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. L'objectif de cet article est de rechercher les composés phénoliques des plantes qui pourraient avoir une action sur les microbes du rumen et limiter la production de méthane. Comme il y a une très grande diversité de structures chimiques, nous avons eu recours à la fouille de données pour faire émerger des composés susceptibles d'avoir un effet significatif. La pertinence des règles d'association a été améliorée grâce à une nouvelle mesure d'intensité qui a permis de sélectionner quelques composés qui seront à identifier. Ainsi, parmi les *1075* composés inconnus présents dans le jeu de plantes, *26* ont émergé desquels *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. Pendant la digestion, la matière végétale est dégradée par l'écosystème microbien du rumen et produit entre autres, des gaz de fermentation dont le méthane. Les composés phénoliques des plantes peuvent avoir un effet sur les fermentations ruminales et diminuer la production de méthane. Cependant, la très grande diversité des structures chimiques possibles pour ces composés ne permettait pas de tester pour tous, leur activité.

Ainsi, notre stratégie a été d'effectuer un essai de criblage par fermentations *in vitro* sur *208* plantes et d'identifier parmi celles-ci, les plantes bio-actives contre le méthane. Parallèlement le profil en composés phénoliques de chacune de ces plantes a été déterminé. L'analyse par *HPLC-DAD* 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 ne pouvaient pas être identifiés car il y avait en moyenne plus d'une centaine de pics par plante. La priorité était de sélectionner quelques composés susceptibles d'être responsables de 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 (Srikant et al., 1997), pour faire émerger les composés susceptibles d'avoir un effet positif. Nous nous focalisons ici sur la recherche des composés actifs et non sur les synergies car il était impératif de sélectionner d'abord un nombre raisonnable de composés. En effet, 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*).

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.

2.1 Les variables prémisses

Les composés phénoliques sont extraits des substrats par un traitement éthanol : eau puis séparés avec une chaîne *HPLC*. Le profil chromatographique de chaque plante est enregistré à *280* nm. Un composé de référence, la flavone, a été utilisé 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. 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. Le nombre de composés différents trouvés en moyenne par plante est de *106*. 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.

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 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 mesurée *in vitro* est 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. Toute plante qui a un index supérieur 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*.

3 Extraction des composés potentiellement prometteurs

La technique d'extraction des règles d'association (Agrawal et Srikant, 1994) nécessite que les données soient sous la forme binaire. Comme nous l'avons dit dans la *section 2*, la particularité de cette base de données est qu'elle est éparse puisque les plantes ne possèdent qu'une centaine de composés en moyenne parmi les *1 075* qui ont été détectés. Nous allons donc discrétiser les variables numériques prémisse 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.

L'extraction a été effectuée en utilisant la bibliothèque `arulesViz` (Hahsler, 2017) du logiciel `R`. Nous avons retenu comme seuil minimum pour le support [1], la valeur de *0,025*, soit vérifié par au moins *6* individus (*substrats*), et comme seuil minimum pour la confiance [2], 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.

Afin d'aider les biologistes dans le choix des composés prometteurs, nous avons proposé la visualisation représentée dans la *figure 1*. 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 :

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.
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) l'équilibre $(conf(X \Rightarrow Y) = 0,5)$ ou l'indépendance $(conf(X \Rightarrow Y) = sup(Y))$ et (2) l'implication logique $(conf(X \Rightarrow Y) = 1)$. Pour plus de détails concernant cette mesure, nous renvoyons le lecteur aux travaux de (Guillaume, 2010).

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

1. Le support $sup(X)$ du motif X évalue la proportion d'individus le vérifiant.
2. La confiance $conf(X \Rightarrow Y)$ de la règle $X \Rightarrow Y$ est la probabilité conditionnelle $P(Y/X)$.

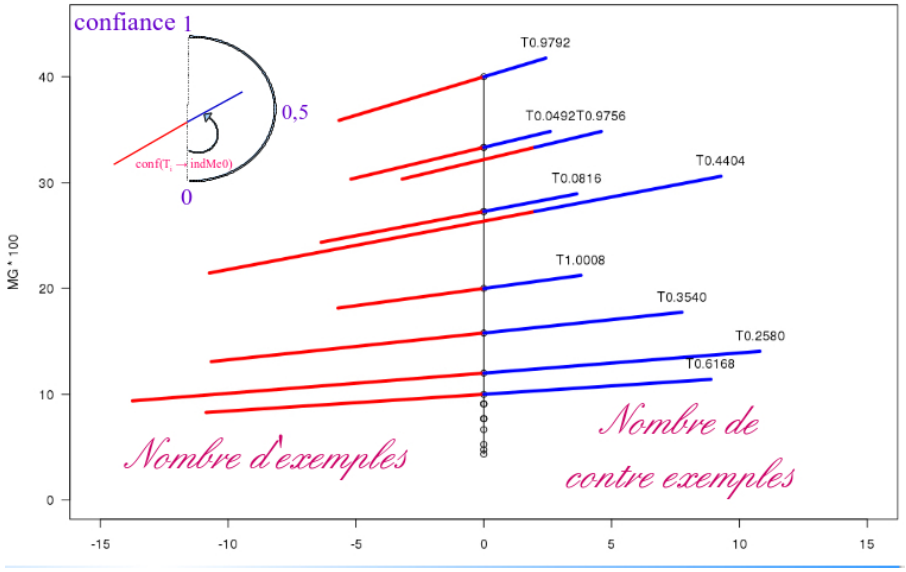

FIGURE 1 – *Visualisation des 9 meilleures règles* $T_i \Rightarrow indMeO$.

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*.

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 2* 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

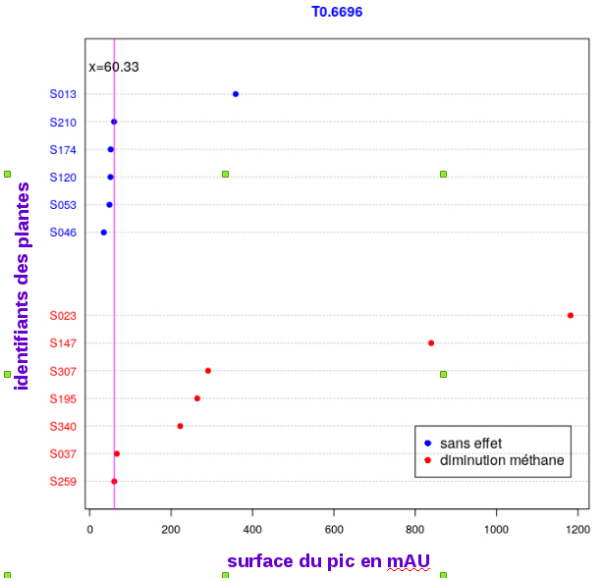

FIGURE 2 – *Distribution des valeurs pour le composé T0.6696 et ceci par catégorie d'effet.*

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 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. 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 :

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

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, *7* composés 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, l'extraction des règles d'association de classe, 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 guide 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.

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.

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.

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.