enable us to find the confidence of the rule, since $conf(T_i \Rightarrow indMeO) = \frac{sup_{abs}(T_i indMeO)}{sup_{abs}(T_i)}$.
5. A `measure of our choice` on the y-axis. We chose here the interest measure M_G, which evaluates the distance between two characteristic states: (i) equilibrium [4] or independence [5] and (ii) logical implication [6]. For more details on this measure, we refer the reader to the work of [Guillaume, 2010].

This kind of representation is inspired by the Venn diagram, which here is condensed and flattened. We recall that for all these rules, the conclusion is the same, item *indMeO*. The absolute support $sup_{abs}(indMeO)$ is not represented on the graph because (i) it has the same value for all the rules and, (ii) the value of its support is approximately six times greater than the value of the support of all the rules extracted.

If two rules have the same value for the chosen measure on the y-axis, then we perform a translation of the representation of the second rule along the x-axis. This is the case, for example, for compounds *T0.9756* and *T0.4404*.

---

[4] Equilibrium is the case where $conf(X \Rightarrow Y) = conf(X \Rightarrow \overline{Y}) = 0.5$.

[5] Independence is the case where $conf(X \Rightarrow Y) = sup(Y)$.

[6] Logical implication is the case where $conf(X \Rightarrow Y) = 1$.

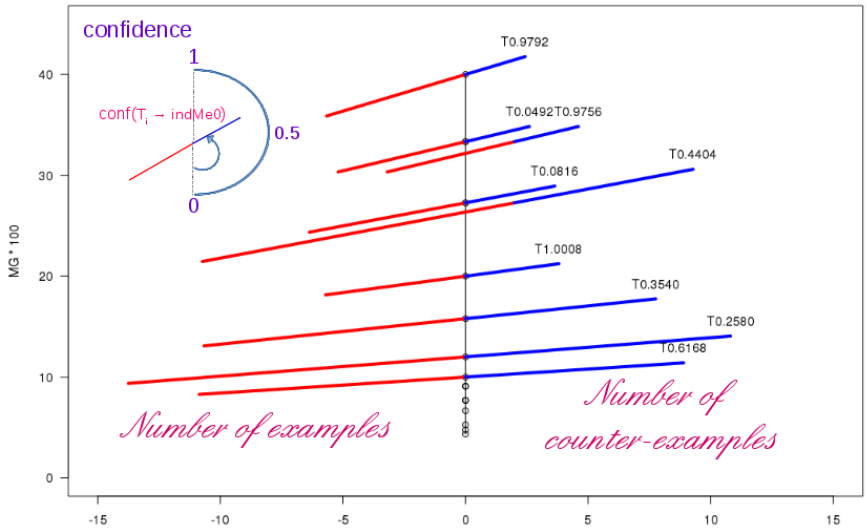

Fig. 7 Visualization of the 9 best rules $T_i \Rightarrow indMeO$.

Such a graph is of interest and readable only for a limited number of rules and with values for the confidence not too close to *1*. If the values of the confidence are close to *1*, then we can represent this information on the *y*-axis and choose another interest measure for the orientation of the line segment.

Clustering of the compounds

We went on to study how these potentially promising compounds clustered, i.e. to what extent they were associated with other compounds. Compounds were classified into *"gregarious"* or *"solitary"* according to whether they were found associated with other compounds or not. To find out whether there were many associations with a given compound, we searched, for each potentially promising compound, the number of maximally frequent itemsets [7] (*noted MMF*) containing it together with the size of the largest MMF (*i.e. the MMF with the largest number of associated items*). The greater the number of MMFs and / or the larger the size of the largest MMF, the more *"gregarious"* will be the compound.

To extract these MMFs, we need to add the parameter `target="maximally frequent itemsets"` in the `apriori` function of `R` software. *3,765* MMFs were detected among the *339,532* frequent itemsets containing the item *indMeO*. Tab. 3 (*columns 1, 6 and 7*) shows the result of this extraction ordered by num-

[7] A frequent itemset is maximal if none of its supersets are frequent and all its subsets are frequent.

ber of MMFs extracted. *8 "solitary"* compounds were detected among the *19* potentially promising compounds. These *8* compounds were truly solitary, not being associated with any other compound. The *"solitary"* compounds were: *T0.9792*, *T0.0492*, *T0.9756*, *T1.0008*, *T0.2126*, *T1.0464*, *T0.6696* and *T0.5784*. As expected, these were usually compounds with low values for the support. The three best rules previously detected had a solitary compound in the premise (*T0.9792, T0.0492 and T0.9756*).

Table 3 Associations of the Potentially Promising Compounds.

Premise	Support	Confidence	Support Premise	Leverage	Number MMFs	Size of the largest MMF
T0.9792	0.03	0.70	0.05	0.02	1	2
T0.0492	0.03	0.67	0.04	0.02	1	2
T0.9756	0.03	0.67	0.04	0.02	1	2
T1.0008	0.03	0.60	0.05	0.01	1	2
T0.2126	0.03	0.54	0.05	0.01	1	2
T1.0464	0.03	0.54	0.05	0.01	1	2
T0.6696	0.03	0.54	0.06	0.01	1	2
T0.5784	0.03	0.54	0.06	0.01	1	2
T1.0896	0.03	0.54	0.05	0.01	1	3
T0.5052	0.03	0.54	0.06	0.01	1	3
T0.0816	0.03	0.64	0.05	0.02	2	3
T0.5928	0.04	0.53	0.07	0.02	2	4
T0.8496	0.05	0.53	0.09	0.02	6	3
T0.8112	0.05	0.52	0.10	0.02	7	3
T0.3540	0.05	0.58	0.09	0.02	11	3
T0.6168	0.05	0.55	0.10	0.02	16	4
T1.0140	0.06	0.52	0.11	0.02	16	5
T0.2580	0.07	0.56	0.12	0.03	23	4
T0.4404	0.07	0.64	0.11	0.03	31	5

To retrieve this information in a more user-friendly way for biologists, we performed an ascending hierarchical classification, choosing as distances the Euclidean distance and the Ward distance. This is the only usefulness of this classification. We did not limit ourselves to these last two items of information (*number of MMF and size of the largest MMF*) but took some information obtained during the extraction, namely: (i) the *support* of the rule (*see Tab. 3 column 2*), (ii) the *confidence* (*see Tab. 3 column 3*), (iii) the *support* of the compound (*see Tab. 3 column 4*), and (iv) the *leverage* of the rule (*see Tab. 3 column 5*). We did not keep the *lift* because it was linearly correlated to the *confidence* when the class association rules had the same conclusion.

Fig. 8 shows the classification obtained. Two categories of compound appear: the first comprises compounds with a low number of MMF (*solitary compounds*), and the second comprises compounds with a large number of MMF (*gregarious compounds*). In each of these two categories, there are two subclasses that are differentiated by the characteristics listed above being somewhat more pronounced in one

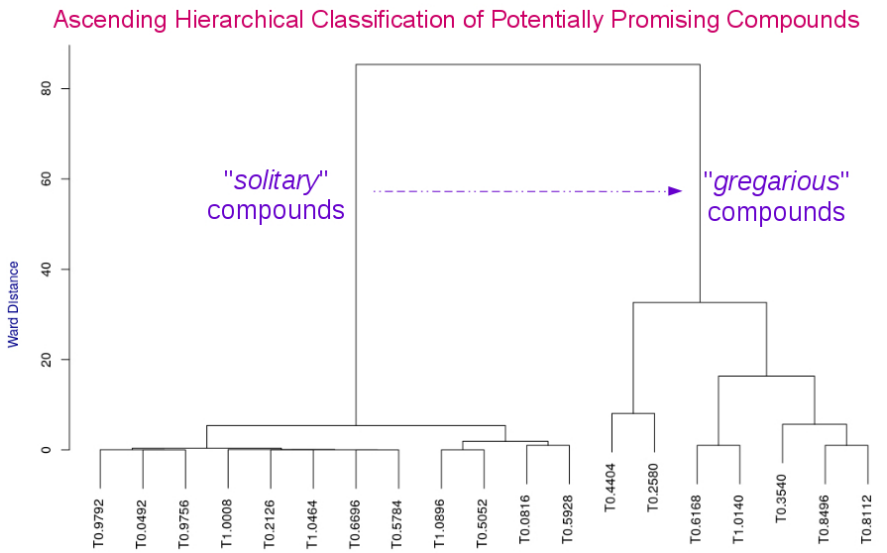

Fig. 8 Ascending Hierarchical Classification of the Potentially Promising Compounds.

of the two subclasses. The further the compounds are on the right, the more MMF they have, and with greater lengths. This classification thus orders the compounds from the most *"solitary"* to the most *"gregarious"*.

We now address the potentially promising associations.

5.2 Mining The Potentially Promising Associations

Tab. 4 displays the 28 best confidence-based potentially promising associations. All these rules had almost the same values for the different interest measures, except for the support. The study of the numerical values taken by these different compounds (*the database* D_1) enabled us to order these rules so as to retain the best ones. This is the subject of Sec. 6.

The graph-based visualization of these *28* best rules is represented in Fig. 9. We can see the limits of such a representation, but it gives us an overview of all the components involved. These are moderately well distributed.

After extracting class rules on the binary data D_3, we compared our results using the numerical data D_1 to take into account the intensity of expression of compounds. This comparison enabled us to select the most promising compounds and associations.

Table 4 The 28 potentially promising associations.

Premise	Support	Confidence	Lift	Count
T0.2724,T0.5928	0.02885	1	3.302	6
T0.7428,T0.8496	0.02885	1	3.302	6
T0.1428,T0.4404	0.02885	1	3.302	6
T0.2196,T0.7512	0.02885	1	3.302	6
T0.3600,T0.7968	0.03365	1	3.302	7
T0.3408,T0.4464	0.02885	1	3.302	6
T0.2268,T1.0980	0.02885	1	3.302	6
T0.2268,T0.7272	0.03365	1	3.302	7
T0.0720,T0.4404,T0.6060	0.02885	1	3.302	6
T0.2172,T0.3060,T0.3168	0.02885	1	3.302	6
T0.3600,T0.7968,T0.8076	0.02885	1	3.302	6
T0.1344,T0.1692,T0.8328	0.02885	1	3.302	6
T0.2100,T0.7968,T0.8076	0.02885	1	3.302	6
T0.1272,T0.1512,T0.2760	0.02885	1	3.302	6
T0.1320,T0.3060,T0.3480	0.02885	1	3.302	6
T0.1512,T0.3252,T0.4140	0.02885	1	3.302	6
T0.1944,T0.2628,T0.4140	0.02885	1	3.302	6
T0.1044,T0.2172,T0.2916	0.02885	1	3.302	6
T0.2172,T0.2484,T0.3060	0.02885	1	3.302	6
T0.2484,T0.3060,T1.0224	0.02885	1	3.302	6
T0.1512,T0.1944,T0.3252	0.02885	1	3.302	6
T0.3228,T0.7092,T0.9024	0.02885	1	3.302	6
T0.3060,T0.3228,T0.7092	0.02885	1	3.302	6
T0.3060,T0.7092,T0.8400	0.02885	1	3.302	6
T0.2556,T0.3060,T0.3480	0.02885	1	3.302	6
T0.1512,T0.3048,T1.0884	0.02885	1	3.302	6
T0.0720,T0.2172,T0.3060	0.02885	1	3.302	6
T0.1512,T0.1944,T0.2628,T0.4140	0.02885	1	3.302	6

6 Mining Class Strongly Expressed Rules

The more abundant a compound is in a plant, the greater will be its peak area. Thus for our purposes, a class rule will be of greater utility the more strongly expressed the phenolic compounds are, i.e. the higher their values are. We name such rules "*class strongly expressed rules*".

First, we describe the discovery of class strongly expressed rules of order *2* to mine the most promising compounds, and then the discovery of class strongly expressed rules of order strictly higher than *2* to mine the most promising compound associations.

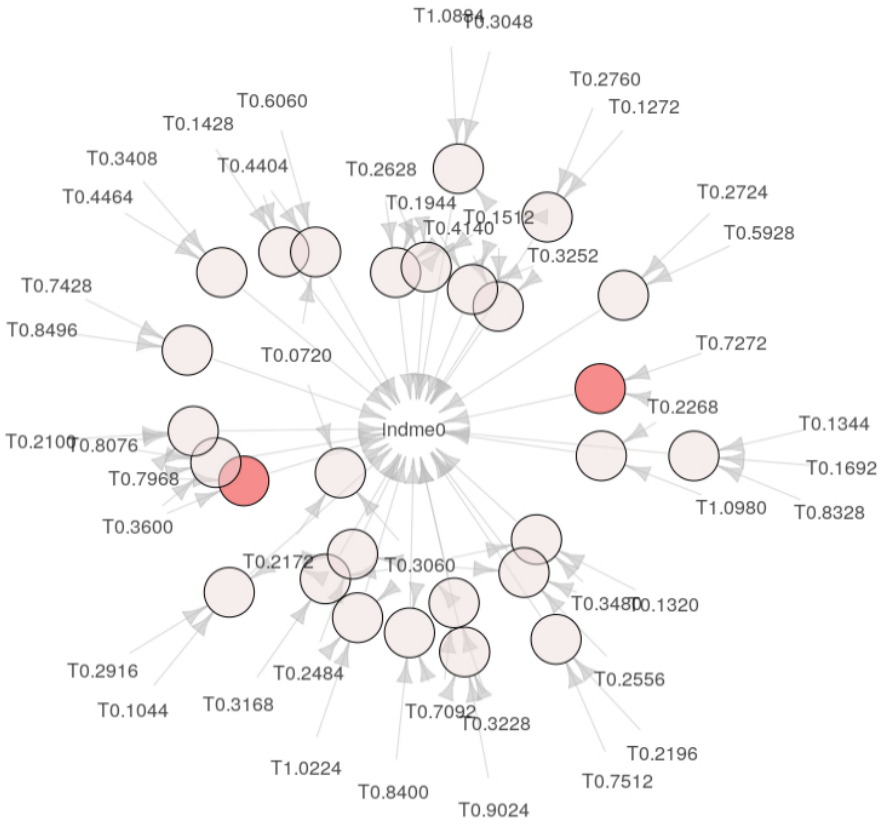

Fig. 9 Graph-based visualization of the 28 best rules.

6.1 Mining the Most Promising Compounds

Rules of *order 2* that were of interest to us were those where the high values taken by the compound T_i were contained in the rule, rules that we could formalize in the following way: $T_i \geq v \Rightarrow indMe0$ with v a value taken by the compound T_i.

To detect this kind of rule, we used the following strategy, which we first explain using an example.

Fig. 10 shows all the positive values taken by compound *T0.6696*, and that by class of substrates, that is to say those for which there was no improvement on methane emissions (*i.e. class 1*) and those for which there was a positive effect (*i.e. class 2: reduction of methane emissions*). We then searched for the optimum value v_{opt} taken by the compound where the proportion of substrates having a positive effect was higher than the proportion of substrates having no effect.

For each value taken by compound T_i (*except the minimum value*), we evaluated the quality of the rule $T_i \geq v \Rightarrow indMe0$ by computing the value of different interest

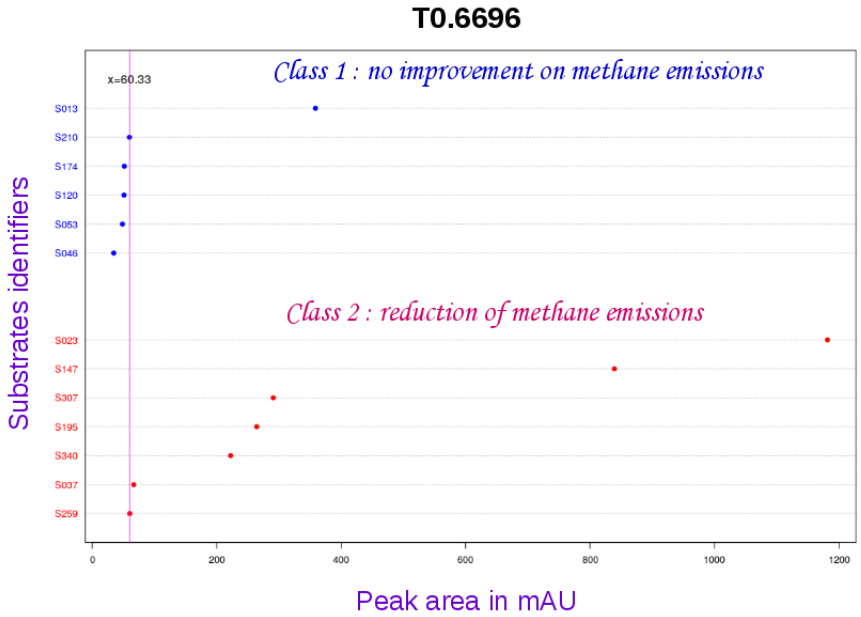

Fig. 10 Distribution of positive values for compound T0.6696 and that by impact class on methane emissions.

measures (*confidence, support and lift*). Since we wanted these rules to verify the user-defined minimum support, we only evaluated a subset of them (*we discarded the* $(min_{sup} \times n)$ *highest values taken by the compound*). To formalize our mining method, we defined the following notations. Let t_i be the number of distinct values taken by the compound T_i and let $\{v_{i1},..,v_{ik},..,v_{it_i}\}$ be the set of ordered values taken by compound ($k \in \{1,..,t_i\}$). Let s be the absolute minimum support defined by the user ($s = min_{sup} \times n$). We are therefore looking for the best rule or rules for a quality measure (*confidence, leverage, etc.*) chosen by the user among all the following rules: $T_i \geq v_{ik} \Rightarrow indMeO$ with $v_{ik} \in \{v_{i2},..,v_{i(t_i-s)}\}$.

To select the promising compounds, we used a new measure, intensity of expression Int_{exp}, which would inform us about the intensity of the rule compared to the mean intensity of the compound. This was the ratio of the mean of the values taken by the numerical rule $T_i \geq v_{ik} \Rightarrow indMe0$, that is to say the mean $mean(T_i \geq v_{ik})$ of the values higher than v_{ik}, to the mean $mean(T_i)$ of values taken by the compound T_i:

$$Int_{exp}(T_i \geq v_{ik} \Rightarrow IndMeO) = \frac{mean(T_i \geq v_{ik})}{mean(T_i)} \tag{3}$$

Thus the more the intensity of expression of the rule is greater than *1*, the better it will be.

Algorithm for mining the most promising compounds

The pseudo-code of the SER algorithm for mining the most promising compounds is presented in *algorithm 1*. We summarize the notations used in Tab. 5.

Table 5 The notations used for the SER algorithm.

Notation	Meaning
n	number of transactions (*substrates*)
DB_1	numerical compounds data
DB_2	target data
R_2	set of valid Rules of order *2*
SER_2	set of Strongly Expressed Rules of order *2*
R_i	set of candidate numerical Rules for compound T_i
min_{sup}	minimum support
$s = min_{sup_{abs}} = min_{sup} \times n$	minimum absolute support
min_{conf}	minimum confidence
$min_{Int_{exp}}$	minimum intensity of expression
t_i	number of values taken by T_i
$\{v_{i1}, .., v_{ik}, .., v_{it_i}\}$	values taken by T_i sorted in ascending order
$s_{ik} = sup(T_i \geq v_{ik} \Rightarrow indMeO)$	support of the rule $T_i \geq v_{ik} \Rightarrow indMeO$
$c_{ik} = conf(T_i \geq v_{ik} \Rightarrow indMeO)$	confidence of the rule $T_i \geq v_{ik} \Rightarrow indMeO$
$Int_{exp_{ik}} = Int_{exp}(T_i \geq v_{ik} \Rightarrow indMeO)$	intensity of expression of the rule $T_i \geq v_{ik} \Rightarrow indMeO$
$s_i = sup(T_i \Rightarrow indMeO)$	support of the rule $T_i \Rightarrow indMeO$
$c_i = conf(T_i \Rightarrow indMeO)$	confidence of the rule $T_i \Rightarrow indMeO$

The algorithm starts by initializing the set SER_2 (*i.e. the set of Strongly Expressed Rules of order 2*) with the empty set (*step 1*). Each compound T_i is then examined successively (*steps 2 to 30*). The set R_i of candidate numerical rules for compound T_i is initialized with the empty set (*step 3*). We search for all the values t_i taken by T_i (*step 4*) and we order them in ascending order (*step 5*). For each value v_{ik} taken by compound T_i (*except the minimum value and the s highest values*), we evaluate the quality of the rule $T_i \geq v_{ik} \Rightarrow indMeO$ (*steps 6 to 21*). We compute the support (*step 7*), the confidence (*step 8*) and the intensity of expression (*step 9*) of the numerical rule $T_i \geq v_{ik} \Rightarrow indMeO$. If the binary rule $T_i \Rightarrow indMeO$ is contained in the set R_2 of valid binary rules of order *2* (*steps 10 to 15*), we search for the support (*step 11*) and the confidence (*step 12*) of the binary rule to verify that it has lower values than the numerical rule (*step 13*). If this is verified and the intensity of expression of the numeric rule is higher than or equal to the minimum intensity of expression (*step 13*), then the numerical rule is inserted into R_i with all the different values of measures (*step 14*). If the binary rule of the compound T_i is not in the set R_2 (*steps 16 to 19*), then we verify that all the values of different measures are higher than the minimum thresholds (*step 17*) and we insert the numerical rule in the set R_i (*step 18*). If the number of candidate numerical rules for the compound T_i is strictly higher than *1*

Algorithm 1 : Mining the most promising compounds with SER

Input : $[DB_1, DB_2]$ (*database*),
R_2 (*set of valid Rules of order 2*),
min_{sup}, min_{conf} and $min_{Int_{exp}}$
Output : SER_2 (*set of Strongly Expressed Rules of order 2*)

1: $SER_2 = \emptyset$ {initialization}
2: **for all** compound T_i **do**
3: $R_i = \emptyset$ {initialization of the set of candidate rules for compound T_i}
4: Search for all the values t_i taken by T_i
5: Sort in ascending order the values taken by $T_i : \{v_{i1}, .., v_{ik}, .., v_{it_i}\}$
6: **for** value $v_{ik} \in \{v_{i2}, .., v_{i(t_i-s)}\}$ **do**
7: Compute $s_{ik} = sup(T_i \geq v_{ik} \Rightarrow indMeO)$
8: Compute $c_{ik} = conf(T_i \geq v_{ik} \Rightarrow indMeO)$
9: Compute $Int_{exp_{ik}} = Int_{exp}(T_i \geq v_{ik} \Rightarrow indMeO)$
10: **if** the rule $(T_i \Rightarrow indMeO) \in R_2$ **then**
11: Search for the support s_i of the rule $(T_i \Rightarrow indMeO)$ in R_2
12: Search for the confidence c_i of the rule $(T_i \Rightarrow indMeO)$ in R_2
13: **if** $(s_{ik} > s_i) \wedge (c_{ik} > c_i) \wedge (Int_{exp_{ik}} \geq min_{Int_{exp}})$ **then**
14: $R_i \leftarrow R_i \cup \{(T_i, v_{ik}, s_{ik}, c_{ik}, Int_{exp_{ik}})\}$
15: **end if** $(s_{ik} > s_i) \wedge (c_{ik} > c_i) \wedge (Int_{exp_{ik}} \geq min_{Int_{exp}})$
16: **else**
17: **if** $(s_{ik} > min_{sup}) \wedge (c_{ik} > min_{conf}) \wedge (Int_{exp_{ik}} \geq min_{Int_{exp}})$ **then**
18: $R_i \leftarrow R_i \cup \{(T_i, v_{ik}, s_{ik}, c_{ik}, Int_{exp_{ik}})\}$
19: **end if** $(s_{ik} > min_{sup}) \wedge (c_{ik} > min_{conf}) \wedge (Int_{exp_{ik}} \geq min_{Int_{exp}})$
20: **end if** $(T_i \Rightarrow indMeO) \in R_2$
21: **end for** $v_{ik} \in \{v_{i2}, .., v_{i(t_i-s)}\}$
{Selection of the best rule for compound T_i}
22: **if** $|R_i| > 1$ **then**
23: sort(R_i, c_{ik} desc, $Int_{exp_{ik}}$ desc, *order* asc)
24: $SER_2 \leftarrow SER_2 \cup \{R_i[1]\}$
25: **else**
26: **if** $|R_i| = 1$ **then**
27: $SER_2 \leftarrow SER_2 \cup \{R_i\}$
28: **end if** $|R_i| = 1$
29: **end if** $|R_i| > 1$
30: **end for** compound T_i
31: **return** SER_2;

(*step 22*), then we sort them first in decreasing order of confidence, then intensity of expression and finally in increasing order of value v_{ik} (*step 23*) and we retain the best one, inserting it into the set SER_2 (*step 24*). If the cardinality of the set R_i is equal to *1* (*step 26*), then we retain the single candidate rule and it is inserted into the set SER_2 (*step 27*). The algorithm finally returns the set SER_2 containing all strongly expressed rules of order *2* (*step 31*).

Here is an example of an extracted rule: $T0.6696 \geq 60.33 \Rightarrow indMeO$ with a value for the confidence equal to *0.875* and a value for the support equal to *0.034*. The value for the confidence of the binary rule $T0.6696 \Rightarrow indMeO$ extracted pre-

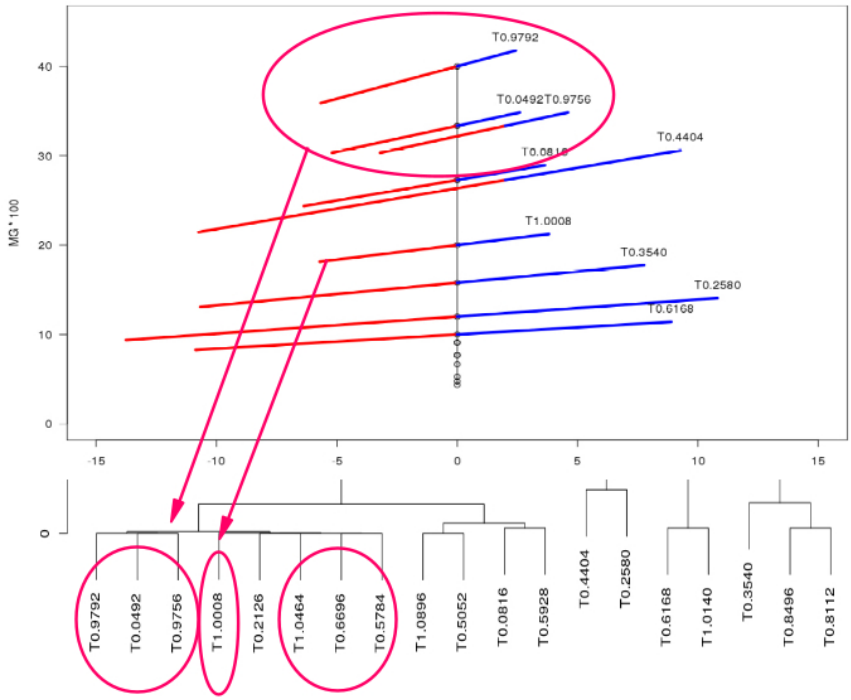

Fig. 11 The best promising compounds.

viously is equal to *0.54* and the value for the support is equal to *0.034*. There is a significant improvement in the confidence when $T0.6696$ is present as a major peak ($> 1000\ mAU$). The intensity of expression of this rule is equal to *1.51*.
$T0.6696 \geq 60.33 \Rightarrow indMeO$ is a promising rule, and therefore $T0.6696$ is a compound to be retained.

At the end of this step, among the *12* solitary phenolic compounds (*see Fig. 8, class 1*), *7* show a threshold effect, which improves the value of the confidence: $T0.9792$, $T0.0492$, $T0.9756$, $T1.0008$, $T1.0464$, $T0.6696$ and $T0.5784$ (*see Fig. 11*).

We then search for the most promising associations by studying the numerical values of the compounds again.

6.2 Mining the Most Promising Compounds Associations

As stated above, we search for the rules where the compounds are strongly expressed. To mine these rules, we propose a new interest measure.

Let $T \Rightarrow indMe0$ be the rule where T is an itemset of compounds T_i. Let sub be a substrate. Let $Sub(T, indMe0)$ be the set of anti-methanogenic substrates where all the compounds T_i have a positive value. This set can be formalized as follows:

$$Sub(T, indMeO) = \{sub \ / \ \forall T_i \subset T \ sub[T_i] > 0 \ \wedge \ sub[indMe0] = 1\} \quad (4)$$

where $sub[T_i]$ is the value taken by the compound T_i for the substrate sub, and $sub[indMe0] = 1$ indicates that sub is an anti-methanogenic substrate.

For all these substrates sub contained in the set $Sub(T, indMe0)$, we compute the mean of the values $sub[T_i]$ of all the compounds T_i contained in T. This is a measure of the mean intensity of rule $T \Rightarrow indMe0$ that we can formalize in the following way:

$$Int(T \Rightarrow indMe0) = \frac{\sum_{sub \in Sub(T,indMe0)} (\sum_{T_i \subset T} sub[T_i])}{|Sub(T, indMe0)| \times |T|} \quad (5)$$

where $|Sub(T, indMe0)|$ is the cardinality of the set $Sub(T, indMe0)$ i.e. the number of substrates, and $|T|$ is the number of compounds T_i contained in T.
We recall that all compounds T_i are expressed in the same unit (*i.e. mAU*).

Let $Int(T^+)$ be the mean of the positive values of all compounds T_i contained in T:

$$Int(T^+) = \frac{\sum_{sub} (\sum_{T_i \subset T} sub[T_i] > 0)}{\sum_{T_i \subset T} sup(T_i)} \quad (6)$$

$Int(T^+)$ is a measure of mean intensity of the positively expressed compounds contained in T.

We propose a new interest measure Int_{exp} which evaluates the strength of expression of the rule $T \Rightarrow indMe0$. This is the ratio of the mean intensity of the rule $Int(T \Rightarrow indMe0)$ to the positive mean intensity $Int(T^+)$ of compounds involved in the rule. This measure can be formalized as follows:

$$Int_{exp}(T \Rightarrow indMe0) = \frac{Int(T \Rightarrow indMe0)}{Int(T^+)} \quad (7)$$

Only rules with an intensity of expression higher than $min_{Int_{exp}}$ will be considered of interest, and therefore retained.
We now present the algorithm SER^+ that mines all class strongly expressed rules

Algorithm 2 : Mining the most promising compound associations with SER^+

Input : $[DB_2, DB_3]$ (*database*),
AR_{2^+} (*set of class Association Rules of order higher than 2*),
$min_{Int_{exp}}$ (*minimum intensity of expression*) and
max_{rules} (*maximum number of rules*)
Output : SER_{2^+} (*set of Strongly Expressed Rules of order higher than 2*)

1: $SER_{2^+} = \emptyset$ {initialisation}
2: **for all** rules $R \in AR_{2^+}$ **do**
3: Compute $Int_{exp}(R)$
4: **if** $Int_{exp}(R) > min_{Int_{exp}}$ **then**
5: $SER_{2^+} \leftarrow SER_{2^+} \cup \{R\}$
6: **end if** $Int_{exp}(R) > min_{Int_{exp}}$
7: **end for** rules $R \in AR_{2^+}$
{selecting the best rules}
8: **if** $|SER_{2^+}| > max_{rules}$ **then**
9: $sort(SER_{2^+}, Int_{exp}(R)desc)$
10: $SER_{2^+} \leftarrow SER_{2^+}[1..max_{rules}]$
11: **end if** $|SER_{2^+}| > max_{rules}$
12: **return** SER_{2^+};

of order higher than *2*.
The pseudo-code of the SER^+ algorithm for mining the most promising compound associations is presented in *algorithm 2*.

The algorithm starts by initializing the set SER_{2^+} (*set of Strongly Expressed Rules of order higher than 2*) with the empty set (*step 1*). For each class association rule R of order higher than *2*, we evaluate the quality of the rule $T \Rightarrow indMeO$ and more specifically, we measure the presence of high values (*steps 2 to 7*). We compute the intensity of expressed $Int_{exp}(R)$ of the rule (*step 3*). If this intensity is higher than the minimum intensity of expression $min_{Int_{exp}}$ (*step 4*) then the rule is retained and inserted into the set SER_{2^+} (*step 5*). If the number of strongly expressed rules is higher than the user-defined maximum number of rules (*steps 8 to 11*), then the max_{rules} best are selected (*step 10*) according to the intensity of expression $Int_{exp}(R)$ (*step 9*). The algorithm finally returns the set SER_{2^+} containing all strongly expressed rules of order higher than *2* (*step 12*).

The 5 best rules of order *2* extracted are shown in Tab. 6.

Tab. 7 displays the values of the intensity of expression for the best association rules of order *2*.

At the end of this extraction, we retained the following associations (*see Tab. 8*). We chose *1* as the minimum value for the intensity of expression. We selected associations of order *2* because they involved fewer compounds.

Table 6 The 5 best rules of order 3.

Premise	$Int_{exp}(R)$	$Int(T \Rightarrow indMe0)$	$Int(T^+)$	Confidence
T0.1248,T0.2232	5.3	5,348.28	1,008.78	0.75
T0.1632,T0.2232	3.72	4,657.17	1,253.09	0.57
T0.1512,T0.2004	3.5	2,671.62	764.19	0.60
T0.3060,T0.3252	3.44	1,139.17	331.46	0.60
T0.1368,T0.3924	3.35	1,697.89	507.03	0.75

Table 7 Values of the intensity of expression for the best association rules of order *3*.

Premise	$Int_{exp}(R)$	$Int(T \Rightarrow indMe0)$	$Int(T^+)$	Confidence
T0.2724,T0.5928	1.24	493.13	396.28	1
T0.7428,T0.8496	0.72	142.31	197.84	1
T0.1428,T0.4404	0.55	217.62	398.56	1
T0.2196,T0.7512	0.3	175.81	579.18	1
T0.3600,T0.7968	0.49	191.25	389.27	1
T0.3408,T0.4464	0.85	870.96	1,023.25	1
T0.2268,T1.0980	1.08	251.45	231.8	1
T0.2268,T0.7272	1.17	428.3	364.95	1

Table 8 The associations selected.

Premise	$Int_{exp}(R)$	$Int(T \Rightarrow indMe0)$	$Int(T^+)$	Confidence
T0.1248,T0.2232	5.3	5,348.28	1,008.78	0.75
T0.1368,T0.3924	3.35	1,697.89	507.03	0.75
T0.2724,T0.5928	1.24	493.13	396.28	1
T0.2268,T0.7272	1.17	428.3	364.95	1
T0.2268,T1.0980	1.08	251.45	231.8	1

7 Conclusion

From the *1,047* unidentified phenolic compounds contained in our database, the discovery of class-association rules generated *676* valid rules, representing far too many compounds for biologists and chemists to identify. The association rules of order *2* allow us to extract the potentially promising compounds. These were classified into two categories using a hierarchical ascending classification: *12* "*solitary*" compounds and *7* "*gregarious*" compounds. Taking into account the intensity of expression of these rules and the new visualization proposed, *7* compounds were finally selected, all belonging to the category of "*solitary*" compounds. The association rules of higher order highlighted *28* potentially promising associations with a confidence equal to *1*. The extraction of the strongly expressed rules enabled us to select *5* promising associations involving *9* new compounds. This comparison of the two kinds of extraction retained only *3* rules out of the *28* rules with a confidence equal to *1* and revealed *2* new promising associations. Examination of the ultraviolet

spectra of the solitary compounds already shows that they belong to (i) the family of cinnamic acids, and (ii) the family of flavonols. They remain to be identified precisely, so that the pure products can be synthesized to determine whether they have true anti-methanogenic effects in fermentation tests. Encouragingly, previous results [Macheboeuf et al., 2008] have shown that cinnamaldehyde has a significant effect on methane release.

Acknowledgements Data used in the KDD process presented in this paper was obtained with the support of Phytosynthèse (*57, avenue Jean Jaurès, 63200 Mozac, FRANCE*) and of the Carnot Institute France Futur Breeding. We thank Guy Lalière (*ethnobotanist 6, rue des Plats, 63000 Clermont-Ferrand, FRANCE*), Arnaud Descheemacker and Romain Pradinas of the Conservatoire Botanique National du Massif Central (*Le Bourg, 43230 Chavaniac-Lafayette, FRANCE*) for their help in plant identification and collection. Special thanks to Agnès Cornu and Clara Leguay (*INRA - UMR Herbivores, 63122 St-Genès-Champanelle, FRANCE*) for their participations in the analysis of plants and to Roger Bergeault for his help during the fermentation experiments.

References

[Aggarwal and Yu, 1998] Aggarwal, C. and Yu, P. (1998). Online generation of association rules. In *Proceedings of the International Conference on Data Engineering*, pages 402–411.

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings SIGMOD conference*, pages 207–216.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th Very Large Data Bases Conference*, pages 487–499.

[Bastide et al., 2000] Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *CL'2000 international conference on Computational Logic*, pages 972–986.

[Berasaluce et al., 2004] Berasaluce, S., Laureno, C., Napoli, A., and Niel, G. (2004). An experiment on knowledge discovery in chemical databases. In Boulicaut, J.-F., Esposito, F., Giannotti, F., and Pedreschi, D., editors, *PKDD, Volume 3202 of Lecture Notes in Computer Science*, pages 39–51.

[Borgelt, 2003] Borgelt, C. (2003). Efficient implementations of apriori and eclat. In *FIMI'03: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, page 90.

[Borgelt and Kruse, 2002] Borgelt, C. and Kruse, R. (2002). Induction of association rules: Apriori implementation. In Physika Verlag, editor, *Proceedings 15th Conference on Computational Statistics*, pages 395–400.

[Brin et al., 1997] Brin, S., Motwani, R., Ullman, J., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264.

[Choi et al., 2005] Choi, D., Ahn, B., and Kim, S. (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. In *Expert Systems with Applications*, volume 29, pages 867–878.

[Dangal et al., 2017] Dangal, S. R. S., Tian, H., Zhang, B., Pan, S., Lu, C., and Yang, J. (2017). Methane emission from global livestock sector during 1890 - 2014: Magnitude, trends and spatiotemporal patterns. In *Global Change Biology*, volume 23, pages 4147–4161.

[Djenouri et al., 2018] Djenouri, Y., Belhadi, A., Fournier-Viger, P., and Chun-Wei Lin, J. (2018). Discovering strong meta association rules using bees swarm optimization. In Ganji M., Rashidi L., Fung B., Wang C., editor, *Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Revised Selected Papers*, volume 11154, pages 195–206.

[Djenouri et al., 2014] Djenouri, Y., Drias, H., and Bendjoudi, A. (2014). Pruning irrelevant association rules using knowledge mining. In Ganji M., Rashidi L., Fung B., Wang C., editor, *International Journal of Business Intelligence and Data Mining*, volume 9, pages 112–144.

[Duin et al., 2016] Duin, E. C., Wagner, T., Shima, S., Prakash, D., Cronin, B., Yáñez-Ruiz, D. R., Duval, S., Rümbeli, R., Stemmler, R. T., Thauer, R. K., and Kindermann, M. (2016). Mode of action uncovered for the specific reduction of methane emissions from ruminants by the small molecule 3-nitrooxypropanol. In National Academy of Sciences, editor, *Proceedings of the National Academy of Sciences*, volume 113, pages 6172–6177.

[Fayyad, U.M. et al., 1996] Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In AAAI Press, editor, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, page 83.

[Fischer and Meinl, 2004] Fischer, I. and Meinl, T. (2004). Graph based molecular data mining - an overview. In Wil Thissen, M., Wieringa, P., and Ludema, M., editors, *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics: The Hague, Netherlands*, pages 4578–4582.

[Freitas, 2000] Freitas, A. A. (2000). Understanding the crucial differences between classification and discovery of association rules - a position paper. In *ACM SIGKDD Explorations Newsletter*, volume 2, pages 65–69.

[Garcia et al., 2000] Garcia, J. L., Patel, B. K., and Ollivier, B. (2000). Taxonomic, phylogenetic, and ecological diversity of methanogenic archaea. In *Anaerobe*, volume 6, pages 205–226.

[Guillaume, 2010] Guillaume, S. (2010). Améliorations de la mesure d'intérêt m_{GK}. In *Actes des XVIIèmes rencontres de la Société Francophone de Classification*, pages 41–45.

[Hahsler, 2015] Hahsler, M. (2015). A probabilistic comparison of commonly used interest measures for association rules. http://michael.hahsler.net/research/association_rules/measures.html.

[Hahsler, 2016] Hahsler, M. (2016). Grouping association rules using lift. In C. Iyigun and R. Moghaddess and A. Oztekin, editor, *11th INFORMS Workshop on Data Mining and Decision Analytics (DM-DA 2016)*.

[Hahsler, 2017] Hahsler, M. (2017). arulesviz: Visualizing association rules with r. In *R Journal, 9(2)*, pages 163–175.

[Hristov et al., 2013] Hristov, A. N., Oh, J., Firkins, J. L., Dijkstra, J., Kebreab, E., Waghorn, G., Makkar, H. P. S., Adesogan, A. T., Yang, W., Lee, C., Gerber, P. J., Henderson, B., and Tricarico, J. M. (2013). SPECIAL TOPICS ? Mitigation of methane and nitrous oxide emissions from animal operations: I. A review of enteric methane mitigation options1. In *Journal of Animal Science*, volume 91, pages 5045–5069.

[Huws et al., 2018] Huws, S. A., Creevey, C. J., Oyama, L. B., Mizrahi, I., Denman, S. E., Popova, M., Muoz-Tamayo, R., Forano, E., Waters, S. M., Hess, M., Tapio, I., Smidt, H., Krizsan, S. J., Yez-Ruiz, D. R., Belanche, A., Guan, L., Gruninger, R. J., McAllister, T. A., Newbold, C. J., Roehe, R., Dewhurst, R. J., Snelling, T. J., Watson, M., Suen, G., Hart, E. H., Kingston-Smith, A. H., Scollan, N. D., do Prado, R. M., Pilau, E. J., Mantovani, H. C., Attwood, G. T., Edwards, J. E., McEwan, N. R., Morrisson, S., Mayorga, O. L., Elliott, C., and Morgavi, D. P. (2018). Addressing global ruminant agricultural challenges through understanding the rumen microbiome: Past, present, and future. In *frontiers in Microbiology*, volume 9, pages 1–33.

[Klemettinen et al., 1994] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. (1994). Finding interesting rules from large sets of discovered association rules. In *Proceedings CIKM conference*, pages 401–407.

[Knapp et al., 2014] Knapp, J., Laur, G., Vadas, P., Weiss, W., and Tricarico, J. (2014). Invited review: Enteric methane in dairy cattle production: Quantifying the opportunities and impact of reducing emissions. In *Journal of Dairy Science*, volume 97, pages 3231–3261.

[Loh, 2011] Loh, W.-Y. (2011). Classification and regression trees. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, volume 1, pages 14–23.

[Lynch, 2019] Lynch, J. (2019). Availability of disaggregated greenhouse gas emissions from beef cattle production: A systematic review. In *Environmental Impact Assessment Review*, volume 76, pages 69–78.

[Macheboeuf et al., 2018] Macheboeuf, D., Cornu, A., Kerros, S., and Recoquillay, F. (2018). An antimethanogenic index for meadow plants consumed by ruminants. In Cambridge University Press, Animal Biosciences, editor, *Herbivore nutrition supporting sustainable intensification and agro-ecological approaches. Proceedings of the 10 th International Symposium on the Nutrition of Herbivores ISNH 2018, Clermont-Ferrand, FRA (2018-09-02 - 2018-09-06)*, page 608.

[Macheboeuf et al., 2008] Macheboeuf, D., Morgavi, D., Papon, Y., Mousset, J.-L., and Arturo-Schaan, M. (2008). Dose-response effects of essential oils on *in vitro* fermentation activity of the rumen microbial population. In Elsevier, editor, *Animal Feed Science and Technology, Volume 145, Issuesin vitro 1-4*, pages 335–350.

[Mizrahi and Jami, 2018] Mizrahi, I. and Jami, E. (2018). Review: The compositional variation of the rumen microbiome and its effect on host performance and methane emission. In *animal*, volume 12, pages s220–s232.

[Ordonez, 2006] Ordonez, C. (2006). Comparing association rules and decision trees for disease prediction. In Li Xiong and Yuni Xia, ACM, editor, *Proceedings of the International Workshop on Healthcare Information and Knowledge Management, HIKM 2006, Arlington, Virginia, USA, November 11, 2006*, pages 17–24.

[Pasquier et al., 1999a] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999a). Discovering frequent closed itemsets for association rules. In *Proceedings of the ICDT International Conference on Database Theory*, pages 398–416.

[Pasquier et al., 1999b] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999b). Efficient mining of association rules using closed itemset lattices. In *Information Systems*, pages 25–46.

[Patra et al., 2017] Patra, A., Park, T., Kim, M., and Yu, Z. (2017). Rumen methanogens and mitigation of methane emission by anti-methanogenic compounds and substances. In *Journal of Animal Science and Biotechnology*, volume 8, page 13.

[Patra and Saxena, 2010] Patra, A. and Saxena, J. (2010). A new perspective on the use of plant secondary metabolites to inhibit methanogenesis in the rumen. In *Phytochemistry*, pages 1198–1222.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro, G. and Frawley, W., editors, *Knowledge Discovery in Databases*, pages 229–248.

[R Development Core Team, 2020] R Development Core Team (2020). R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing.

[Sakakibara et al., 2003] Sakakibara, H., Honda, Y., Nakagawa, S., Ashida, H., and Kanazawa, K. (2003). Simultaneous determination of all polyphenols in vegetables, fruits, and teas. In *Journal of Agricultural and Food Chemistry*, pages 571–581.

[Seno and Karypis, 2005] Seno, R. and Karypis, G. (2005). Finding frequent itemsets using length-decreasing support constraint. In *Data Mining and Knowledge Discovery*, pages 197–228.

[Shukla et al., 2019] Shukla, S., Mohanty, B., , and Kumar, A. (2019). A fuzzy approach to prioritise dea ranked association rules. In *International Journal of Business Intelligence and Data Mining*, volume 14, pages 155–176.

[Unwin et al., 2001] Unwin, A., Hofmann, H., and Bernt, K. (2001). The twokey plot for multiple association rules control. In Springer-Verlag, editor, *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 472–483.

[Vougas et al., 2019] Vougas, K., Sakellaropoulos, T., Kotsinas, A., Foukas, G.-R. P., Ntargaras, A., Koinis, F., Polyzos, A., Myrianthopoulos, V., Zhou, H., Narang, S., Georgoulias, V., Alexopoulos, L., Aifantis, I., Townsend, P. A., Sfikakis, P., Fitzgerald, R., Thanos, D., Bartek, J., Petty, R., Tsirigos, A., and Gorgoulis, V. G. (2019). Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining. In *Pharmacology & Therapeutics*, volume 203, page 107395.

[Wu et al., 2018] Wu, D.-Y., Zhang, X.-Y., and Zhou, X.-L. (2018). Mining and correlation analysis of association rules between properties and therapeutic efficacy of chinese materia medica based on strategy pattern. In *Chinese medical journal*, volume 131, pages 2755–2757.

[Zaki, 2000] Zaki, M. (2000). Generating non-redundant association rules. In *Proceedings of the ACM SIGKDD*, pages 34–43.

[Zaki, 2004] Zaki, M. (2004). Mining non-redundant association rules. In Kluwer Academic, editor, *Data Mining and Knowledge Discovery, 9*, pages 223–248.