# Extraction de composés phénoliques végétaux susceptibles de limiter les émissions de méthane chez les ruminants

Sylvie Guillaume, Didier Macheboeuf

# Extraction de composés phénoliques végétaux susceptibles de limiter les émissions de méthane chez les ruminants

Sylvie-Guillaume*, Didier Macheboeuf**

*CNRS, UMR 6158, LIMOS, Université Clermont Auvergne, F-63173 Aubière, France
sylvie.guillaume@uca.fr
**Université Clermont Auvergne, INRA, VetAgro Sup, UMR Herbivores,
F-63122 Saint-Genès-Champanelle, France
Didier.Macheboeuf@inra.fr

**Résumé.** Le méthane est un puissant gaz à effet de serre. En Europe, les émissions de méthane proviennent principalement de l'élevage, et en particulier, des ruminants qui hébergent dans leur panse un écosystème microbien qui fermente la matière végétale. L'objectif de cet article est de rechercher les composés phénoliques des plantes qui pourraient avoir une action sur ces microbes et limiter la production de méthane, afin de proposer des alternatives naturelles d'alimentation. Comme on trouve chez ces composés une très grande diversité de structures chimiques, il n'est pas possible de les tester tous. Ainsi nous avons eu recours à la fouille de données, et plus particulièrement aux règles d'association de classe, pour faire émerger des composés susceptibles d'avoir un effet significatif. Par ailleurs, la nouvelle représentation graphique des règles générales et une nouvelle mesure d'intensité qui sont proposées, ont permis d'affiner la pertinence des règles et de réduire la sélection à quelques composés qu'il sera alors possible d'identifier. Ainsi, parmi les *1 075* composés inconnus présents dans le jeu de plantes, *26* ont émergé desquels *11* sont de vrais solitaires dont *7* ont un effet seuil.

## 1 Introduction

Le méthane est le deuxième gaz à effet de serre après le dioxyde de carbone mais il a un pouvoir de réchauffement global *23* fois supérieur. En Europe, la quasi-totalité des émissions de méthane sont d'origine agricole parmi lesquelles deux tiers proviennent de l'élevage des ruminants. Dans le cadre de la mise en place de systèmes d'élevages durables respectueux de l'environnement, des recherches sont menées depuis une vingtaine d'années pour réduire le méthane produit par les ruminants. Elles répondent bien sûr à un objectif environnemental mais aussi à des objectifs économiques et de productivité. En effet, le méthane émis dans l'atmosphère peut représenter dans certains cas, jusqu'à *10 %* de l'énergie de la ration ingérée et constitue de fait un manque d'efficacité digestive pour l'animal et une perte économique pour l'éleveur.

Pendant la digestion, la matière végétale est dégradée par l'écosystème microbien du rumen [1]. Le processus de fermentation aboutit, d'une part, à des acides gras volatils (*AGV*) qui sont les nutriments absorbés par l'animal, et d'autre part, à des gaz de fermentation (*CO2 et CH4*), libérés dans l'atmosphère par éructation. Le méthane est produit par les *Archaea*, un type de micro-organisme qui a la particularité de posséder les voies métaboliques de la méthanogénèse. Pour manipuler les fermentations ruminales et limiter l'action des *Archaea*, les recherches se dirigent vers les métabolites secondaires des plantes, car l'utilisation des antibiotiques et tous les additifs de synthèse dans l'alimentation animale, sont interdits dans l'Union Européenne depuis 2006 (regulation 1831/2003/EC).

Les métabolites secondaires sont en effet des composés chimiques qui ne sont pas essentiels à la constitution des plantes mais que celles-ci peuvent produire dans certains cas en réponse à des stress (*hydrique par exemple*) ou à des agressions (*insectes, micro-organismes, herbivores*) ou lors de leur reproduction. Il serait donc intéressant de connaître l'effet de ces composés sur la production de méthane de l'écosystème ruminal. Cependant la très grande diversité des structures chimiques, estimée à plus de *200 000*, ne permet pas de tester l'activité de tous ces composés.

Ainsi, notre stratégie a été d'effectuer un essai de criblage [2] par fermentations *in vitro* sur *208* plantes et d'identifier parmi celles-ci, les plantes bio-actives contre le méthane. Parallèlement, et pour chacune de ces plantes, le profil en composés phénoliques de petite taille a été déterminé. Nous nous sommes intéressés pour l'instant aux composés phénoliques de petite taille qui ne représentent qu'une petite partie des métabolites secondaires. L'analyse par *HPLC-DAD* [3] donnait un profil qui confirmait la présence d'un composé par un pic. Celui-ci était alors caractérisé par son temps de rétention sur la colonne et son spectre dans l'ultra violet. A ce stade, les composés n'étaient pas identifiés, compte-tenu qu'il y avait en moyenne plus d'une centaine de pics par plante. La priorité était la sélection de quelques composés susceptibles d'être responsables ou de participer à l'effet observé sur le méthane.

Compte-tenu de la très grande fluctuation des profils en termes d'importance relative des composés et de leur faible taux de présence dans les plantes, il n'était pas possible d'établir des corrélations entre les composés et l'effet observé par les méthodes classiques de l'analyse statistique. Nous avons donc eu recours à la fouille de données, et plus particulièrement aux règles d'association de classe, pour faire émerger des composés susceptibles d'avoir un effet positif. Dans cet article, nous nous focalisons sur la recherche des composés actifs et non sur les synergies possibles entre-eux. Ces dernières seront traitées ultérieurement. Il est impératif de sélectionner tout d'abord, un nombre raisonnable de composés ayant une forte chance d'être impliqué dans le pouvoir anti-méthanogène de la plante, car les phases suivantes qui sont, d'une part, l'identification des composés et, d'autre part, la vérification *in vitro* de l'effet escompté, sont onéreuses en temps et en coût.

L'article s'organise donc de la façon suivante. La *section 2* présente les données et la façon dont elles ont été acquises. Le reste de l'article se consacre à la procédure d'extraction des composés prometteurs (*sections 3 et 4*). Ainsi, dans un premier temps, nous utilisons la tech-

1. Le rumen est le plus important des compartiments digestifs des ruminants de par sa fonction et son volume (*100 litres pour un animal de 600 kg*). Il héberge en symbiose un écosystème microbien dense et complexe qui lui confère la possibilité de dégrader les parois végétales par fermentation.
2. Technique qui vise à tester un grand nombre d'individus, ici des plantes, et à faire ressortir celles qui répondent au critère souhaité.
3. Chromatographie liquide haute pression avec détecteur à barrette de diodes.

nique des règles d'association de classe (*technique rappelée brièvement dans la section 3.1*) pour extraire les composés potentiellement prometteurs (*section 3.2*). A cette occasion, nous présentons une nouvelle représentation des règles. Dans un second temps, nous sélectionnons parmi ces composés extraits précédemment, ceux qui sont les plus prometteurs en étudiant leur intensité d'expression grâce à une technique de discrétisation contextuelle des règles que nous présentons dans la *section 4*. Nous concluons ce travail en indiquant les familles de molécules prometteuses qui ont été mises en évidence et en exposant la suite de ce travail, à savoir l'étude des synergies des composés.

# 2 Présentation des données

Les substrats qui ont été utilisés pour les fermentations *in vitro* et pour la détermination des profils en composés phénoliques, ont été obtenus à partir de *208* espèces de plantes qui ont été récoltées dans le Massif Central, congelées dans l'azote liquide pour figer les métabolites secondaires, puis lyophilisées et broyées.

## 2.1 Les variables prémisses

Les composés phénoliques sont extraits des substrats par un traitement éthanol : eau puis séparés avec une chaîne *HPLC* pendant *95* minutes sur une colonne *C18* avec un gradient eau : méthanol selon une méthode adaptée de (Sakakibara et al., 2003). L'intensité du signal est mesurée avec un détecteur à barrette de diodes de *200* à *400* nm. Le profil chromatographique de chaque plante est enregistré à *280* et à *320* nm. Un mélange de *21* composés phénoliques connus et utilisés comme standards, est systématiquement injecté dans les séquences d'analyses. Parmi ces standards, la flavone est utilisée comme référence pour le calcul des temps de rétention relatifs ($T_i$) des pics. L'alignement des séquences est réalisé en repositionnant les $T_i$ des standards. Ainsi, le temps de rétention de la flavone étant en moyenne de *84.52* min, le domaine de variation des $T_i$ se trouve dans l'intervalle $[0, 1.124]$ pour une séparation sur *95* min. Par conséquent, les composés des plantes étant inconnus, ils sont identifiés (*noms des variables prémisses*) par leur temps de rétention relatif $T_i$, $i \in [0, 1.124]$.

Il a été détecté au total dans le jeu de *208* plantes, *1 075* composés différents (*soit autant de variables prémisses*). Le nombre de composés différents trouvés en moyenne par plante est de *106* avec une étendue allant de *29* à *161*. Le nombre d'occurrences d'un composé dans le jeu de plantes est très variable allant d'une unique apparition à une fréquence d'apparition de près de *58%*. En moyenne, la fréquence d'apparition est de *10%*. Les données des variables prémisses sont les aires des pics si le composé est présent. Là aussi, les variations sont extrêmement importantes allant de *3 mAU* (*seuil de détection du système d'analyse*) à *105 200 mAU*[4]. L'aire du pic médian est de *100 mAU*. La valeur de *10* fois l'aire du pic médian est choisie pour distinguer les pics mineurs (*<1 000 mAU*) des pics majeurs. Le nombre de pics majeurs est en moyenne de *9* par plante avec une étendue de *0* à *43*.

Les données brutes sont donc structurées en une matrice *208* (*plantes*) x *1 075* (*composés*) à *280 nm* contenant les valeurs numériques des aires des pics. Cette matrice a un taux de remplissage faible de *10%*. Les composés omniprésents dans les plantes (*fréquence > 30%*) ont

4. mAU : milli-arbitrary unit, unité d'aire arbitraire utilisée lors de l'intégration du chromatogramme

été retirés du jeux de données pour éviter le risque de faux-positifs, c'est-à-dire *28* composés. Les données ont ensuite été binarisées avant la fouille comme indiqué dans la partie suivante.

### 2.2 La variable cible

La particularité de ce travail de fouille de données est qu'il ne comporte qu'une seule variable cible : le méthane. La production de méthane de l'écosystème microbien a été mesurée pour les *208* substrats après *24h* de fermentation dans des systèmes *in vitro* simulant le rumen. Toutes les fermentations ont été répétées *3* fois. Les productions ont été rapportées à la production de méthane obtenue pour le substrat témoin qui est le *Ray Grass Anglais*. La production de méthane est donc un vecteur colonne de dimension *208* sans aucune données manquantes, dont les valeurs (*moyenne de 3 répétitions*) sont des ratios compris entre *0,10* et *1,33*. Cette variable a été transformée pour calculer un index anti-méthanogène. L'index est calculé comme la déviation entre la production de méthane d'une plante qui a été mesurée et la valeur qui aurait été obtenue dans les conditions d'une fermentation normale (*sans inhibition de la méthanogénèse*) à production d'*AGV* équivalente, divisée par la plus grande des déviations observées (Macheboeuf et al., 2018). L'index varie de *-0,74* à *1,00* avec une moyenne de *-0,09*. Toute plante qui obtient un index supérieur strictement à *0* a un effet anti-méthanogène très significatif ($p$<*0,01*). L'index est converti en données binaires et nommé $indMeO$. L'effet anti-méthanogène est présent (*index >0*) chez *64* plantes, soit environ *30%* de l'effectif, pour lesquelles $indMeO$ a pris la valeur *1*.

Après avoir exposé les données, nous effectuons l'extraction des composés phénoliques susceptibles d'avoir un effet positif, en utilisant tout d'abord la technique des règles d'association de classe. Nous les appellerons les composés potentiellement prometteurs.

## 3 Extraction des composés potentiellement prometteurs

Avant d'exposer les résultats de cette extraction, nous faisons un bref rappel de la technique d'extraction des règles d'association, les règles de classe étant un cas particulier de celles-ci.

### 3.1 Extraction des règles d'association de classe

L'extraction des règles d'association (Agrawal et Srikant, 1994) consiste à découvrir des corrélations entre les attributs (*ou variables*) d'une base de données $BD$ composée de $n$ individus. Une règle d'association est une implication de la forme $X \Rightarrow Y$, où $X$ (*prémisse ou antécédent*) et $Y$ (*conclusion ou conséquent*) sont deux ensembles ($X = \{x_1, .., x_i, .., x_p\}$ *et* $Y = \{y_1, .., y_j, .., y_q\}$) disjoints d'items ($X \cap Y = \varnothing$). Un item $x_i$ ou $y_j$ avec $(i \in \{1, .., p\})$ et $(j \in \{1, .., q\})$ est une variable binaire de la base de données et nous parlons de motif lorsque nous sommes en présence d'un ensemble d'items ; $X = \{x_1, .., x_i, .., x_p\}$ et $Y = \{y_1, .., y_j, .., y_q\}$ sont donc deux motifs. La règle $X \Rightarrow Y$ signifie que les individus qui vérifient tous les items (*ou caractéristiques*) $x_i$ $(i \in \{1, .., p\})$ de la prémisse $X$ vérifient également en général tous les items $y_j$ $(j \in \{1, .., q\})$ de la conclusion $Y$. Par exemple, $crêpes, beurre \Rightarrow cidre$ est une règle d'association révélant que lorsqu'un consommateur achète à la fois des *crêpes* et du *beurre*, alors il achète également en général du *cidre*. Afin de quantifier l'intérêt d'une règle $X \Rightarrow Y$, on utilise en général deux mesures d'intérêt

objectives : le support et la confiance dont nous rappelons la définition.
Le support de la règle $X \Rightarrow Y$ est le support du motif $X \cup Y = \{x_1, .., x_i, .., x_p, \ y_1, .., y_j, .., y_q\}$ c'est-à-dire le pourcentage d'individus qui vérifient à la fois tous les items de $X$ et tous les items de $Y$, ou encore la probabilité d'apparition de $X \cup Y$ : $sup(X \Rightarrow Y) = sup(X \cup Y) = P(X \cup Y) = \frac{n_{XY}}{n}$ avec $n_{XY}$ le nombre d'individus vérifiant le motif $X \cup Y$. Pour simplifier les notations, nous noterons l'ensemble $X \cup Y$ par $XY$.
Le support permet d'évaluer la portée de la règle en révélant la proportion d'individus de la base de données concernée par cette règle. On parle de support absolu dans le cas où nous nous intéressons au nombre d'individus vérifiant la règle et non à la fréquence : $sup_{abs}(X \Rightarrow Y) = n_{XY} = n \times sup(X \Rightarrow Y)$
La confiance de la règle $X \Rightarrow Y$ est le pourcentage d'individus qui vérifient tous les items de $Y$ parmi ceux qui vérifient tous les items de $X$. C'est également la probabilité conditionnelle de $Y$ sachant $X$ et par conséquent, c'est le rapport entre le support de $XY$ et le support de $X$ : $conf(X \Rightarrow Y) = P(Y/X) = \frac{sup(XY)}{sup(X)} = \frac{n_{XY}}{n_X}$.
La confiance permet d'évaluer la force de la règle.
Afin de retenir les règles les plus intéressantes, l'utilisateur fixe deux seuils minimaux $min_{sup}$ et $min_{conf}$ pour respectivement le support et la confiance. Les règles vérifiant ces deux contraintes $(Ct_1) : sup(X \Rightarrow Y) \geqslant min_{sup}$ et $(Ct_2) : conf(X \Rightarrow Y) \geqslant min_{conf}$ seront celles qui seront retenues et restituées par l'algorithme d'extraction. Les règles vérifiant ces deux contraintes sont appelées règles valides.
La confiance est une mesure très intéressante puisqu'elle est simple à comprendre pour les utilisateurs et qu'elle possède une propriété d'anti-monotonicité évitant de parcourir tout l'espace de recherche des règles. Cependant, elle peut restituer des règles non pertinentes, et pour pallier ce problème, on a recours classiquement à deux autres mesures : le lift (Brin et al., 1997) et le leverage (Piatetsky-Shapiro, 1991). Ces deux mesures évaluent l'écart de la règle à l'indépendance, par un rapport pour la première mesure et avec une différence pour la seconde. Nous rappelons leurs définitions : $lift(X \Rightarrow Y) = \frac{P(Y/X)}{P(Y)}$ et $leverage(X \Rightarrow Y) = sup(X \Rightarrow Y) - sup(X)sup(Y)$.
L'algorithme d'extraction des règles d'association se décompose en deux phases :

1. L'extraction des motifs fréquents
2. L'extraction des règles à partir des motifs fréquents

L'extraction des motifs fréquents consiste à rechercher tous les motifs vérifiant la contrainte $(Ct_1)$, c'est-à-dire tous les motifs ayant une valeur pour le support supérieure au seuil minimum $min_{sup}$. Dans ce cas, on parle de motifs fréquents.
L'extraction des règles à partir des motifs fréquents consiste à rechercher toutes les règles vérifiant la contrainte $(Ct_2)$, c'est-à-dire toutes les règles ayant une valeur pour la confiance supérieure au seuil minimum $min_{conf}$. A l'issue de cette phase, on obtient les règles valides comme cela a été défini précédemment.
Dans cet article, nous nous focalisons sur les règles d'association de classe du type $X \Rightarrow y_j$. Nous recherchons donc les règles qui se concluent sur une seule variable binaire (Srikant et al., 1997). Si nous utilisons les notations définies dans la *section 2*, nous recherchons donc les règles du type $T \Rightarrow indMeO$, où $T$ est un ensemble de composés $T_i$.
Après avoir rappelé la technique d'extraction des règles d'association, nous présentons la première phase d'extraction : celles des composés potentiellement prometteurs.

## 3.2 Extraction des composés potentiellement prometteurs

La technique d'extraction des règles d'association nécessite que les données soient sous la forme binaire. En présence de variables numériques, une discrétisation suivie d'un codage disjonctif complet sont nécessaires.

Comme nous l'avons dit dans la *section 2*, la particularité de cette base de données est qu'elle est éparse (*i.e. elle possède beaucoup de valeurs nulles*) puisque les plantes ne possèdent qu'une centaine de composés en moyenne parmi les *1 075* qui ont été détectés. La fréquence moyenne d'apparition des composés dans notre base est de *10%*, soit une moyenne de *21* expressions par composé. Étant donnée cette particularité, nous allons discrétiser les variables numériques de la façon suivante : la valeur *1* sera attribuée pour tous les composés exprimés, c'est-à-dire pour toutes les valeurs supérieures strictement à *0* ; et la valeur *0* pour les composés non exprimés. Une discrétisation plus fine, c'est-à-dire avec plus de deux intervalles, conduirait à un nombre très faible de règles, voire aucune règle.

L'extraction a été effectuée en utilisant la bibliothèque `arulesViz` (Hahsler, 2017) du logiciel `R` (R Development Core Team, 2005). Nous avons retenu comme seuil minimum pour le support, la valeur de *0,025*, soit vérifié par au moins *6* individus (*substrats*), et comme seuil minimum pour la confiance, la valeur de *0,50*. La base de données possède *30%* de plantes anti-méthanogènes ce qui se traduit par *sup(indMe0)=0,30*. Par conséquent, le seuil retenu pour la confiance nous garantit que les règles extraites sont obligatoirement dans la zone attractive, c'est-à-dire la zone où $conf(T \Rightarrow indMeO) > sup(indMeO)$.

*2 892* règles de classe ont été extraites dont *26* de niveau *2*, c'est-à-dire les règles composées de *2* items et par conséquent avec un seul composé en prémisse.

Le graphique de la *figure 1* nous restitue l'ensemble de ces *26* règles. La taille du cercle est proportionnelle à la valeur de la confiance (*plus la taille est importante, plus la valeur de la confiance est élevée*) et l'intensité de la couleur du cercle est proportionnelle à la valeur du support (*plus la couleur est foncée, plus la valeur du support est élevée*). Ce graphique a été réalisé avec la bibliothèque `arulesViz`.

La meilleure règle extraite du point de vue de la confiance est la règle $T0.9792 \Rightarrow indMeO$ qui a une confiance égale à *0,70*, un support égal à *0,034* et un lift de *2,31*.

Le graphique de la *figure 1* nous permet de comparer les règles pour les deux principales mesures, ce qui nous donne une bonne vision générale des règles. Cependant cette visualisation ne permet pas d'ordonner les règles extraites. De plus, dans le cas d'une sélection des composés les plus prometteurs, il est intéressant de connaître, parmi les individus possédant le composé, combien d'entre eux ont un effet positif (*information donnée par la valeur du support absolu de la règle*) et combien ont un effet négatif (*méthanogène*). La première information correspond au nombre d'exemples d'une règle (*i.e. le nombre d'individus vérifiant la prémisse et la conclusion*) et la seconde au nombre de contre-exemples (*i.e. le nombre d'individus vérifiant la prémisse mais ne vérifiant pas la conclusion*). La connaissance de la proportion de l'un par rapport à l'autre, peut devenir une aide précieuse quant à la sélection des règles les plus prometteuses.
Ainsi, nous proposons la visualisation représentée par la *figure 2*. Un tel graphique n'est intéressant et lisible que dans le cas d'un nombre limité de règles et avec des valeurs pour la confiance pas trop proches de *1*.

Ce graphique nous restitue les informations suivantes :

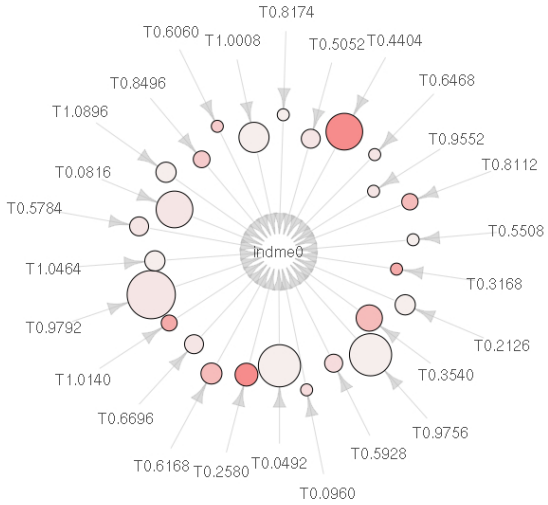


FIGURE 1 – *Règles générales du type* $T_i \Rightarrow indMeO$.

1. Le nombre d'individus vérifiant la prémisse $T_i$ ou support absolu $sup_{abs}(T_i)$ du composé $T_i$ grâce à la longueur du segment de droite (*segments rouge et bleu*).
2. Le nombre d'exemples ou le support absolu $sup_{abs}(T_i indMeO)$ de la règle $T_i \Rightarrow indMeO$ grâce à la longueur du segment de droite qui se situe à gauche de la droite d'équation $x = 0$ (*segment rouge*).
3. Le nombre de contre-exemples ou le support absolu $sup_{abs}(T_i\overline{indMeO})$ grâce à la longueur du segment de droite qui se situe à droite de la droite d'équation $x = 0$ (*segment bleu*).
   On rappelle que $sup_{abs}(T_i) = sup_{abs}(T_i indMeO) + sup_{abs}(T_i\overline{indMe0})$.
4. La confiance de la règle visible grâce à l'orientation du segment de droite : plus le segment de droite est vertical, plus la confiance de la règle est importante. Un segment de droite horizontal indique une valeur égale à *0,50* pour la confiance, et un segment de droite vertical indique une valeur égale à *1* pour la confiance.
   Les supports $sup_{abs}(T_i)$ et $sup_{abs}(T_i indMeO)$ visibles sur le graphique nous permettent de retrouver la confiance de la règle puisque $conf(T_i \Rightarrow indMeO) = \dfrac{sup_{abs}(T_i indMeO)}{sup_{abs}(T_i)}$.
5. Une mesure de notre choix sur l'axe des ordonnées. Nous avons choisi ici la mesure $M_G$ qui mesure la distance entre deux points caractéristiques : (1) soit l'équilibre [5] ou soit l'indépendance [6] et (2) l'implication logique [7]. Pour plus de détails concernant cette mesure, nous renvoyons le lecteur aux travaux de (Guillaume, 2010).

5. L'équilibre est le cas où $conf(X \Rightarrow Y) = conf(X \Rightarrow \overline{Y}) = 0,5$.
6. L'indépendance est le cas où $conf(X \Rightarrow Y) = sup(Y)$.
7. L'implication logique est le cas où $conf(X \Rightarrow Y) = 1$.

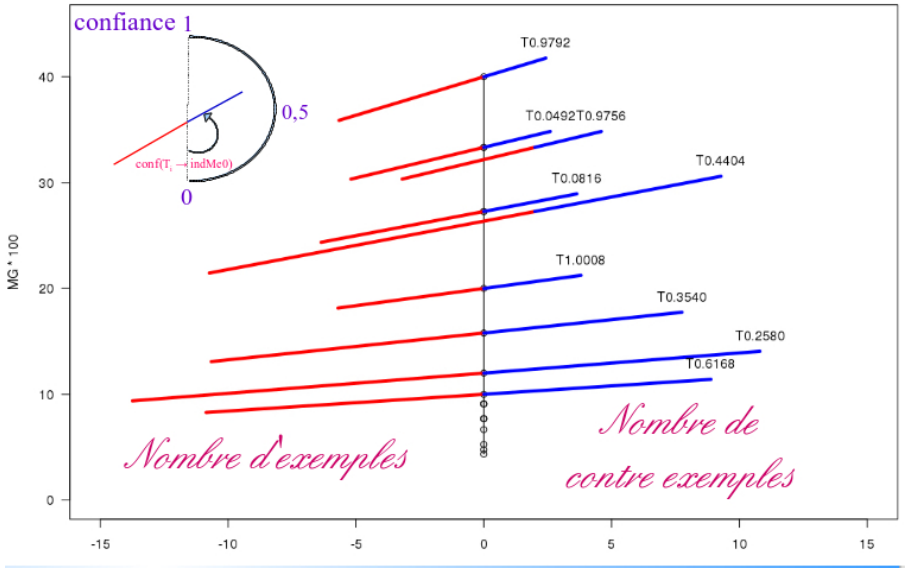


FIGURE 2 – *Visualisation des 9 meilleures règles* $T_i \Rightarrow indMeO$.

Ce mode de représentation s'inspire du diagramme de Venn, qui est ici condensé et aplati. On rappelle que pour toutes ces règles, la conclusion est la même, l'item $indMeO$. Le support absolu $sup_{abs}(indMeO)$ n'est pas représenté sur le graphique car d'une part, c'est la même valeur pour toutes les règles, et d'autre part, la valeur de son support est environ six fois plus élevée que la valeur du support de toutes les règles extraites.

On note que dans le cas où deux règles ont la même valeur pour la mesure choisie sur l'axe des ordonnées, nous effectuons une translation de la représentation de la seconde règle selon l'axe des abscisses. C'est le cas notamment pour les composés *T0.9756* et *T0.4404*.

Suite à l'extraction et à la visualisation de ces composés potentiellement prometteurs, nous souhaitons poursuivre notre étude sur ceux-ci en révélant leur comportement avec les autres composés. Nous souhaitons savoir si ceux-ci sont associés avec beaucoup d'autres composés, et dans ce cas nous parlerons de composés "*grégaires*", ou au contraire, s'ils ne sont associés à aucun autre composé, et nous parlerons de composés "*solitaires*". Outre l'obtention de connaissances supplémentaires sur les composés, cette information nous sera utile lorsque nous étudierons les synergies entre composés.

Pour savoir s'il y a de nombreuses associations avec un composé donné, nous recherchons pour chaque composé potentiellement prometteur, le nombre de motifs maximalement fréquents [8] (*que nous noterons MMF par souci de simplification d'écriture*) le contenant ainsi que la taille du plus grand MMF (*i.e. le MMF possédant le plus grand nombre d'items associés*). Plus le nombre de MMF est élevé et/ou plus la taille du plus grand MMF est importante, plus le composé sera "*grégaire*".

8. Un motif fréquent est maximal si aucun de ses super ensembles n'est fréquent.

Pour extraire ces MMF, il suffit d'ajouter le paramètre `target="maximally frequent itemsets"` dans la fonction `apriori` de `R`. *3 765* MMF ont été détectés parmi les *339 532* motifs fréquents contenant l'item $indMeO$. *11* composés "*solitaires*" ont été détectés parmi les *26* composés potentiellement prometteurs. Ces *11* composés sont de véritables solitaires puisqu'ils ne sont associés à aucun autre composé. Les composés "*solitaires*" sont les suivants : *T0.9792*, *T0.0492*, *T0.9756*, *T1.0008*, *T0.2126*, *T1.0464*, *T0.5784*, *T0.6696*, *T0.5508*, *T0.6468* et *T0.8174*. Ce sont généralement des composés ayant de faibles valeurs pour le support, ce qui n'est pas surprenant. Les trois meilleures règles au regard de la confiance sont constituées de composés solitaires (*T0.9792, T0.0492 et T0.9756*).

Afin de restituer cette information de façon plus intelligible aux utilisateurs, nous avons effectué une classification hiérarchique ascendante en retenant comme distances : la distance euclidienne et la distance de Ward. Nous ne nous sommes pas limités à ces deux dernières informations (*nombre de MMF et taille du plus grand MMF*) mais avons repris toutes les informations obtenues lors de l'extraction, à savoir : (1) le *support* de la règle, (2) la *confiance*, (3) le *support* du composé et (4) le *leverage*. Nous avons supprimé le lift qui est lié de façon linéaire à la confiance lorsque les règles d'association de classe ont la même conclusion.

La *figure 3* nous restitue la classification obtenue. Deux grandes catégories de composés se dégagent dans la constitution des règles générales. La première catégorie est constituée de composés ayant peu de MMF (*les composés solitaires*), et la seconde catégorie, de composés ayant beaucoup de MMF (*les composés grégaires*). Dans chacune de ces deux catégories, on trouve deux sous classes qui se distinguent par le fait que les caractéristiques énoncées précédemment sont un peu plus prononcées dans l'une des deux sous classes. Plus on va vers les composés de droite, plus ceux-ci ont des MMF et avec des longueurs plus importantes. Cette classification a donc ordonné les composés des plus "*solitaires*" vers les plus "*grégaires*".

Après avoir extrait les composés potentiellement prometteurs sur la base de données binaires, nous confrontons ces résultats avec les données initiales, c'est-à-dire avec les données numériques, afin de prendre en compte l'intensité d'expression des composés.

# 4 Sélection des composés les plus prometteurs

Plus un composé est fortement présent dans une plante, plus la valeur de celui-ci sera élevée. Ainsi, pour notre problématique, une règle de classe sera d'autant plus intéressante que le composé phénolique sera fortement exprimé, donc aura de fortes valeurs. Par conséquent, les règles qui vont particulièrement nous intéresser sont celles où les fortes valeurs pour le composé $T_i$ sont présentes, règles que nous pouvons formaliser de la façon suivante :
$T_i \geq v \Rightarrow indMe0$ avec $v$ une valeur prise par le composé $T_i$.

Afin de détecter ce type de règles, nous retenons la stratégie suivante que nous expliquons en nous appuyant sur un exemple.

La *figure 4* restitue l'ensemble des valeurs prises par le composé *T0.6696*, et ceci par catégorie de substrats, c'est-à-dire ceux pour lesquels il n'y a aucun effet sur les émissions de méthane et ceux pour lesquels il y a un effet positif (*diminution de méthane*). Nous recherchons donc la valeur optimale $v_{opt}$ du composé où la proportion de substrats ayant un effet positif est supérieure à la proportion de substrats n'ayant aucun effet.

Pour se faire, nous allons évaluer toutes les règles pour chacune des valeurs prises par le composé $T_i$, excepté évidemment la valeur minimale. Comme nous voulons que ces règles

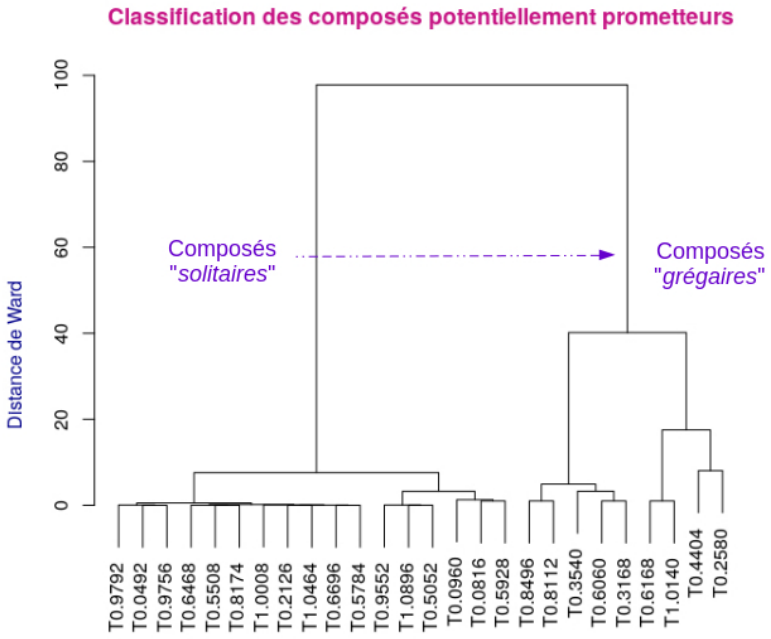


FIGURE 3 – *Classification ascendante hiérarchique des composés phénoliques prometteurs.*

vérifient le support minimum déterminé par l'utilisateur, nous n'allons évaluer qu'un sous-ensemble des règles possibles. Afin de formaliser notre stratégie d'extraction, nous définissons les notations suivantes. Soit $t_i$ le nombre de valeurs distinctes prises par le composé $T_i$ et soit $\{v_{i1}, .., v_{ik}, .., v_{it_i}\}$ avec $k \in \{1, .., t_i\}$ l'ensemble des valeurs ordonnées prises par le composé. Soit $s$ le support absolu minimum déterminé par l'utilisateur ($s = min_{sup} \times n$). Nous recherchons donc la ou les meilleures règles au regard d'une mesure de qualité (*confiance, leverage, ...*) choisie par l'utilisateur parmi toutes les règles suivantes : $T_i \geq v_{ik} \Rightarrow indMeO$ avec $v_{ik} \in \{v_{i2}, .., v_{i(t_i - s)}\}$.

Voici un exemple de règle extraite : $T0.6696 \geq 60,33 \Rightarrow indMeO$ avec une valeur pour la confiance de *0,875* et une valeur pour le support de *0,034*. La valeur de la confiance de la règle binaire $T0.6696 \Rightarrow indMeO$ extraite précédemment est de *0,54* et la valeur du support de *0,034*. Il y a une amélioration importante de la confiance lorsque le composé $T0.6696$ est présent sous la forme d'un pic majeur (> *1 000 mAU*).

Afin de nous guider dans le choix final de ces règles, nous utilisons une nouvelle mesure, l'intensité d'expression de la règle $Int_{exp}$, qui va nous renseigner sur l'intensité de la règle par rapport à l'intensité moyenne du composé. C'est le rapport entre la moyenne des valeurs prises par la règle numérique, c'est-à-dire la moyenne des valeurs supérieures à $v_{ik}$, et la moyenne des valeurs prises par le composé $T_i$ :

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

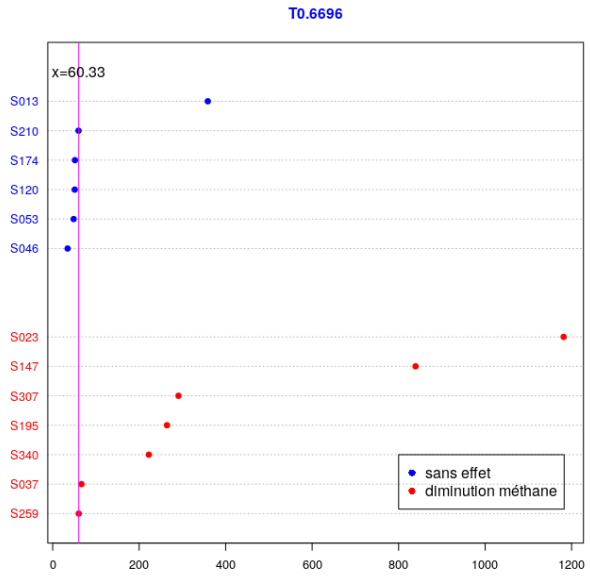


FIGURE 4 – *Distribution des valeurs pour le composé T0.6696 et ceci par catégorie d'effet.*

Ainsi, plus l'intensité de la règle est supérieure à *1*, meilleure sera celle-ci.

La règle $T0.6696 \geq 60,33 \Rightarrow indMeO$ a une intensité d'expression $Int_{exp}$ de *1,51*. C'est une règle prometteuse, donc un composé à étudier.

A l'issue de cette étape, parmi les *11* composés phénoliques solitaires, *7* ont montré un effet seuil qui permet d'améliorer encore la confiance : $T0.9792$, $T0.0492$, $T0.9756$, $T1.0008$, $T1.0464$, $T0.6696$, $T0.5784$.

# 5 Conclusion

A partir des *1 075* composés phénoliques non identifiés qui étaient présents dans notre jeu de plantes, la fouille des données par la technique d'extraction des règles d'association de classes, a restitué dans un premier temps *26* composés prometteurs. La nouvelle visualisation des règles qui est proposée, permet d'intégrer dans la représentation graphique cinq mesures importantes et guident l'utilisateur plus efficacement dans ses choix et donc dans la sélection des composés. Enfin, après une discrétisation contextuelle des règles, l'évaluation de l'intensité des règles par la nouvelle mesure proposée, permet encore d'affiner la pertinence des règles, et de réduire la sélection à quelques composés qu'il sera alors possible d'identifier. L'examen des spectres ultra-violet de ces composés montre déjà qu'ils appartiennent à la famille des acides cinnamiques pour une part, et à la famille des flavonols d'autre part. Il restera à les identifier précisément, à obtenir les produits purs par synthèse et à vérifier qu'ils sont effectivement responsables d'un effet anti-méthanogène en reproduisant l'essai de fermentation avec les produits de synthèse. Nous souhaitons par ailleurs continuer le travail d'extraction des

connaissances à partir de la matrice des données afin de dégager les associations de composés qui augmentent la probabilité d'avoir un effet anti-méthanogène, et faire ressortir des synergies entre les composés.

## Références

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

Brin, S., R. Motwani, J. Ullman, et S. Tsur (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD international conference on management of data*, pp. 255–264.

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

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

Macheboeuf, D., A. Cornu, S. Kerros, et F. Recoquillay (2018). An antimethanogenic index for meadow plants consumed by ruminants. In R. Baumont, M. Silberberg, et I. Cassar-Malek (Eds.), *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).*, Volume 9, pp. 608. Cambridge University Press.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases 1991*, pp. 229–248. MIT Press.

R Development Core Team (2005). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*.

Sakakibara, H., Y. Honda, S. Nakagawa, H. Ashida, et K. Kanazawa (2003). Simultaneous determination of all polyphenols in vegetables, fruits, and teas. In *Journal of Agricultural and Food Chemistry*, Volume 51, pp. 571–581.

Srikant, R., Q. Vu, et R. Agrawal (1997). Mining association rules with item constraints. In *Proceedings ACM SIGKDD'97*, pp. 67–73.

## Summary

Methane is a powerful greenhouse gas. In Europe, methane emissions come mainly from breeding, and in particular from ruminants that have in their paunch a microbial ecosystem that ferments the plant matter. The purpose of this paper is to find the phenolic compounds of plants that could have an action on these microbes and limit the production of methane, in order to propose natural alternatives of feedstuff. As these compounds have a very wide variety of chemical structures, it is not possible to test them all. So we used data mining, and more specifically class association rules, to bring out compounds that could have a significant effect.