Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Seleção de Genes da Vitamina B2 Utilizando o Algoritmo *splmm*

Lucas E. P. de Melo[1], Daniela C. R. de Oliveira [2]
UFSJ, São João del-Rei, MG

A riboflavina, também conhecida como vitamina B2 do complexo B, é obtida principalmente por meio do consumo de carnes e laticínios e desempenha um papel essencial como componente de duas coenzimas: flavina-adenina-dinucleotídeo (FAD) e flavina-mononucleotídeo (FMN) [1]. Essas coenzimas estão envolvidas em processos fundamentais do metabolismo energético, incluindo a produção de energia, a regulação do relógio biológico e o reparo do DNA. A deficiência de riboflavina pode levar a diversos problemas de saúde, como lesões na pele e na boca, queda de cabelo, disfunções reprodutivas e degeneração do fígado e do sistema nervoso [4].

Dessa forma, considerando o papel fundamental da riboflavina na manutenção do bem-estar humano, a seleção e manipulação de genes associados à sua produção representam uma estratégia promissora para otimizar a síntese dessa vitamina pelas células. Assim, utilizando a base de dados *riboflavinV100*, disponibilizado por [6], que contém os principais genes que influenciam a taxa de produção da vitamina B2 na bactéria unicelular *Bacillus subtilis* (presente no sistema digestivo humano), este estudo tem como objetivo ajustar um modelo linear misto (LMM, do inglês *Linear Mixed Model*) para selecionar os genes mais relevantes na produção de riboflavina.

Mais especificamente, como os dados têm um caráter longitudinal, foi considerado o seguinte modelo linear de efeito misto [5]:

$$y_{ij} = \sum_{k=1}^{100} \beta_k x_{ijk} + \beta_{101} x_{ij101} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, i = 1, ..., n, j = 1, ..., n_i, \tag{1}$$

onde a variável resposta y_{ij} representa o logaritmo da taxa de produção de riboflavina; as 100 covariáveis correspondem ao logaritmo do nível de expressão de 100 genes, além da covariável tempo ($x_{ijk}, k = 1, \ldots, 101$). Essa base de dados contém $n = 28$ tipos da bactéria *Bacillus subtilis*, medidos entre 2 e 4 vezes ao longo de 96 horas, totalizando $N = 71$ observações. Além disso, a variável resposta e todas as covariáveis foram padronizadas para terem média 0 e variância 1. O vetor de resposta, a matriz de planejamento correspondente aos efeitos fixos e o vetor de efeitos fixos possuem, respectivamente, as dimensões: $N \times 1$, $N \times p$ e $p \times 1$; já a matriz de planejamento dos efeitos aleatórios, vetor b de efeitos aleatórios, e o vetor ε de erros aleatórios possuem, respectivamente, as dimensões $N \times q$, $q \times 1$ e $N \times 1$. O modelo possui as seguintes suposições: $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ e $b \sim N_q(0, D)$.

A seleção dos genes foi realizada utilizando o método de regularização Lasso, a definição de um intervalo para o λ e o uso do Critério de Infomação Bayesiano (BIC) [3]. Destacamos que, como o número de covariáveis é maior que o tamanho amostral ($p > n$, caracterizando uma situação de alta dimensão), o método de regularização Lasso possui o papel de reduzir a dimensão das covariáveis. Essa etapa foi realizada com o auxílio do pacote *splmm* do R (versão 1.2.0) [7]. Para a verificação

[1] lucasepml@aluno.ufsj.edu.br
[2] daniela@ufsj.edu.br

das suposições de normalidade e avaliação do poder preditivo com as variáveis selecionadas pelo *splmm*, foi utilizado o pacote *lme4* do R [2].

Por fim, foi realizado um estudo de simulação com diversos cenários para avaliar as métricas de sensibilidade, especificidade, acurácia e erro quadrático médio, além da seleção das covariáveis associadas à produção de riboflavina. Os resultados indicaram que o algoritmo apresentou um bom desempenho nessas métricas, fornecendo valores iguais ou próximos de 1 e um erro próximo de zero.

O *splmm* selecionou 12 genes relevantes para a produção de riboflavina. Utilizando esse algoritmo para a seleção de variáveis, em conjunto com o pacote *lme4*, foi observado um bom desempenho na predição da vitamina, evidenciado pelo baixo valor do erro quadrático médio na previsão de y. Além disso, o modelo selecionado atendeu as suposições de normalidade, conforme analisado por meio dos gráficos quantil-quantil (QQ plots) dos efeitos aleatórios, e apresentou uma baixa variância residual.

Agradecimentos

Os autores desse trabalho foram apoiados pelo projeto de pesquisa nº APQ-04657-23 da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig).

Referências

[1] C. A. Abbas e A. A. Sibirny. "Genetic Control of Biosynthesis and Transport of Riboflavin and Flavin Nucleotides and Construction of Robust Biotechnological Producers". Em: **Microbiology and Molecular Biology Reviews** (2011). Aceito. DOI: 10.1128/mmbr.00030-10.

[2] D. M. Bates et al. **lme4: Linear Mixed-Effects Models using 'Eigen' and S4**. Online. Acessado em 10/03/2025, https://cran.r-project.org/web/packages/lme4/index.html.

[3] T. Hastie, R. Tibshirani e J. Friedman. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2009. ISBN: 978-0387848570.

[4] National Institutes of Health. **Riboflavin**. Online. Acessado em 10/03/2025, https://ods.od.nih.gov/factsheets/Riboflavin-HealthProfessional/.

[5] J. C. Pinheiro e D. M. Bates. **Mixed-Effects Models in S and S-Plus**. Springer, 2005. ISBN: 978-0-387-98957-0.

[6] J. Schelldorfer, P. Bühlmann e S. van de Geer. "Estimation for High-Dimensional Linear Mixed-Effects Models Using l1-Penalization". Em: **Scandinavian Journal of Statistics** (2011). Aceito. DOI: 10.1111/j.1467-9469.2011.00740.x.

[7] L. Yang, E. Sun e T. T. Wu. **splmm: Simultaneous Penalized Linear Mixed Effects Models**. Online. Acessado em 10/03/2025, https://cran.r-project.org/web/packages/splmm/index.html.