# A knowledge-driven network-based analytical framework for the identification of rumen metabolites

Wang, Mengyuan; Wang, Haiying; Zheng, Huiru; Dewhurst, Richard J.; Roehe, Rainer

*Published in:*
IEEE Transactions on Nanobioscience

*DOI:*
10.1109/TNB.2020.2991577

Print publication: 01/07/2020

*Document Version*
Peer reviewed version

# A knowledge-driven network-based analytical framework for the identification of rumen metabolites

| | |
|---|---|
| Journal: | *Transactions on NanoBioscience* |
| Manuscript ID | Draft |
| Manuscript Type: | BIBM |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Wang, Mengyuan; Ulster University Faculty of Computing Engineering and The Built Environment, Computing<br>Wang, Haiying; University of Ulster, Faculty of Engineering<br>Zheng, Huiru; University of Ulster, School of Computing and MAthematics<br>Dewhurst, Richard; Scotland's Rural College<br>Roehe, Rainer; Scotland's Rural College |
| Keywords: | Metabolomics, NMR analysis, Mutual information, Network analysis, Metagenomics |
| | |

# A knowledge-driven network-based analytical framework for the identification of rumen metabolites

Mengyuan Wang
*School of Computing*
*Ulster University*
United Kingdom
wang-m5@ulster.ac.uk

Haiying Wang
*School of Computing*
*Ulster University*
United Kingdom
hy.wang@ulster.ac.uk

Huiru Zheng*
*School of Computing*
*Ulster University*
United Kingdom
h.zheng@ulster.ac.uk

Richard J. Dewhurst
*Scotland's Rural College*
Edinburgh, United Kingdom
Richard.Dewhurst@sruc.ac.uk

Rainer Roehe
*Scotland's Rural College*
Edinburgh, United Kingdom
Rainer.Roehe@sruc.ac.uk

***Abstract*—Metabolites are the final production of biochemical reactions in the rumen micro-ecological system and are very sensitive to changes in rumen microbes. Nuclear magnetic resonance (NMR) spectroscopy could both identify and quantify the metabolic composition of the ruminal fluid, which reflects the interaction between rumen microbes and diet. The main challenge of untargeted metabolomics is the compound annotation. Based on non-linear and linear associations between microbial gene abundances and integrals derived from NMR spectra, combined with knowledge of enzymatic reaction from the KEGG database, this study developed a knowledge-driven network-based analytical framework for the inference of metabolites. There were 89 potential metabolites inferred from the integral co-occurrence network. The results are supported by dissimilarity network analysis. The coexistence of non-linear and linear associations between microbial gene abundances and spectral integrals was detected. The study successfully found the corresponding integrals for acetate, butyrate and propionate, which are the major volatile fatty acids (VFA) in the rumen. This novel framework could very efficiently infer metabolites to corresponding integrals from NMR spectra.**

***Keywords—Metabolomics, NMR analysis, KEGG pathway, Mutual information, Rumen Microbe, Network analysis***

## I. INTRODUCTION

The rumen is the primary organ for microbial fermentation of ingested plant material for domestic ruminant livestock. Diet and feed quality are essential for rumen health which is critical to the growth and high-quality of livestock production [1]. The metabolites in the ruminal fluid could reflect the health of interaction between rumen microbes and the diet [1]. Since health status, meat production and milk quality of cattle are directly dependent on rumen metabolites, the chemical composition analysis of ruminal fluid could offer valuable biochemical insights into the rumen-diet microbial interactions [1].

Quantitative metabolomics studies metabolites in cells, tissues, organs, or microorganisms and is widely applied in the host phenotype prediction, biomarker selection, clinical diagnosis, and drug discovery [2]. In metabolic analysis methods, Nuclear magnetic resonance (NMR) does not require elaborate sample preparation while could directly detect the composition of metabolites in the host [3]. The nucleus (i.e., $^{1}H$ or $^{13}C$) with a non-zero nuclear magnetic moment is recorded to absorb the specific resonance frequency under the action of an external magnetic field, thereby predicting the molecular structure of the compound [4]. Non-targeted metabolomics can detect hundreds or thousands of metabolites from a single sample, however, identification and quantification of metabolites in non-targeted metabolomics results remain difficult. A statistical survey of the metabolomics platform published by Metabolomics Workbench and MetaboLites [5] showed that metabolites were identified in non-targeted metabolomics studies only accounted for 20%-30% of the total detected metabolites [6]. There are as many as 2,500 metabolites in animals [10], and most of the database is limited. A survey reported that 246 rumen metabolites were present in the bovine ruminal fluid metabolome database (BRDB) [1], and the updated database now includes 335 metabolites. Due to the shortage of known correspondence between signal spectrum and compound, identification of metabolites in non-targeted metabolomics is still challenging [7]. This research constructed a framework for the identification of metabolites by the knowledge-driven network-based analytical methodology.

The main contributions of this research are summarized below.

- In contrast to the traditional methods in which identify compounds by matching their proton or NMR spectra of molecular structures in the spectral library [8], this study developed a novel approach to infer metabolites to spectral integrals using microbial gene abundances determined by whole metagenomic sequencing. The integrals are the relative intensity of signals in the NMR analysis, which is close to the ratio expected for a pure compound [9]. This study is an innovative attempt to annotate integrals by combining genomics information with the quantitative association and metabolic pathway knowledge.
- This research also involves innovations in the application and integration of classical research methods. The current metabolomics methods [8] always identify metabolites through database matching and chemical structure comparisons. In this study, as same as the Euclidean distance dissimilarity network, the network-based Markov clustering is firstly applied to mining integrals from the same compound. This study explored the linear and non-linear quantitative relationships between

This research is jointly supported by Ulster University and Scotland's Rural College, U.K. and partially supported by the MetaPlat project, (www.metaplat.eu), funded by H2020-MSCA-RISE-2015.

microbial genes and integrals, and integrated both associations for subsequent analysis through a novel ranking pipeline.

## II. METHODOLOGY

The pipeline for integrated multi-omics inference methodology adopted in this research is illustrated in Fig. 1. Based on the 128 integrals from non-targeted metabolomics by NMR technology and relative abundance of 1461 microbial genes identified in a metagenomics analysis, a network-based approach has been developed to infer integral-gene-compound association. By incorporating domain knowledge from KEGG pathways, the integrals network and integral-gene-compound network was further investigated in terms of biological relevance and topological structure.

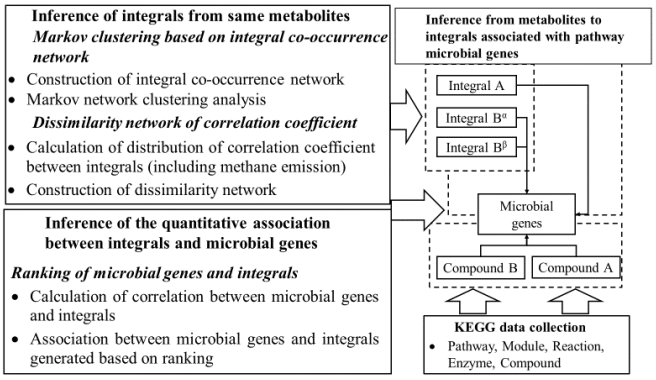


Fig. 1. Pipeline for the inference of metabolites based on the association between the integrals and microbial gene abundances.

### A. Microbial Multi-omics Data Description

Data were from a 2×4×4 factorial experiment of diets, genotypes, and additives in beef cattle that were designed by Roehe et al. [11-13]. The experiment was conducted by the Beef and Sheep Research Centre of Scotland's Rural College (SRUC, Edinburgh, UK). After removing the samples with missing metabolite data, a total of 36 rumen fluid samples was included in the study.

Data used in this study include methane emissions, relative abundance of rumen microbial genes, and integrals associated with the rumen metabolites.

#### 1) Metagenomics data

There were 1461 microbial genes with a relative abundance higher than 0.001%. The reader is referred to [11], [14] for a detailed description of data generation.

#### 2) Metabolomics data

Integrals were derived from the peak of $^{1}$H NMR (Nuclear Magnetic Resonance) spectroscopy, which is in correspondence with the hydrogen atom of each signal in the NMR spectrum. The relative intensity of the signal in the NMR analysis reflects the relative content of each component in the sample. In this paper, 128 integrals obtained by NMR analysis were used. One integral was identified as acetate, and nine integrals were identified as possibly related to propionate and butyrate. The metabolites associated with the remaining 118 spectral integrals are still unknown.

### B. Markov clustering based on integral co-occurrence network

Before conducting network analysis, an exploratory study of the data was performed. The difference in the integral sample distribution was tested by the Kolmogorov-Smirnov two-tail test [36]. The *x*-*y* scatter plots exhibit the similarity between integrals.

#### 1) Construction of integral co-occurrence network

The integral co-occurrence network was obtained by pipeline in Fig. 2 to reduce false-positive correlations [34].

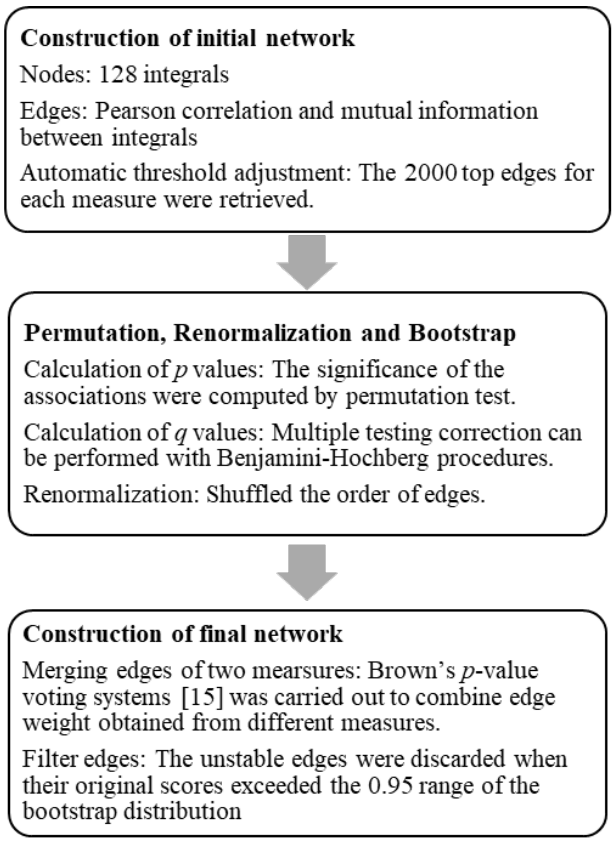


Fig. 2. Pipeline for construction of integral co-occurrence network.

As shown in Table I, this research score the association strength by linear correlations (i.e., Pearson correlation and Spearman correlation), similarity (i.e., Mutual information) and dissimilarity (i.e., Euclidean distance).

TABLE I
SUMMARY OF QUANTITATIVE ANALYSIS METHODS

| Measurements | Range | Definition |
|---|---|---|
| Pearson correlations | [-1, +1] | $\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$ |
| Spearman correlations | [-1, +1] | $\rho_{rx,ry} = \frac{cov(r_x, r_y)}{\sigma_{rx}\sigma_{ry}}$ |
| Mutual information | [0, INF] | $I(x; y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\left(\frac{P(x,y)}{P(x),P(y)}\right)$ |
| Euclidean Distance | [0, INF] | $D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ |

[a] $cov(x,y)$ is the covariance; $\sigma_x$ is the standard deviation of X; $\sigma_y$ is the standard deviation of Y. $r_x$ *and* $r_y$ are ranks of each observation; $P(x,y)$ *is the joint probability mass function of* $X$ *and* $Y$; P(x) and P(y) are the marginal probability mass functions of X and Y respectively. $x_i$ represents the Pearson correlation between the integral x and the integral *i*.

*2) Markov network clustering analysis*

The Markov Clustering algorithm (MCL) was performed on the integral co-occurrence network to identify natural clusters by mathematical bootstrapping procedures [16]. It carries out random walks within a network by the operation of expansion and inflation [17]. The specific implementation process is the correlation matrix pass through iterative rounds of matrix expansion and matrix inflation until the matrix stop changing. The matrix is finally interpreted as clustering results [18]. This study performed the Markov clustering based on the weights of each edge. The inflation parameter was manually adjusted to 3 when all the clusters obtained the average shortest path length less than or equal to 1.

### *C. Construction of dissimilarity network based on the correlation coefficient between integrals*

*1) Calculation of correlation coefficient between integrals (including methane emission)*

Different from mutual information, linear correlation measures could assign a positive or negative trait to a predicted relationship, which reflects whether a variable is consistently increasing with another variable or decreasing. Therefore, this study calculated Pearson correlations between the integrals (including methane emission levels). A 129 × 129 correlation coefficient matrix was constructed, each row representing the correlation of an integral with the other 127 integrals and methane emission observed in the samples.

*2) Construction of the dissimilarity network*

Euclidean distance (Table 1) as an important indicator for comparing species distributions in biological research was used to measure the dissimilarities between correlation coefficient distributions [19]. Following the similar processes used in the construction of the integral co-occurrence network, a dissimilarity-based network was generated in which nodes are integrals and edges represent Euclidean distances between integral correlation coefficient distributions. Only edges with a significant *p*-value after multiple tests were included.

### *D. Inference of the quantitative association between integrals and microbial gene abundance*

The quantitative associations between microbial genes and integrals were finally inferred by the pipeline illustrated in Fig. 3.

### *E. KEGG data collection*

As one of the largest data sources and powerful tools for metabolic system analysis and network research, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database [20] contains various pieces of biochemical information such as genome sequences, biochemical reactions and pathway data of various species. The database consists of several databases such as PATHWAY, GENES and LIGAND. The PATHWAY database contains detailed pathway information, including metabolic pathways (i.e., glucose metabolism, lipid metabolism and amino acid metabolism). It also stores genes, enzymes, metabolites, reactions, and relationship data between them by species and pathway. In this study, the metabolic pathways, metabolic modules, enzymatic reactions, enzymes related to these microbial genes, and corresponding metabolites were collected through the KEGG database API.

There are 961 microbial genes (Fig.4) involved in 219 metabolic pathways (excluding global metabolic pathways). Among these metabolic pathways, map03010 (ribosome) contains 87 microbial genes, followed by map00680 (methane metabolism), which contains 67 microbial genes. There were 386 microbial genes mapped to 164 modules, of which M00567 (methanogenesis) contained the most microbial genes, with a total of 39. Only 876 of the 1461 microbial genes were mapped to the corresponding 730 enzymes. Among them, 730 enzymes were mapped to the corresponding 1554 reactions and 1870 corresponding compounds.

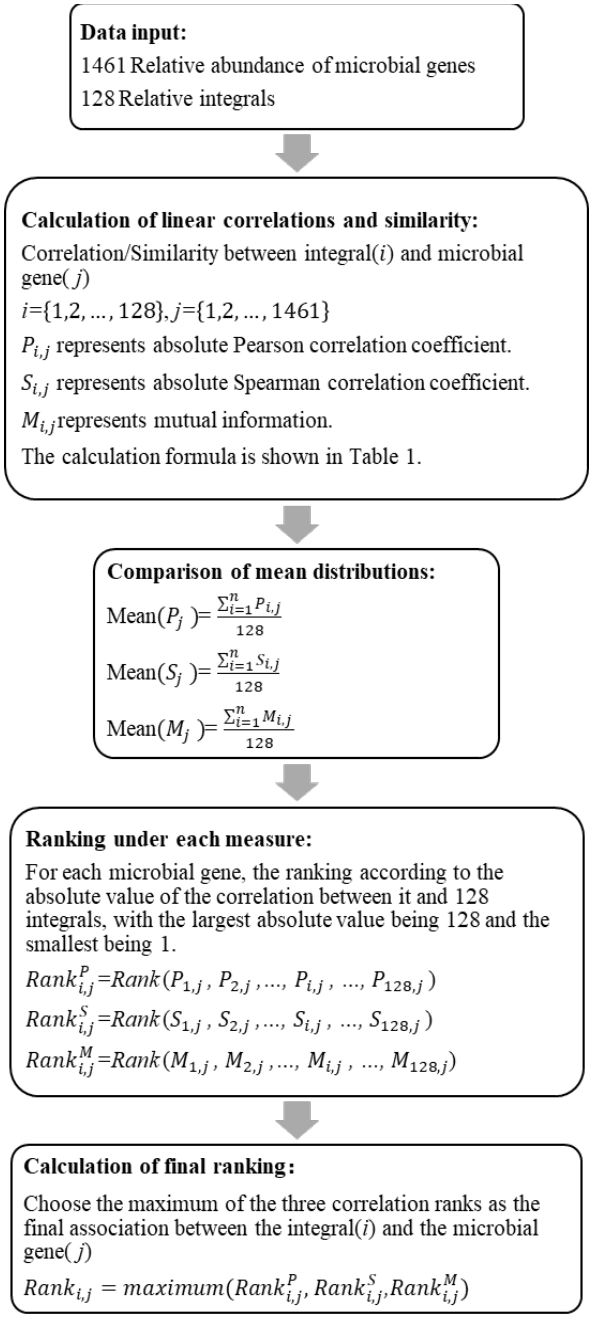


Fig. 3. Pipeline for the calculation of rankings based on the association between the integrals and microbial genes.

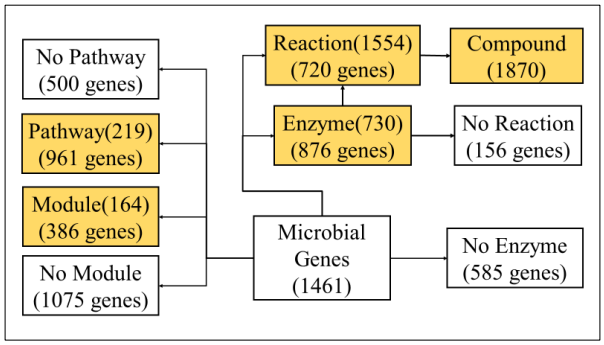


Fig. 4. KEGG data collection details. (The Orange square represents successfully collected data)

### *F. Inference from KEGG compound to the corresponding integrals.*

The compounds of interest can be found in the information collected from the KEGG database [20] so that microbial genes involved in the corresponding reactions could be mapped. Compound-gene-integral networks can then be constructed with compounds, microbial genes, and integrals as nodes.

There are two ways to generate edges of the network. The link between a compound and a microbial gene indicates whether they are involved in the same biochemical reaction found in the KEGG database. The weights of edges are represented by the ranking (1-128) between microbial genes and integrals calculated by the procedure depicted in Fig. 3.

### *G. Software and tools*

Correlation (Pearson, Spearman), similarity (mutual information), and dissimilarity (Euclidean distance) calculations were carried out with the CoNet app [21], which was also used to do data renormalization, multiple tests, *p*-value correction, bootstrapping, and Brown's p-value voting process. MCL was performed by the Cytoscape plugin clusterMaker [23]. The network topological analysis and network visualization were done by Cytoscape 3.7.1 [22]. Rank and other data analysis were implemented through Matlab. The pathway was mapped by Pathview package in R software [24].

## III. RESULTS

### *A. Inference of integrals from the same metabolites*

Since each spectral integral corresponds to hydrogen atoms of each metabolite in each environment, it is highly probable that each metabolite is composed of more than one integral [6]. The first step in the study is to infer the association between integrals to find which integrals belong to the same metabolite. The formation of different peaks in NMR analysis depends on the structure of molecules, solvent, temperature, the strength of the magnetic field used in the NMR analysis, and other adjacent functional groups [35]. The hypothesis here is that the samples were all obtained in an experimental environment with indifferent solvents and temperatures. The molecular structures are the basis of signal peak formation. Therefore, integrals from the same compound should maintain consistent trends across all the samples, when the concentration of the corresponding compound changing.

The distributions of the relative values of integrals across samples are significant ($p$<0.05) different. In contrast, integrals from the same compounds exhibit a similar trend, as the examples in Fig.5, which the distributions of Integral $B^{\alpha}$ and Integral $B^{\beta}$ belonging to the same compound and the Integral A from a different compound. In this study, a network based on Pearson correlation and mutual information between integrals was used to capture the association between the integrals of the same compound.

#### *1) The integral co-occurrence network based on Pearson correlation and mutual information*

The original network includes 128 integrals as nodes, and 4008 edges were calculated by Pearson correlation and mutual information. The final co-occurrence network consists of 1027 significant associations among 107 integrals after filtering through multiple tests (Fig. 6). Twenty-one integrals have been filtered out in the final network. Because of the absence of close associations with other integrals, they should be from different metabolites. The network eventually consisted of a dense cluster and five integral pairs, as shown in Fig. 6. For instance, the pair of ButyrateCH3.1 and ButyrateCH2b.1 has been separated from the others, suggesting that ButyrateCH3.1 and ButyrateCH2b.1 are most likely to constitute one metabolite.

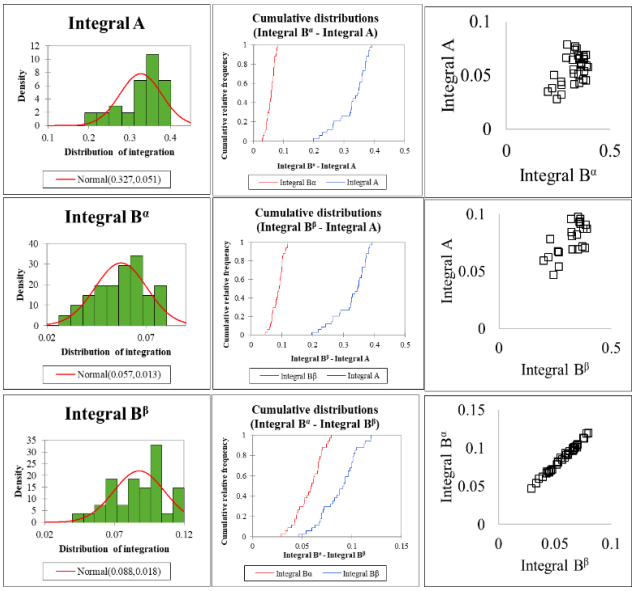


Fig. 5. Distribution of integral relative values and scatter plot of the relationship between integrals. (Integral A represents the integral of compound A. Integral $B^{\alpha}$, and Integral $B^{\beta}$ represent two integrals of compound B.)

#### *2) Markov clustering on integral network*

In an attempt to identify integrals which were obtained from the same metabolite, Markov clustering was performed, and a total of 22 clusters were formed, including 61 nodes and 70 links between nodes (Fig. 7). The average shortest path of each cluster should be 1, which means that the members in each cluster are connected equally, and they are likely to be from the same metabolite. In this study, 128 integrals were inferred to be associated with 89 potentially independent metabolites. These results need to be further verified by experiments.

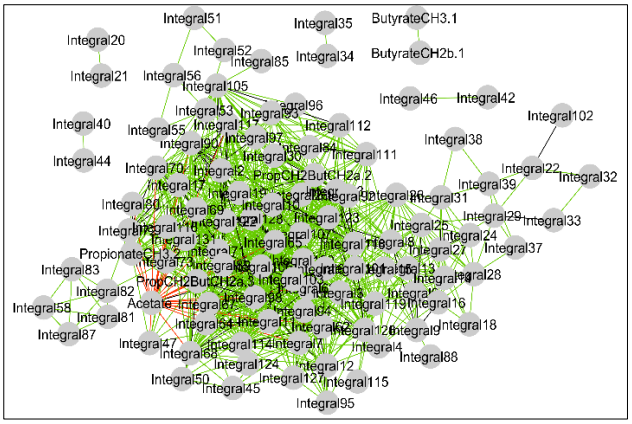


Fig. 6. Co-occurrence network of integrals. (The nodes represent integrals, the edges represent linear correlations and similarity, the absence of edge represent no correlations after multiple tests).

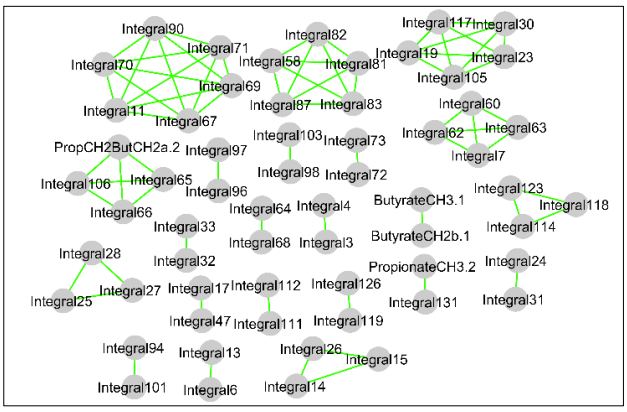


Fig. 7. Clustering results of a Markov clustering on the integral co-occurrence network. (The nodes represent integrals, the edges represent linear correlations and similarity, the absence of edge represent no correlations after multiple tests).

### *3) A dissimilarity network of correlation coefficient based on Euclidean distance*

In addition to the above relationship about the changing trend of the integrals across the samples, two other conditions can be used to judge whether the two integrals are from different compounds. If the two integrals are inversely correlated (i.e., a positive and a negative) to the third integral or the compound, then they should come from different compounds. Alternatively, the two groups of correlations have the same direction (i.e., Both positive or both negative), but the correlation coefficients are too different, then it is very likely that these two integrals are from different compounds. To further verify the integrals in both cases, this study calculated the Pearson correlation between all integrals (including methane emissions). Since methane emissions corresponding to the same batch of samples showed significantly different levels in the previous study [25], it is necessary to take the levels of methane emission into account when estimating correlation profiles between integrals. Then the distributions of correlation coefficients between the integrals were investigated. As illustrated in Fig. 8, two spectral integrals belonging to the same compound (i.e., Integral $B^{\alpha}$ and Integral $B^{\beta}$) exhibit a similar correlation distribution. In contrast, the integrals derived from different compounds (i.e., Integral A and Integral $B^{\alpha}$, or Integral A and Integral $B^{\beta}$) have significantly different distributions ($p$<0.05).

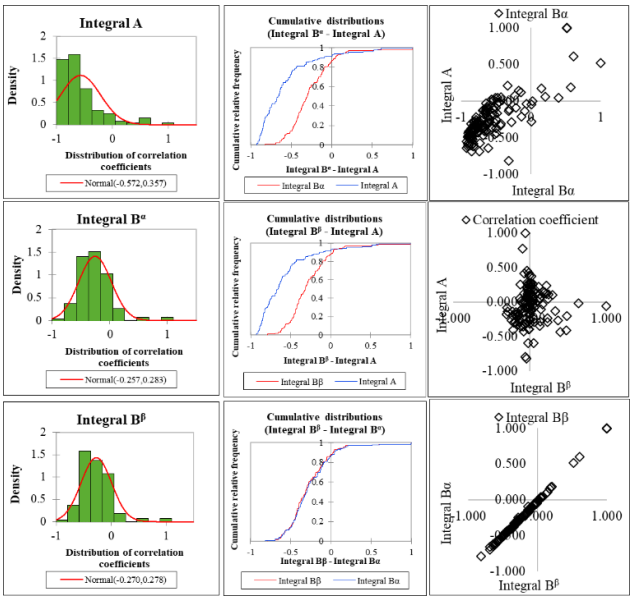


Fig. 8. Distributions of correlation coefficients and scatter plots of the correlation coefficient between integrals. (Integral A is the only spectral integral from Compound A. Both Integral $B^{\alpha}$ and Integral $B^{\beta}$ are from the same compound.

Thus, a dissimilarity network was constructed in which nodes denote integrals and links represent the Euclidean distance between correlation profiles associated with each integral. As shown in Fig. 9, the similarity network constructed exhibits a visible modular structure. A total of 17 integrals was founded in the network forming six well-separated clusters, suggesting these integrals are most likely from 6 different metabolites. The observation is mostly consistent with those obtained from the above clustering analysis of the co-occurrence network except for the cluster containing Integrals 22 and 39.

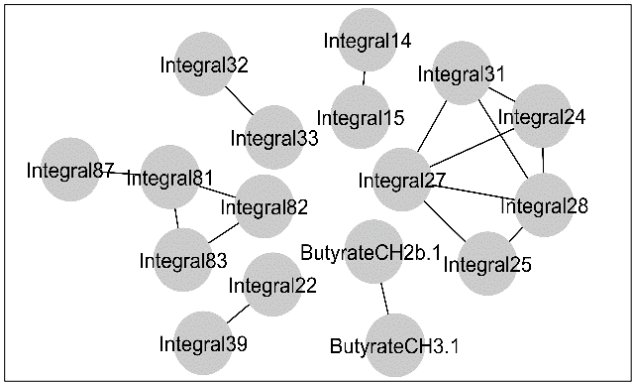


Fig. 9. A Euclidean distance network based on the distribution of correlation coefficients between integrals. (Including methane emission.)

### *B. Inference of the quantitative association between integrals and microbial gene abundance*

The means of the association between the microbial gene ($j$) and 128 integrals were calculated by Pearson correlation, Spearman correlation and mutual information, respectively. Different from mutual information values range from 0 to positive infinity, Pearson correlation and Spearman correlation generates a positive or negative value from -1 to 1 (Table I). Each box in Fig. 10 represents the distribution of the mean of each measure, highlighting the

distribution of the average values based on mutual information has lower variance. The standard deviation of mutual information is 0.027, followed by the standard deviation of Spearman correlation distribution is 0.084, and the standard deviation of Pearson correlation distribution is 0.091.

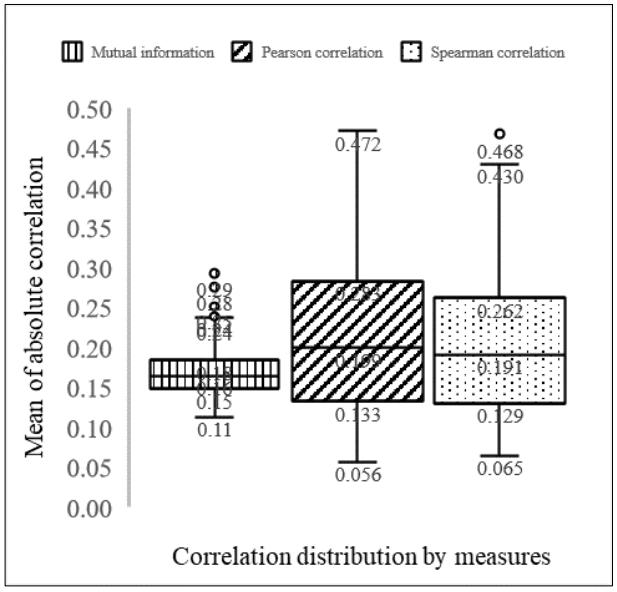


Fig. 10. Comparison of non-linear and linear correlation distributions between microbial genes and integrals. (The standard deviations (SD) for Mean ($M_j$ ) range from 0.039 to 0.089, the SD of Mean ($P_j$ ) range from 0.044 to 0.199 and the SD of Mean ($S_j$ ) range from 0.048 to 0.179.)

Taking acetate as an example, we found that about 20% of microbial genes have a stronger linear correlation than nonlinear association measured by mutual information as depicted in Fig. 11.

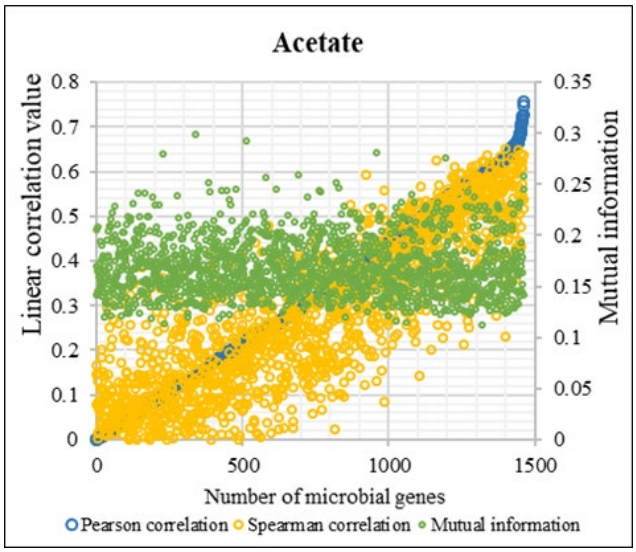


Fig. 11. A scatter plot of the associations between acetate and 1461 microbial gene abundance based on three measures.

This study used different measures to infer the maximum rank between each microbial gene and 128 integrals. There are three types of quantitative relationships between microbial genes and integrals. They respectively are (I) Linear correlation is stronger than the non-linear correlation. (II) Non-linear correlation is stronger than the linear correlation. (III) Non-linear and linear correlations are both strong. (IV) Both linear and nonlinear are both weak. Of the 187008 pairs of final rankings, 60.1% of the rankings came from linear correlation, and 38.9% of the rankings came from mutual information (Fig.12). Furthermore, 1840 final rankings are linear equal to nonlinear, accounting for 1% of the total (Fig.12), which including both the III type and IV type of association between microbial genes and integrals. Among microbial genes, 70% ranked in the range of 95-128, which indicates their close associations.

The relationship between acetate and 67 microbial genes mapped to the methane metabolism pathway is shown in Fig. 13. Of these microbial genes, 43 had strong linear correlations with acetate, and 13 genes had non-linear correlations with acetate. The linear and nonlinear associations between 11 microbial genes and acetate were equally ranked. The microbial genes that were non-linearly associated with acetate were directly involved in the corresponding enzymatic reactions (Fig.13).

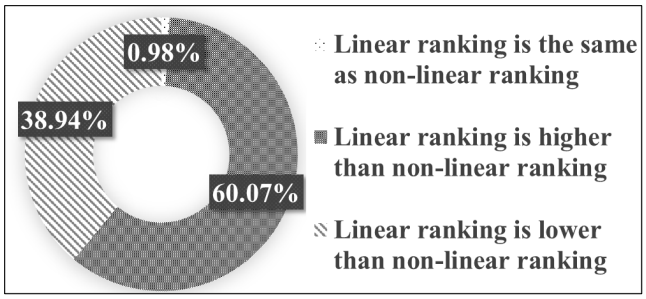


Fig. 12. Composition of relationships in the final rankings.

### C. *Inference of the integrals to the corresponding metabolites*

To infer the association between integrals and metabolites, we mapped 1461 microbial genes in the KEGG database to obtain information about the metabolites in the corresponding reactions. To demonstrate the feasibility of the study, we used acetate, with known spectral integral, as an example (Fig.14). Seven microbial genes (i.e., K00128, K00925, K01738, K01438, K01740, K12410, and K01895) associated with acetic acid (C00033) were found. In this study, the top 10 integrals of the seven microbial genes were selected to enter the network. ButyrateCH3.1 associated with six of the microbial genes, acetate associated with five of the microbial genes, and three integrals associated with the four microbial genes. The remaining integrals were associated with no more than two microbial genes. Acetate has strong non-linear rankings with three of the microbial genes (i.e., K01895, K00128 and K00925), and was strongly linearly associated with K01738. Notably, it has been observed that K01740 exhibits a strong link with acetate in terms of both linear and nonlinear association.

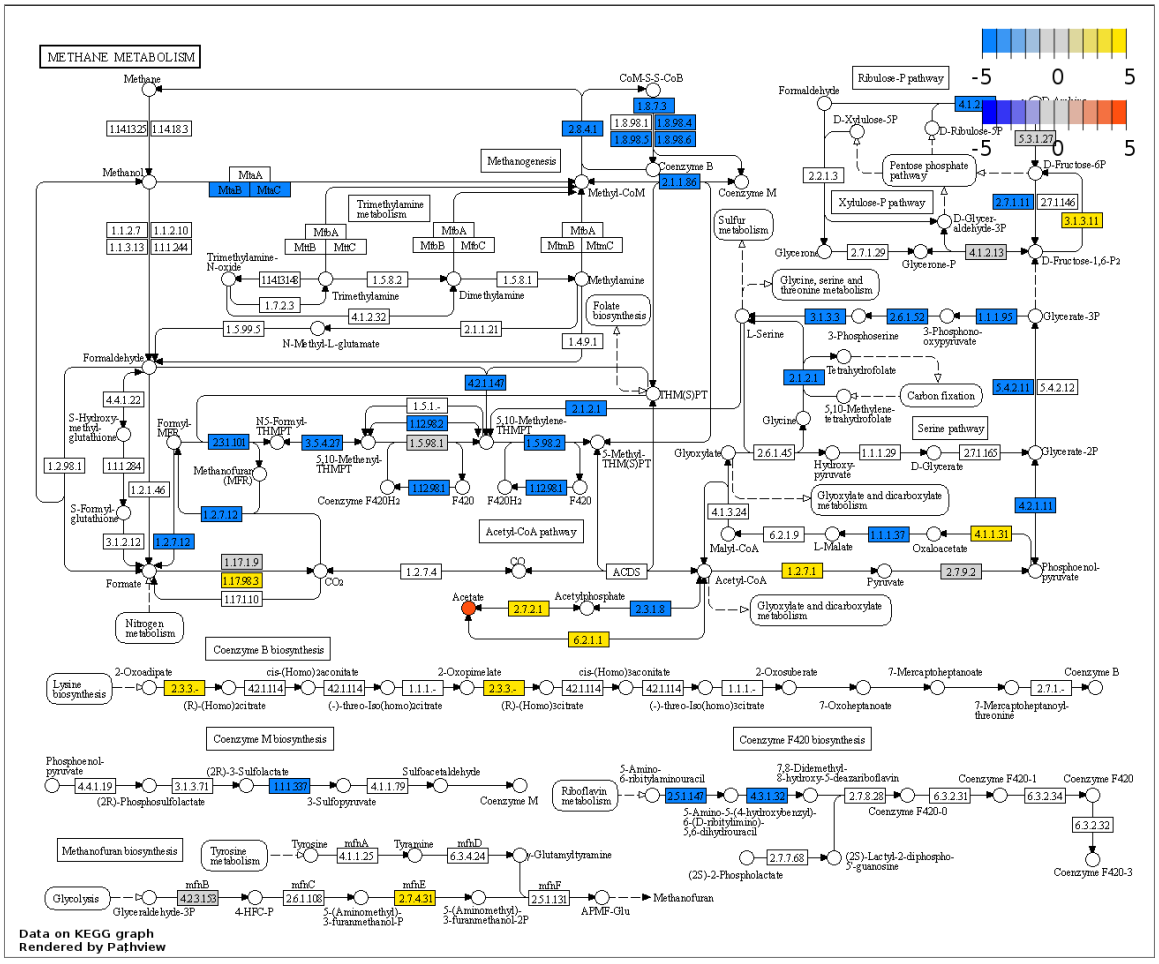


Fig. 13. The final ranking relationship between acetate and microbial genes involved in the methane metabolism pathway (map00680). (Orange node represented acetate. Blue boxes represented microbial genes that are strongly linearly associated with acetate. Yellow box represented microbial genes had strong non-linear associations with acetate. Grey boxes represented the microbial genes had equally linear and non-linear associations with acetate. White boxes indicated that no microbial genes were mapped.)

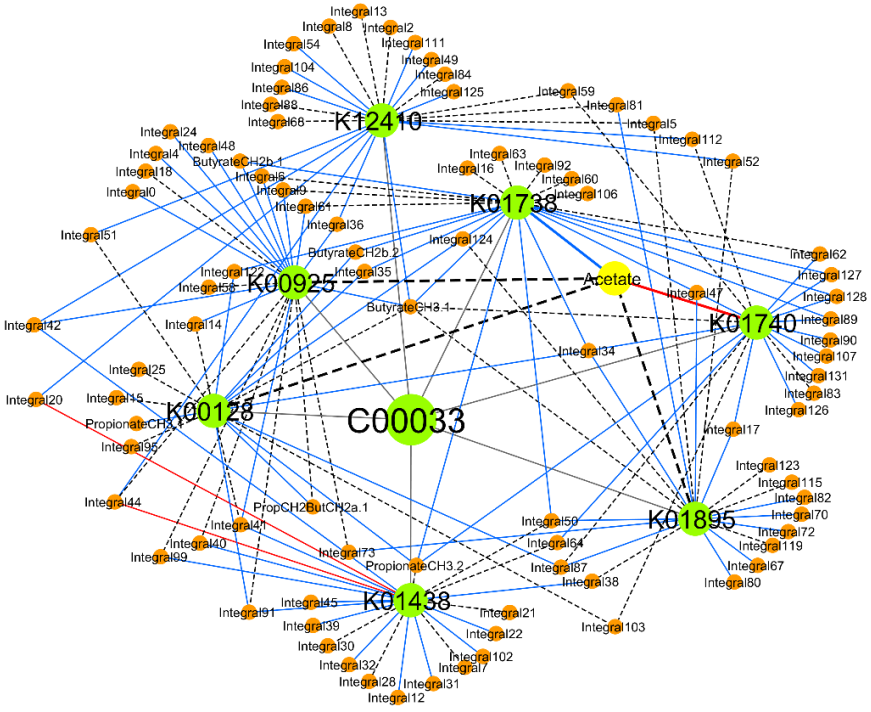


Fig. 14. Compound-Gene-Integral inference network of acetate (Nodes represent KEGG compounds, microbial genes, and integrals. Edges of black dot line and solid blue line indicate non-linear and linear rankings respectively. Red edges represent non-linear and linear equal strong rankings. Black edges represent that microbial genes and compounds involved in the same enzymatic reaction.)

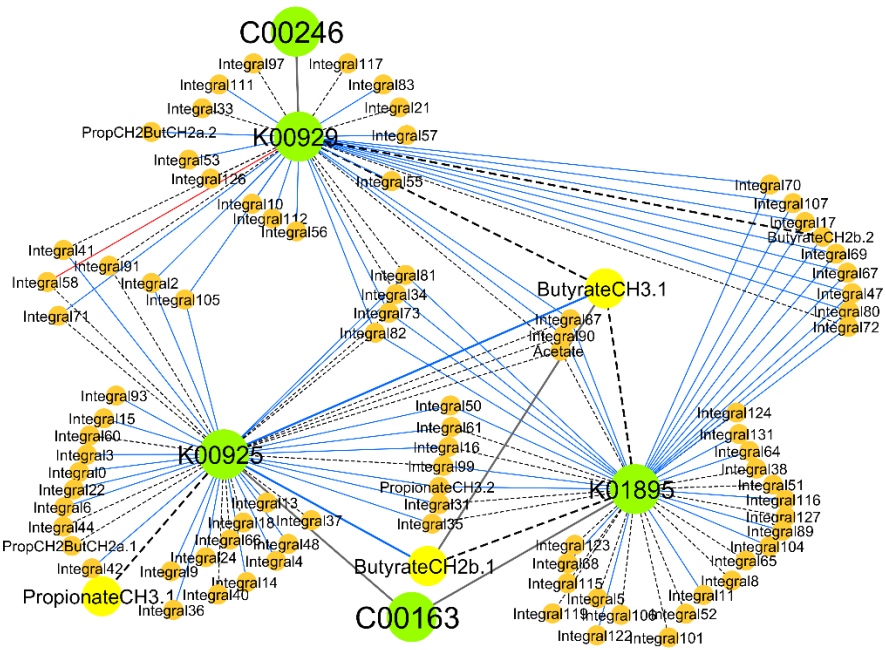


Fig. 15. Compound-Gene-Integral inference network of butyrate and propionate. (Nodes represent KEGG compounds, microbial genes, and integrals. Edges of black dot line and solid blue line indicate non-linear and linear rankings respectively. Red edges represent non-linear and linear equal strong rankings. Solid black edges represent that microbial genes and compounds involved in the same enzymatic reaction.)

Similarly, propionate and butyrate, which are essential metabolites in the rumen, were also shortlisted (Fig.15). K00925 and K01895 are involved in the propionate reaction in the same way as acetate. It can be seen from Fig. 15 that PropionateCH3.1 is included in the top 30 integrals, which should be propionate (C00163). PropionateCH3.1 only has a non-linear association with K00925. The rank of PropionateCH3.1 for K01895 is 69.

Regarding butyrate, there was only K00929 directly involved in the corresponding reaction. ButyrateCH3.1 and ButyrateCH2b.1 were inferred as one metabolite. Moreover, they were identified as butyrate (C00246) using the experimental methodology and was also included in the network.

## IV. Discussion and Conclusion

The identification of metabolites in non-targeted metabolomics usually relies on searching the spectral database and matching signal peaks by software (i.e., Chenomx) [26]. Because of the limitations of NMR technology and the imperfection of professional rumen metabolite databases, the annotation of NMR integrals in the quantitative metabolomics remains a challenging task [27]. This study developed a knowledge-driven network-based analytical framework for the inference of rumen metabolites.

Each compound in the NMR analysis can include more than one integrals which represent different hydrogen functional groups [26]. The first step in this study is the identification of the integrals belongs to same metabolite. The underlying traits of the integral value proved that although the digital distributions of the integrals are different, the integrals from the same metabolite have a significant correlation.

The co-occurrence network method is widely used in research on complex biological networks [28]. The multiple testing and bootstrap steps of the co-occurrence network method have been proven to effectively reduce false associations caused by the relative abundance of genes [29]. Similar to relative gene abundance, relative integrals are also composition type data. This research combined the Pearson linear correlation and mutual information to construct an integral network to capture their associations. Then Markov clustering successfully generated integral clusters which could be different potential metabolites. Part potential compounds have been verified by traditional NMR spectral analysis software (i.e., ButyrateCH3.1 and Butyrate2b.1).

Furthermore, this study found that the distributions of correlation between the integrals of same compound and other integrals are similar, and there is a functional relationship between the two groups of correlation coefficients. The dissimilarity network based on the Euclidean distance resulted in correlation coefficients between the integrals that are highly similar to Markov clustering. Although the analysis needs further verification by experiments, the analysis of two different procedures strongly supports each other by similar results of the integral clusters. The dissimilarity network seems to be more stringent than the co-occurrence based Markov clustering because it only produces a limited number of integral pairs.

Metabolites are produced under the synergy of microbial genes and are affected by the host, diet and environment, and are quickly consumed by secondary metabolic utilization [32]. The close connection between biochemical processes leads to a complicated relationship between metabolites and microbial genes. This study showed there are coexisting linear and non-linear relationships between metabolites and microbial genes. The research results found that the association between microbial genes and metabolites are different in the overall numerical range. For example, the maximum of correlations between K03294 and all integrals is smaller than the case of K00145. However, this does not indicate that most integrals have

closer biological relationships with K00145 than K03294. May be due to changes in the rumen environment caused by the functions of specific microbial genes, such as changing the overall metabolic biochemical response by influencing the pH [31] by dietary intervention. The challenges of the analysis were that microbial genes do not directly act on metabolites, and the production, transformation, and absorption of metabolites often occurs very rapidly [30]. It is necessary to take the overall situation between microbial genes and integrals into account when comparing the linear and nonlinear associations [33]. The ranking method proposed in this study well balanced the differences between linear and non-linear correlation calculations.

Combinations of label-free metabolic profiling and multi-omics analysis lead to an increased resolution in the bottlenecks of metabolic identification. The inferential analysis framework has great potential to improve the efficiency of the identification of metabolites using NMR. Limitations of existing microbial gene and metabolite databases were discussed in this study. There are at least 50% microbial genes in our data found corresponding biochemical reactions and compounds. After verification by traditional spectroscopy, three primary rumen volatile fatty acids were successfully identified by this study. While encouraging results have been obtained, it is worth noting that the identification of metabolites using spectral integrals could be used as an additional measure for metabolite identification before getting more experimental verification.

## REFERENCES


[1] F. Saleem *et al.*, "The Bovine Ruminal Fluid Metabolome," *Metabolomics*, vol. 9, no. 2, pp. 360–378, Apr. 2013.

[2] E. Riekeberg and R. Powers, "New frontiers in metabolomics: from measurement to insight," F1000Res, vol. 6, Jul. 2017.

[3] S. Pasek, J.-L. Risler, and P. Brézellec, "The Role of Domain Redundancy in Genetic Robustness Against Null Mutations," Journal of Molecular Biology, vol. 362, no. 2, pp. 184–191, Sep. 2006.

[4] J. G. M. Pontes, A. J. M. Brasil, G. C. F. Cruz, R. N. de Souza, and L. Tasic, "NMR-based metabolomics strategies: plants, animals, and humans," Anal. Methods, vol. 9, no. 7, pp. 1078–1096, Feb. 2017.

[5] A.-H. Em was et al., "NMR Spectroscopy for Metabolomics Research," Metabolites, vol. 9, no. 7, p. 123, Jun. 2019.

[6] A. Sharma, N. S. R. Kumari, and E. Menghani, "BIOACTIVE SECONDARY METABOLITES: AN OVERVIEW," 2014.

[7] C. J. Hurst, R. L. Crawford, J. L. Garland, and D. A. Lipson, Manual of Environmental Microbiology. American Society for Microbiology Press, 2007.

[8] R. Dunkel and X. Wu, "Identification of organic molecules from a structured database using proton and carbon NMR analysis results," Journal of Magnetic Resonance, vol. 188, no. 1, pp. 97–110, Sep. 2007.

[9] B. Fürtig, C. Richter, J. Wöhnert, and H. Schwalbe, "NMR Spectroscopy of RNA," ChemBioChem, vol. 4, no. 10, pp. 936–962, 2003.

[10] A. Sharma, N. S. R. Kumari, and E. Menghani, "BIOACTIVE SECONDARY METABOLITES: AN OVERVIEW," 2014.

[11] M. D. Auffret et al., "Identification, Comparison, and Validation of Robust Rumen Microbial Biomarkers for Methane Emissions Using

[12] Diverse Bos Taurus Breeds and Basal Diets," Frontiers in Microbiology, vol. 8, Jan. 2018.

[13] R. Roehe et al., "Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance," PLOS Genetics, vol. 12, no. 2, p. e1005846, Feb. 2016.

[14] R. J. Wallace et al., "The rumen microbial metagenome associated with high methane production in cattle, " BMC Genomics, vol. 16, no. 1, Dec. 2015.

[15] W. Poole, D. L. Gibbs, I. Shmulevich, B. Bernard, and T. A. Knijnenburg, "Combining dependent *P*- values with an empirical adaptation of Brown's method," *Bioinformatics*, vol. 32, no. 17, pp. i430–i436, Sep. 2016.

[16] S. Brohée and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 488, Nov. 2006.

[17] A. J. Enright, "An efficient algorithm for large-scale detection of protein families," Nucleic Acids Research, vol. 30, no. 7, pp. 1575–1584, Apr. 2002.

[18] S. van Dongen, "A New Cluster Algorithm for Graphs," p. 42.

[19] B. Lacevic and E. Amaldi, "Ectropy of diversity measures for populations in Euclidean space," *Information Sciences*, vol. 181, no. 11, pp. 2316–2339, Jun. 2011.

[20] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes, " Nucleic Acids Res, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[21] K. Faust and J. Raes, "CoNet app: inference of biological association networks using Cytoscape," *F1000Res*, vol. 5, Oct. 2016.

[22] P. Shannon et al., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, " Genome Res., vol. 13, no. 11, pp. 2498–2504, Jan. 2003.

[23] J. H. Morris *et al.*, "clusterMaker: a multi-algorithm clustering plugin for Cytoscape," *BMC Bioinformatics*, vol. 12, no. 1, p. 436, Dec. 2011.

[24] W. Luo and C. Brouwer, "Pathview: an R/Bioconductor package for pathway-based data integration and visualization, " Bioinformatics, vol. 29, no. 14, pp. 1830–1831, Jul. 2013.

[25] M. Wang et al., "Understanding the relationships between rumen microbiome genes and metabolites to be used for prediction of cattle phenotypes, " BIBE 2019 von Chengyu Liu | ISBN 978-3-8007-5026-9 | Bei Lehmanns online kaufen - Lehmanns.de." [Online]. [Accessed: 12-Oct-2019].

[26] D. S. Wishart, "Quantitative metabolomics using NMR," *TrAC Trends in Analytical Chemistry*, vol. 27, no. 3, pp. 228–237, Mar. 2008.

[27] S. Kostidis, R. D. Addie, H. Morreau, O. A. Mayboroda, and M. Giera, "Quantitative NMR analysis of intra- and extracellular metabolism of mammalian cells: A tutorial," *Analytica Chimica Acta*, vol. 980, pp. 1–24, Aug. 2017.

[28] K. Faust *et al.*, "Microbial co-occurrence relationships in the human microbiome," *PLoS Comput. Biol.*, vol. 8, no. 7, p. e1002606, 2012.

[29] J. Friedman and E. J. Alm, "Inferring Correlation Networks from Genomic Survey Data," *PLoS Comput Biol*, vol. 8, no. 9, p. e1002687, Sep. 2012.

[30] S. A. Goldansaz, A. C. Guo, T. Sajed, M. A. Steele, G. S. Plastow, and D. S. Wishart, "Livestock metabolomics and the livestock metabolome: A systematic review, " PLOS ONE, vol. 12, no. 5, p. e0177675, May 2017.

[31] R. Laaksonen et al., "A Systems Biology Strategy Reveals Biological Pathways and Plasma Biomarker Candidates for Potentially Toxic Statin-Induced Changes in Muscle, " PLOS ONE, vol. 1, no. 1, p. e97. Dec. 2006.

[32] B. Worley and R. Powers, "Multivariate Analysis in Metabolomics," Current Metabolomics, vol. 1, no. 1, pp. 92-107(16), 2013.

[33] J. Xi, Y. Niu, and L. Liu, "Multivariate Phase Space Reconstruction Based on Combination of Nonlinear Correlation Degree and ICA," in *Future Communication, Computing, Control, and Management: Volume 2*, Y. Zhang, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 465–472.

[34] H. Wang, H. Zheng, R. J. Dewhurst, and R. Roehe, "Microbial co-presence and mutual-exclusion networks in the Bovine rumen microbiome," in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, Nov. 2017, pp. 114–119, doi: 10.1109/BIBM.2017.8217635.

[35] D. Marion, "An Introduction to Biological NMR Spectroscopy," *Mol Cell Proteomics*, vol. 12, no. 11, pp. 3006–3025, Nov. 2013.

[36] J. Townend, Practical Statistics for Environmental and Biological Scientists. John Wiley & Sons, 2013.