Predicción de Metano en Vacuno de Carne con Microbioma y Algoritmos de Aprendizaje Automático

S. N. Saez Torillo[1], T. Q Nguyen[2], J. Lima[2], R.r Roehe[2], M. Martínez-Álvaro[1]*

[1] Instituto de Ciencia Animal, Universitat Politècnica de València (UPV), 46022, Valencia, España.
[2] Scotland's Rural College (SRUC), West Mains Road, EH9 3JG, Edinburgh, UK.
** Corresponding autor: snsaetor@posgrado.upv.es*

Resumen

La industria cárnica tiene un interés urgente en desarrollar estrategias para mejorar la eficiencia de la producción de carne y al mismo tiempo minimizar las emisiones de metano. La selección de animales para reducir sus emisiones requiere una recopilación de datos a gran escala con una precisión aceptable a costos asequibles. Los datos metagenómicos podrían servir como indicador para la predicción indirecta de las emisiones de metano, pero para que los modelos sean de aplicación general, deben validarse externamente. En este estudio se propone un modelo de predicción/clasificación (5 categorías) de diferentes fenotipos de metano utilizando datos de composición de microbiota ruminal de 283 animales utilizando diferentes algoritmos de *machine learning*. Los animales fueron alimentados con diferentes dietas, por lo que hubo que corregir este efecto tanto en el microbioma como en los fenotipos de metano. Nuestros resultados preliminares mostraron que la precisión de las predicciones/clasificaciones fue muy buena en los sets de entrenamiento. El modelo Random Forest fue el que mejor predijo el subset de entrenamiento en predicción ($R^2 \geq 0.95$), mientras que XGboost tuvo un mejor rendimiento en clasificación de metano en general (precisión ≥ 0.97). Sin embargo, el rendimiento de los modelos prediciendo/clasificando un subset de datos externo fue en general baja (la R^2 máxima fue 0.07±0.05, y la precisión de clasificación máxima fue de 0.51±0.05). Aumentar la cantidad de datos y optimizar aún más el procedimiento para evitar el sobreajuste, explorar patrones ocultos dentro de los animales o probar otros algoritmos de aprendizaje automático podrían ser posibles estrategias para mejorar los rendimientos de estos modelos predictivos o de clasificación.

Keywords: Metano, Machine Learning, Rumen, Microbiota, validación externa

Introducción

En 2023, la concentración de metano atmosférico ha aumentado en un 160% con respecto a la era pre-industrial (NOAA, 2024). La actividad agrícola es responsable de aproximadamente el 50-60% de las emisiones globales de metano, de las cuales, la fermentación entérica del ganado contribuye en un ~ 32% (IPCC, 2023). Se espera que en 2050 la población mundial supere los 9 mil millones y el consumo de carne vacuno aumente en un 153% con respecto a 2010 (FAO, 2011), principalmente en países en vía de desarrollo. La industria ganadera tiene un gran interés en desarrollar estrategias para mejorar la eficiencia de producción de carne y a la vez minimizar sus emisiones de metano.

La mejora genética es una estrategia atractiva para lograr animales neutros en carbono, ya que la respuesta obtenida es permanente y acumulativa en el tiempo. Sin embargo, para llevarse a cabo, se requiere una gran cantidad de datos. La medición directa del metano producido por los animales utilizando cámaras de respiración se considera el método estándar debido a su precisión, pero no es una alternativa viable en la producción a gran escala debido a que es económicamente muy costoso (Hammond et al., 2016). Otros métodos de medición directos, como los trazadores de hexafluoruro de azufre o las estaciones *Green-feed* son más económicos, pero menos precisos o dependen de las visitas voluntarias de los animales al alimentador (Dressler et al., 2024). Por lo tanto, el desarrollo de métodos indirectos para estimar el metano basados en modelos matemáticos que utilizan *proxies* fáciles de medir como predictores es una opción atractiva. Ya se han desarrollado varios modelos basados en variables como la ingesta de alimento, caracteres productivos (Niu et al., 2018; van Lingen et al., 2019), aunque están basados en factores de emisión promedios que ignoran la variabilidad individual de los animales. La disminución de los costos de obtención de fenotipos intermedios u "-ómicas" permite su uso como *proxies* para predecir el metano. Entre ellas, el microbioma podría tener un gran potencial ya que el metano

emitido por los rumiantes se produce como un subproducto de la fermentación microbiana de los alimentos en el rumen (Tapio et al., 2017). Investigaciones previas han mostrado que la precisión de predicción del microbioma (R^2) para el metano es alrededor de **0.2** utilizando diferentes algoritmos (Ross et al., 2013; Wallace et al., 2015). Al igual que con la predicción genómica, las ecuaciones basadas en datos metagenómicos podrían ser específicas para poblaciones de una raza y dieta en particular, u otros factores ambientales, y pocos estudios proporcionan una validación externa de sus modelos de predicción ajustados con datos diferentes razas, dietas y sistemas de manejo. El objetivo de este estudio es desarrollar ecuaciones de predicción para las emisiones de metano utilizando abundancias de géneros y genes microbianos ruminales y evaluar su rendimiento mediante validación externa, usando datos de animales de diferentes razas y alimentadas con diferentes dietas tras una corrección estadística de dichos efectos. Para abordar este desafío, se exploró el uso de diferentes algoritmos de aprendizaje automático.

Materiales y métodos

Animales, medición de metano y metagenómica.

Los datos se obtuvieron de 283 novillos de diferentes razas (cruce rotacional de razas Aberdeen Angus (n=76) y Limousin (n=72), cruces Charolais (n=68) y raza pura Luing (n=67)) alimentados con dos dietas basales que constan de proporciones de forraje:concentrado 480:520 y 80:920 denominadas forraje (n=182) y dietas concentradas (n=101) respectivamente. Estos animales participaron en diferentes experimentos (Duthie et al., 2016, 2017, 2018; Rooke et al., 2014) realizados durante 5 años (2011, 2012, 2013, 2014 y 2017) en la misma granja y bajo las mismas condiciones de hospedaje. El experimento fue aprobado por el Comité de Experimentación Animal de SRUC y se llevó a cabo siguiendo los requisitos de la Ley de Animales (Procedimientos Científicos) del Reino Unido de 1986. Las emisiones de metano se midieron individualmente durante 48 h dentro de seis cámaras de respiración de circuito abierto indirecto (Rooke et al., 2014). Dentro de cada experimento, los animales fueron asignados a las cámaras de respiración en un diseño aleatorio según raza y dieta. Los animales fueron alimentados una vez al día y se registró la ingesta diaria de materia seca (DFI). Las emisiones de metano se expresaron como producción de metano diaria (MP, g CH_4/día); rendimiento de metano (MY, g CH_4/kg DFI) y metano residual (RM, g CH_4/dia) (Herd et al., 2014), calculado como el residuo del modelo:

$$RM\ \left(\frac{g}{dia}\right) = \beta_0 \mp\ \beta_1 \cdot Peso \mp \beta_2 \cdot DFI + \mathrm{e}$$

Donde el Peso es el tomado al ingreso de la cámara de metano y el DFI es la ingesta de materia seca diaria durante el experimento.

Posteriormente, se recolectaron 5 mL de líquido ruminal inmediatamente después del sacrificio, y se mezclaron con 10 mL de PBS y glicerol (87%). La extracción de ADN se realizó siguiendo el protocolo de Yu y Morrison (Zhongtang Yu and Mark Morrison, 2004). Las librerías de ADN Illumina TruSeq se prepararon a partir de ADN genómico y se secuenciaron en sistemas Illumina HiSeq 4000. Para la anotación filogenética, las lecturas se alinearon con las bases de datos Hungate 1000 (Seshadri et al., 2018) y RefSeq (Pruitt et al., 2007) utilizando el software Kraken (Wood & Salzberg, 2014). Para la anotación funcional, las lecturas se alinearon con la base de datos de Kyoto Encyclopedia of Genes and Genomes Orthologue utilizando el programa KOunt (Mattock et al., 2023). De los 1178 géneros y 7976 genes microbianos identificados, seleccionamos 1136 géneros y 3632 genes microbianos presentes en al menos el 70% de los animales. Los ceros restantes se imputaron basándose en un método multiplicativo bayesiano. Los géneros microbianos y las abundancias de genes se transformaron utilizando la transformación log-ratio centrada (1136 clr-MT) y aditiva (3631 alr-MG), respectivamente, utilizando el gen microbiano K01783 como referencia en alr, elegido usando los criterios definidos en (Greenacre et al., (2021).

Métodos estadísticos.

Se realizó un análisis exploratorio del microbioma usando un análisis componentes principales (PCA) para evaluar patrones entre los datos. Al analizar los efectos, se descubrió que la dieta afectaba tanto el metano como la microbiota (*p-valor*=0.002) por lo que se procedió a hacer una corrección lineal de este efecto en cada una de las variables microbianas y metano (Figura 1). Los 3 fenotipos de metano fueron ajustados como variable dependiente de forma

continua y categórica, dividiendo los animales en 5 grupos de similar número de individuos para cada fenotipo de metano, tras corregirlo por el efecto dieta (Tabla 1). Los 1136 clr-MT y 3631 alr-MG se utilizaron de manera conjunta como variables predictoras en todos los algoritmos (4767 variables microbianas). Se utilizaron 3 algoritmos de predicción/clasificación: *Partial Least Square* (PLS) y *Discriminant Analysis* PLS (PLS-DA) usando el paquete de R mixOmics, (Kim-Anh Le Cao et al., 2016), *Random Forest* (RF) usando el paquete de R *randomForest* (Andy Liaw & Matthew Wiener, 2002) y *eXtreme Gradient Boosting* (XGB), usando el paquete de R *xgboost* (Chen & Guestrin, 2016). Además, se testó la utilización de métodos combinados de PCA+RF y PCA+XGB para disminuir la dimensionalidad de las variables predictoras antes de ajustar los algoritmos.

Tabla 1. Distribución de individuos por categorías, con su media y desviación típica (sd).

	CATEGORIA	**Nº individuos**	**Media ± sd**
MY [g CH_4/kg DFI]	1	53	9.89 ± 0.96
	2	44	12.1 ± 0.60
	3	60	13.9 ± 0.52
	4	53	15.7 ± 0.55
	5	61	19.1 ± 2.2
	TOTAL	271	14.3 ±3.4
MP [g CH_4/día]	1	54	89.2 ± 17
	2	85	130 ±9.6
	3	86	162 ±10
	4	36	197 ±10
	5	19	277 ±56
	TOTAL	280	151 ± 51
RM [g CH_4/día]	1	32	-59.2 ± 9.8
	2	75	-28 ± 9.6
	3	84	0.923 ± 8.9
	4	39	29.7 ±7.9
	5	27	102 ± 59
	TOTAL	257	0 ± 48

El rendimiento de todos los algoritmos se evaluó mediante validación externa. Primero, los datos se dividieron en dos subsets, TRAIN (4/5) y TEST (1/5). El TRAIN se utilizó para optimizar los hiperparámetros de cada uno de los métodos por validación cruzada con 5 *k-folds* y 10 repeticiones. Para la clasificación de MP, MY y RM seleccionamos la combinación de hiperparámetros que maximizó el área bajo la curva ROC (AUC), mientras que, para su predicción, elegimos los hiperparámetros que maximizaban el R^2 de la predicción en los conjuntos de prueba. Con los hiperparámetros optimizados, se ajustó un modelo con el subset de datos TRAIN. Este modelo se utilizó para predecir/clasificar el TRAIN y el TEST y se calculó el AUC para clasificación y el R^2 para las predicciones continuas. Además, para estimar un rango del rendimiento de los diferentes algoritmos, se realizaron 100 iteraciones dividiendo cada vez la base de datos en diferentes TRAIN y TEST, y se predijeron ambos subsets de datos.

Resultados y discusión

El análisis PCA del microbioma mostró dos grupos de animales que correspondían a las diferentes dietas (Figura 1A). Otros efectos, como la raza y el año de experimentación, no causaron ningún patrón o efecto significativo en el microbioma. Antes de ajustar los algoritmos, se realizó una corrección lineal por dieta de cada una de las 4767 variables (Figura 1B).

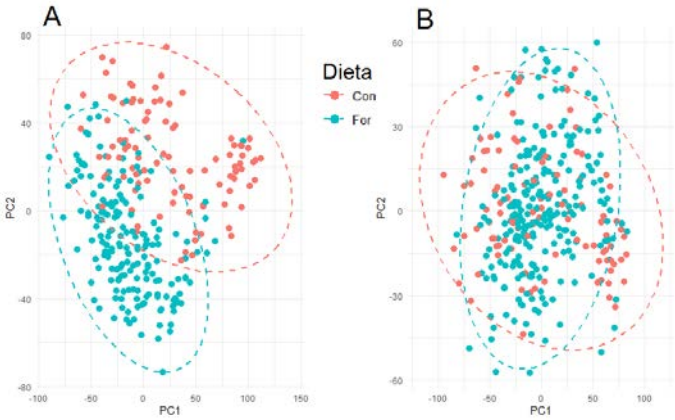

Figura 1. Análisis de componentes principales del microbioma (clr-MT y alr-MG). **A.** Abundancias crudas. **B.** Corregidas por dieta.

Tabla 2. Rendimiento de los distintos algoritmos clasificación y predicción con los datos de entrenamiento (TRAIN) y los datos de prueba (TEST).

		PREDICCIÓN		CLASIFICACIÓN	
		R^2_{train}	R^2_{test}	$Precisión_{train}$	$Precisión_{test}$
MY [g CH_4/kg DFI]	PLS (PLS-DA)	0.29±0.28	0.02±0.02	0.82±0.05	0.21±0.04
	RF	**0.97±0.005**	**0.06±0.04**	0.62±0.03	0.20±0.04
	XGB	0.40±0.06	0.06±0.04	**0.99±0.01**	**0.32±0.05**
	PCA-RF	0.95±0.01	0.03±0.02	0.71±0.12	0.17±0.03
	PCA+XGB	0.79±0.03	0.02±0.03	0.78±0.03	0.21±0.04
MP [g CH_4/día]	PLS (PLS-DA)	0.18±0.11	0.04±0.02	0.90±0.02	0.29±0.04
	RF	**0.98±0.003**	**0.07±0.05**	0.51±0.02	**0.35±0.05**
	XGB	0.91±0.01	0.06±0.04	**0.97±0.01**	0.21±0.04
	PCA-RF	0.96±0.01	0.04±0.03	0.58±0.05	0.25±0.05
	PCA+XGB	**0.99±0.001**	0.03±0.03	0.45±0.04	0.29±0.05
RM [g CH_4/día]	PLS (PLS-DA)	0.23±0.19	0.04±0.05	0.89±0.03	0.27±0.04
	RF	**0.95±0.01**	**0.05±0.05**	0.62±0.02	0.33±0.05
	XGB	0.88±0.02	0.05±0.05	0.73±0.03	**0.51±0.05**
	PCA+RF	0.91±0.01	0.02±0.02	0.60±0.04	0.32±0.07
	PCA+XGB	0.92±0.01	0.02±0.02	**0.98±0.01**	0.31±0.05

En general se obtuvieron buenos resultados para la predicción o clasificación del TRAIN, sin embargo, los resultados de validación externa (TEST) fueron mucho más bajas. Cabe resaltar los problemas de sobreajuste de los modelos, a pesar de la optimización de los mismos a través de validación cruzada y de ofrecer solamente parámetros que dieran lugar a algoritmos simples. En predicción, el mejor método fue RF de manera consistente para MY, MP y RM con $R^2 \geq 0.95$, aunque su capacidad predictiva en el TEST fue cercana a 0. En clasificación, XGB fue el algoritmo con mayor precisión de clasificación en el TRAIN para MY y MP (Precisión ≥ 0.97), y PCA+XGB para RM (Precisión ≥ 0.98). Sin embargo, la mejor precisión de clasificación en el TEST fue lograda por XGB en MY (0.32), RF en MP (0.35) y XGB en RM (0.51). Nuestros resultados son pesimistas en cuanto a la posibilidad de utilizar información metagenómica para obtener mediciones indirectas de emisiones de metano en ganado vacuno de diferentes dietas o razas, y resalta la importancia de validaciones externas a la hora de reportar el rendimiento de los algoritmos. Aumentar el conjunto de datos podría mejorar la comprensión de las conexiones entre la microbiota y la producción de metano a través de algoritmos de aprendizaje automático.

Referencias

Liaw, & Matthew Wiener. (2002). Machine Learning, 45(1), 5–32. doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). Proceedings of the ACM SIGKDD 785–794. doi.org/10.1145/2939672.2939785

Dressler, E. A., et al. (2024). Translational Animal Science, 8. doi.org/10.1093/tas/txae014

Duthie, C. A., et al. (2017). Animal, 11(10), 1762–1771. doi.org/10.1017/S1751731117000301

Duthie, C. A., et al. (2016). Animal, 10(5), 786–795. doi.org/10.1017/S1751731115002657

Duthie, C. A., et al. (2018). Animal, 12(2), 280–287. doi.org/10.1017/S175173111700146X

Food and Agriculture Organization of the United Nations (FAO). Retrieved February 29, 2024, from www.fao.org/3/i2373e/i2373e.pdf

Global Carbon and Other Biogeochemical Cycles and Feedbacks. (2023). In Climate Change 2021 – The Physical Science Basis (pp. 673–816). doi.org/10.1017/9781009157896.007

Global Monitoring Laboratory of the National Oceanic and Atmospheric Administration. doi.org/https://doi.org/10.25925/20231001

Greenacre, M., Martínez-Álvaro, M., & Blasco, A. (2021). F. in Microb., 12. doi.org/10.3389/fmicb.2021.727398

Hammond, K. J., et al. (2016). Animal Feed Science and Technology (Vol. 219, pp. 13–30). Elsevier B.V. doi.org/10.1016/j.anifeedsci.2016.05.018

Herd, R. M, et al. (2014). J. Anim. Sci, 92, 5267–5274. doi.org/10.2527/jas2014-8273

Kim-Anh Le Cao, et al. (2016). mixOmics: Omics Data Integration Project. R package (version 6.1.1.).

Mattock, J., et al. (2023). Bioinformatics. doi.org/https://doi.org/10.1093/bioinformatics/btad483

Niu, M., et al. (2018). Global Change Biology, 24(8), 3368–3389. doi.org/10.1111/gcb.14094

Pruitt, K. D., et al. (2007). Nucleic Acids Research, 35(SUPPL. 1). doi.org/10.1093/nar/gkl842

R Foundation for Statistical Computing. (2012). www.R-project.org/.

Rooke, J. A., et al. (2014). British Journal of Nutrition, 112(3), 398–407. doi.org/10.1017/S0007114514000932

Ross, E. M., et al. (2013). PLoS ONE, 8(9). doi.org/10.1371/journal.pone.0073056

Seshadri, R., et al. (2018). Nature Biotechnology (Vol. 36, Issue 4, pp. 359–367). doi.org/10.1038/nbt.4110

Tapio, I., et al. (2017). Journal of Animal Science and Biotechnology (Vol. 8, Issue 1). doi.org/10.1186/s40104-017-0141-0

van Lingen, H. J., et al. (2019). Agriculture, Ecosystems and Environment, 283. doi.org/10.1016/j.agee.2019.106575

Wallace, R. J., et al. (2015). The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics, 16(1). doi.org/10.1186/s12864-015-2032-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. doi.org/doi:10.1186/gb-2014-15-3-r46

Zhongtang Yu and Mark Morrison. (2004). BioTechniques , 36, 808–812.