USO DE MACHINE LEARNING PARA ANÁLISE DE miRNAs ASSOCIADOS A QTLS NA ESPÉCIE *Bos taurus*

HADASSA GABRIELA ORTIZ[1]; FREDERICO SCHMITT KREMER[2]; VINÍCIUS FARIAS CAMPOS[3]

[1]Universidade Federal de Pelotas – hortizhadassa@gmail.com[1]
[2]Universidade Federal de Pelotas – fred.s.kremer@gmail.com[2]
[3]Universidade Federal de Pelotas – fariascampos@gmail.com[3]

1. INTRODUÇÃO

MicroRNAs (miRNAs) são moléculas de RNA não codificantes de aproximadamente 22 nucleotídeos responsáveis pela regulação da expressão gênica a nível pós-traducional. Essa regulação se dá por uma modulação negativa, a qual pode ser pela degradação do RNA mensageiro (mRNA) alvo ou pelo impedimento da tradução. Apesar de se ligarem por complementaridade ao mRNA alvo, é sabido que a rede de interação entre miRNAs e mRNAs é bastante complexa e que um único miRNA pode afetar diversos mRNAs (RICARTE FILHO, 2006). Já as QTLs são regiões cromossômicas específicas de um locus gênico que estão relacionadas de forma estatística com uma variação fenotípica, ou seja, são seções de DNA que apresentam correlação estatística com alguma característica visível. Alguns estudos apontam a presença de QTLs em regiões não codificantes do genoma, o que torna intrigante a busca por miRNAs nessas localizações (MILES e WAYNE, 2008). Além disso, tem-se evidências de que é possível tanto um miRNA regular mais de uma característica quanto uma característica ser regulada diversos miRNA (PENG et al., 2019).

Ademais, é possível dizer que o mapeamento de QTLs é o passo inicial em estudos que visam identificar genes e também polimorfismos responsáveis por características de interesse em animais. Um exemplo da aplicação desse conhecimento é o uso de genes e marcadores moleculares (e.g. SNPs) em programas de reprodução assistida para geração de animais com interesse econômico. A análise de QTLs se dá, majoritariamente, através de experimentos de bancada pela avaliação dos marcadores moleculares , e.g. SNPs e INDELs, através de técnicas como PCR , Microarray etc. Existem bancos de dados específicos para QTLs em animais, como por exemplo o Animal QTL db (https://www.animalgenome.org/cgi-bin/QTLdb/index). No caso da espécie *Bos taurus,* há uma meta-análise que descreve que grande parte do volume de informações sobre QTLs em domínio público refere-se a características relacionadas à produção e composição de leite (PAUSCH et al., 2016).

Com o avanço das técnicas moleculares, é possível obter mais informação de forma cada vez mais rápida. Desse modo, a quantidade de informações disponíveis em bancos de dados também aumentou. No caso da espécie *Bos taurus*, tem-se uma grande quantidade de informações genômicas em domínio público, justamente por essa ser uma espécie de interesse comercial. Alguns dos exemplos de bancos de dados públicos são Bovinemine (http://128.206.116.13:8080/bovinemine/begin.do), Ensembl (https://www.ensembl.org/), Animal GenomeDB (animalgenome.org) etc. Essa quantidade de dados comunitários nos permite explorar e analisar informações

que já foram coletadas para que seja possível adquirir novos conhecimentos a partir delas.

Contudo, lidar com dados em abundância não é uma tarefa tão simples, ainda mais em situações em que a relação entre os dados não está delimitada. Sendo assim, faz-se necessário o uso de ferramentas de inteligência artificial, mais especificamente do aprendizado de máquina não-supervisionado, para a realização de tal tarefa. O aprendizado de máquina não supervisionado é uma subárea do aprendizado de máquina que visa obter padrões, antes ocultos, em um conjunto de dados não rotulados. (GÉRON, 2019)

A partir disso, tem-se a possibilidade do uso dos dados já disponíveis e das ferramentas de aprendizado de máquina não supervisionado a fim de investigar tanto a presença de miRNAs em QTLs quanto investigar e obter novos insights sobre as associações de miRNAs e QTLS.

2. METODOLOGIA

Para a coleta dos dados sobre as QTLs e sobre os miRNAs descritos para a espécie *Bos taurus* foram utilizados os bancos de dados Ensembl (anotação UMD3.1.1) e BovineMine, respectivamente. Os dados foram coletados utilizando a linguagem de programação Python (https://www.python.org) e as bibliotecas Biopython (https://biopython.org), Intermine (https://intermine.readthedocs.io/en/latest/) e Pandas (https://pandas.pydata.org). Em seguida eles foram cruzados utilizando a biblioteca SQLite3 (https://docs.python.org/3/library/sqlite3.html), a fim de obter os miRNAs presentes nas QTLs descritas para *Bos taurus*, e passaram por uma etapa de limpeza para remoção de valores nulos.

Posteriormente, deu-se início à análise exploratória dos dados. A análise exploratória consistiu em avaliar algumas distribuições estatísticas e também em utilizar algoritmos de aprendizado de máquina não supervisionado para realizar análises de redução de dimensionalidade seguidas de métodos não lineares de clusterização. Para que as ferramentas de aprendizado de máquina não supervisionado pudessem ser utilizadas, as features “identificação do transcrito de miRNA” e “característica associada à QTL” foram relacionadas através da representação de 0 e 1, para ausência ou presença de cada miRNAs em cada uma das QTLs, respectivamente. O algoritmo utilizado para a análise de redução de dimensionalidade foi o UMAP e os escolhidos para as análises de clusterização foi o DBSCAN e HDBSCAN, as bibliotecas utilizadas foram UMAP (https://umap-learn.readthedocs.io/en/latest/) e Sci-Kit Learn (https://scikit-learn.org/stable/index.html), respectivamente. Antes da etapa de clusterização, os dados passaram por um pré-processamento para ajuste de escala. Ademais, os parâmetros utilizados no DBSCAN foram eps= 0.03 e eps=0.5. Por fim, após a obtenção dos clusters, os cinco miRNAs mais presentes em cada um deles foram obtidos.

3. RESULTADOS E DISCUSSÃO

O primeiro resultado refere-se à presença de 157 miRNAs dentro de 481 QTLs. Também foi verificado que as combinações desses miRNAs se deram da seguinte maneira: foram encontradas situações de um miRNA presente em diversas QTLs, de uma QTL relacionada a diferentes miRNAs e também de um único miRNA estar associado a uma única QTL. Esse resultado corrobora com o

que já foi descrito por Peng et al. (vol. 10,6 (2019): e1556). Outrossim, grande parte desses miRNAs estão presentes em QTLs relacionadas com a produção e composição de leite, porém como já mencionado, isso pode ter ocorrido devido à maior quantidade de dados dessas QTLs em domínio público.

Durante a aplicação dos algoritmos de redução de dimensionalidade, foi possível perceber que os dados não apresentaram característica linear e portanto o algoritmo UMAP foi escolhido para prosseguir com as análises. No modelo de clusterização utilizando o parâmetro eps= 0.03 (Figura 1), os dados apresentaram 7 clusters distintos e também a presença de outliers (cluster = -1). Outra característica que esses dados apresentaram foi uma maior homogeneidade e maior densidade. Ao utilizar o modelo com o parâmetro eps=0.5 (Figura 2), foram encontrados apenas 2 clusters diferentes e ausência de outliers, porém esses clusters apresentaram menor homogeneidade e menor densidade. Em síntese, é possível observar que os dados demonstram não só a existência de 2 grupos muito distantes entre si mas também revelam a presença de subdivisões, indicando que as amostras têm características distintas que podem ser exploradas.

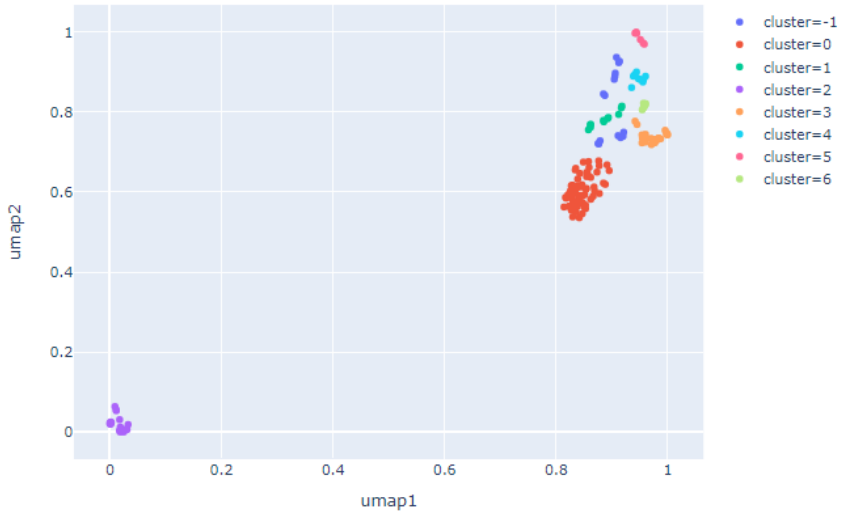

Figura 1

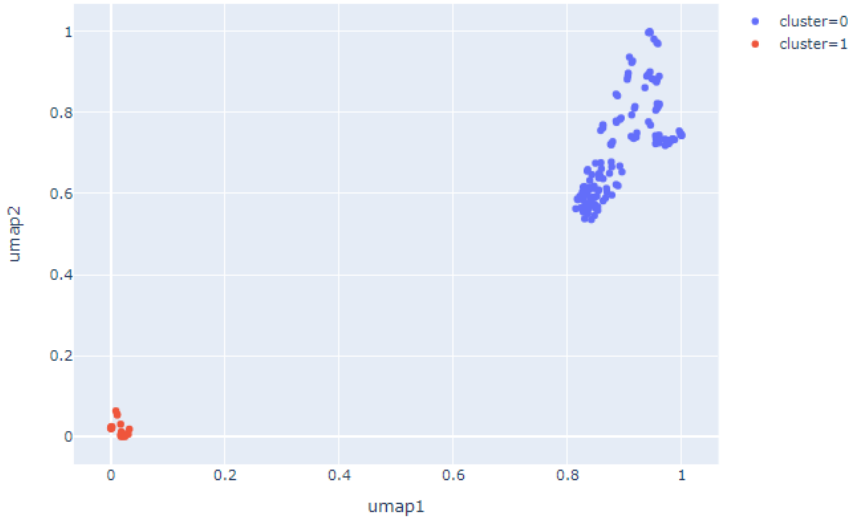

Figura 2

4. CONCLUSÕES

A partir dos resultados obtidos, foi possível obter informações acerca das possíveis relações entre miRNAs e QTLs. Além do mais, os resultados encontrados também correspondem com o que já foi descrito na literatura. Por último, tem-se a perspectiva da realização de uma análise de enriquecimento funcional a partir dos genes parentais e genes alvos dos miRNAs mais presentes em cada um dos clusters encontrados, com o intuito de avaliar a relação entre os padrões encontrados e os processos biológicos afetados.

5. REFERÊNCIAS BIBLIOGRÁFICAS

SHAMIMUZZAMAN, Md et al. Bovine Genome Database: new annotation tools for a new reference genome. **Nucleic acids research**, v. 48, n. D1, p. D676-D681, 2020.

MILES, C. & WAYNE, M. (2008) Quantitative trait locus (QTL) analysis. Nature Education 1(1):208.

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. O'Reilly Media, Inc., 2019.

RICARTE FILHO, Júlio C.M.; KIMURA, Edna Teruko. MicroRNAs: nova classe de reguladores gênicos envolvidos na função endócrina e câncer. **Arq Bras Endocrinol Metab**, São Paulo , v. 50, n. 6, p. 1102-1107, Dec. 2006

PAUSCH, Hubert, EMMERLING, Reiner, SCHWARZENBACHER, Hermann et al. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. **Genet Sel Evol** 48, 14 (2016).

PENG, Ting et al. “MicroRNAs meet with quantitative trait loci: Small powerful players in regulating quantitative yield traits in rice.” **Wiley interdisciplinary reviews. RNA** vol. 10,6 (2019): e1556.