Universidade Federal do Pará
Instituto de Ciências Biológicas - ICB
Programa de Pós-graduação em Genética e Biologia Molecular

Professor: Dr. Gilderlanio Santana de Araújo

Aluno: Victor Benedito Costa Ferreira

Projeto Final das Disciplinas Bioinformática Aplicada à Genômica e Machine Learning

Análise das interações genéticas entre patógenos resistentes a antibióticos e os hospedeiros humanos.

A Organização Mundial da Saúde (OMS) reportou no ano de 2020 que a resistência antimicrobiana (AMR) é uma das 10 principais ameaças à saúde pública e ao desenvolvimento global enfrentadas pela humanidade no séc. 21. Considerando essa preocupação, o monitoramento ambiental e clínico de AMR já está sendo realizado por diversos grupos de pesquisa ao redor do mundo, e, inclusive, no Brasil.

As infecções por microrganismos resistentes podem ser responsáveis por infecções comuns na comunidade, como as do trato respiratório e diarreia, contudo, as infecções da corrente sanguínea causadas por organismos gram-negativos resistentes podem atingir uma taxa de mortalidade de 80 a 85%. Atualmente, as taxas de infecções causadas por bactérias resistentes têm aumentado constantemente, com pelo menos 23.000 mortes por infecções bacterianas multirresistentes sendo relatadas por ano nos Estados Unidos, e o número de pessoas afetadas por essas infecções anualmente é estimado em 2 milhões.

As interações genéticas entre os patógenos resistentes a antibióticos e o genoma humano envolvem diversos mecanismos complexos. Genes de resistência a antibióticos (ARGs) podem ser codificados em elementos genéticos móveis (MGEs) em bactérias, facilitando a transferência horizontal de genes para os microrganismos que compõem a microbiota humana.

O uso indevido de antibióticos exerce pressão seletiva, levando ao surgimento de resistência a múltiplas drogas em genomas bacterianos. Compreender as redes genéticas e a regulação AMR é crucial, pois as bactérias desenvolvem mecanismos de defesa, como bombas de efluxo, formação de biofilme e regulação positiva de genes em resposta à pressão do antibiótico. As redes de interação proteína-proteína podem auxiliar na identificação de ARGs e suas interações funcionais, destacando associações com MGEs e genes vizinhos menos móveis. Essa análise abrangente ressalta a importância de estudar as interações genéticas para combater a propagação da resistência aos antibióticos.

Assim, este trabalho busca realizar a análise das interações genéticas entre patógenos resistentes a antibióticos e de hospedeiros humanos provenientes de amostras clínicas.

METODOLOGIA

Dataset para análises

Foram utilizados dados de 126 amostras de expressão gênica depositados no repositório Gene Expression Omnibus (GEO) do National Center for Biotechnology Information (NCBI). Os indivíduos são crianças com idades entre 0.25 e 9 anos, sendo 75 do sexo masculino e 51 do sexo feminino, 98 são casos, infectados por *Streptococcus. aureus* e 29 são controle. Além disso, informações sobre etnia, origem geográfica e outras características demográficas relevantes foram registradas. Também serão utilizadas 44 amostras de expressão gênica de *Streptococcus Aureus* oriundas do GEO.

Pré-processamento

Para a realização do pré-processamento dos dados foram realizados os seguintes passos:

- **Preparação da Matriz de Dados:** Os dados são oriundos de 3 tipos de tecnologias de Microarrays (Affymetrix Human Genome U133A e U133B Array e Sentrix Human-6 v2 Expression BeadChip). Desta forma, foi realizado o download dos três arquivos de expressão genica e suas respectivas matrizes de anotação. Esses arquivos foram manipulados para comporem a matriz geral de expressão.
- **Limpeza de Dados**: Foi realizada a verificação de valores ausentes, que não foram encontrados, dispensando qualquer tipo de imputação de dados.
- **Normalização**: Os dados de expressão foram normalizados utilizando a técnica de Z-score para garantir comparabilidade entre as amostras.

Análise Exploratória

Foi realizada uma análise exploratória inicial para identificar padrões e distribuições nos dados de expressão gênica. Assim, um *heatmap* das correlações entre os genes foi gerado para identificar relações entre os genes (Figura 1).

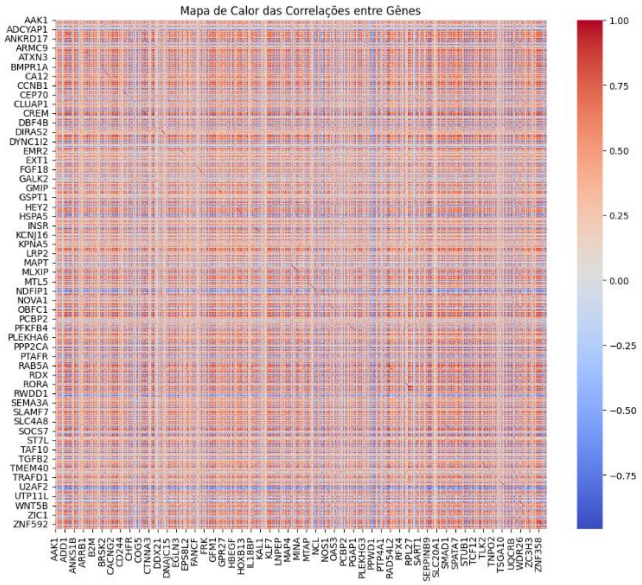

Figura 1 Heatmap da Correlação de Genes

Análise Diferencial de Expressão

Para a análise diferencial, foi realizada uma comparação estatística entre os grupos infectados e não infectados utilizando o teste t de Student. Os genes com valores de p significativos ($p < 0.05$) foram considerados diferencialmente expressos.

Aplicação de Machine Learning

Foram aplicados algoritmos de *machine learning* para classificar amostras como infectadas ou não infectadas com base nos dados de expressão gênica e a partir disso, foram utilizadas técnicas de *feature selection* para identificar os genes mais importantes que contribuem para a diferenciação entre os estados de infecção. Para isso, os dados foram inicialmente preparados para o treinamento do modelo. Em seguida, foi escolhido o modelo de machine learning Árvore de Decisão, para classificar as amostras. Os resultados do modelo, como a matriz de confusão, relatório de classificação e acurácia, foram avaliados para se entender o desempenho do modelo.

Resultados

A análise diferencial de expressão gênica entre os grupos infectados e não infectados revelou que vários genes foram diferencialmente expressos. Os genes com valores de p significativos ($p < 0.05$) foram considerados diferencialmente expressos. A Tabela 1 apresenta os top 10 genes diferencialmente expressos:

Tabela 1. Top 10 genes diferencialmente expressos

Gene	t_stat	p_value
ABCC3	3.885420	0.000165
ABCC4	2.174858	0.031539
ABHD2	2.723865	0.007384
ABHD4	3.053282	0.002770
ABHD5	3.585213	0.000483
ACAA2	1.987927	0.049023
ACOT8	2.848288	0.005147
ACPP	4.779890	0.000005
ACTB	2.805292	0.005839
ACTR6	-3.434553	0.000808

Para aprofundar a análise dos genes diferencialmente expressos, utilizamos o GeneMANIA para criar uma rede de interação gênica (Figura 2) com os top 10 genes identificados. Esta é uma ferramenta poderosa que permite prever interações funcionais entre genes com base em múltiplos tipos de dados, como co-expressão, co-localização e interações físicas.

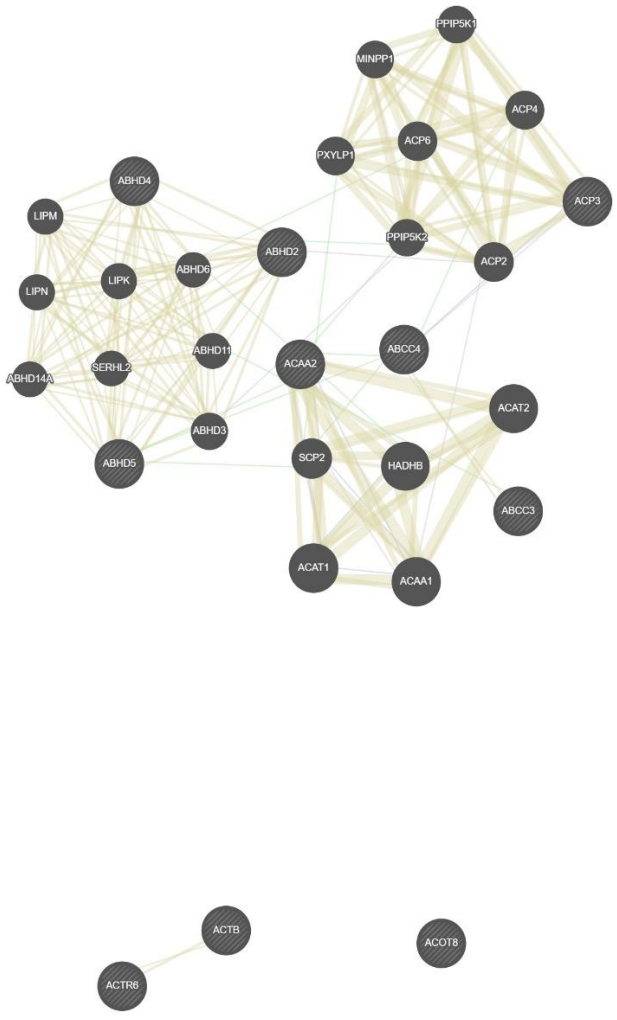

Figura 2. Rede de Interação Gênica

A rede de interação gênica gerada pelo GeneMANIA, revela interações funcionais importantes que podem fornecer mais insights sobre os mecanismos subjacentes à resposta à infecção por *Staphylococcus Aureus*. Os genes **ABCC3 e ABCC4** são genes que estão conectados a vários outros genes na rede, sugerindo um papel central na resposta à infecção. As interações com outros genes envolvidos em transporte e metabolismo sugerem que estes transportadores ABC podem estar crucialmente envolvidos na resistência antimicrobiana. O **ACAA2** está altamente conectado com genes envolvidos no metabolismo de ácidos graxos e processos oxidativos, como ACAT1, ACAA1 e HADHB. Isso reforça a ideia de que mudanças no metabolismo energético são uma resposta importante à infecção bacteriana. Os genes **ABHD2, ABHD4 e ABHD5** estão agrupados com outros genes da família ABHD e genes relacionados ao metabolismo de lipídios, como LIPK e LIPN. Isso sugere uma função coordenada na resposta ao estresse celular e na regulação de processos inflamatórios. Já os genes **ACTB e ACTR6** estão interconectados, indicando que mudanças na dinâmica do citoesqueleto e são uma resposta importante à infecção. Isso pode afetar a motilidade celular e a capacidade de resposta do sistema imunológico. Por fim, a interação do **ACPP** com genes como ACP4, ACP6, e PPIP5K1 indica um papel na regulação de processos celulares complexos e potencialmente na modulação da resposta imunológica.

A aplicação de um modelo de árvore de decisão para classificar amostras como infectadas ou não infectadas baseadas nos dados de expressão gênica apresentou os seguintes resultados:

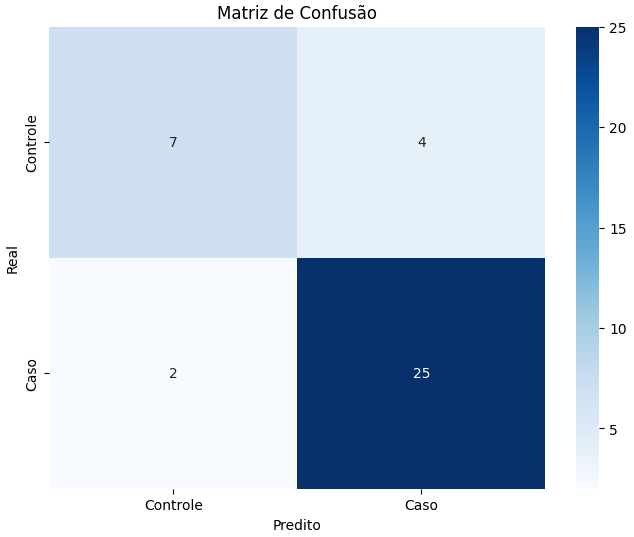

Figura 3. Matriz de Confusão

A matriz de confusão (Figura 3) mostra que o modelo classifica corretamente a maioria das amostras, com alguns erros de classificação. Esses resultados reforçam a importância de considerar múltiplas métricas de desempenho, como precisão e F1-score, para obter uma compreensão completa do desempenho do modelo. O F1-score fornece uma avaliação mais balanceada do desempenho do modelo, especialmente útil quando temos classes desbalanceadas. No caso apresentado, o F1-score ponderado foi de aproximadamente 0.84, alinhado com a acurácia que foi de também aproximadamente 0.84. Isso indica que o modelo está performando bem em geral, mas ainda há espaço para melhorias, especialmente na classe de amostras não infectadas, onde a precisão e o recall são mais baixos.

A análise de importância das características identificou os genes mais importantes para a classificação. Os principais biomarcadores identificados estão na Tabela 2.

Gene	Importância
MMP8	0.584954
RBM14	0.285185
RAB1A	0.068750
POLDIP2	0.061111
AAK1	0.000000
PTPRG	0.000000
PTPN2	0.000000
PTPN21	0.000000
PTPN22	0.000000
PTPN3	0.000000

Tabela 2. Genes mais importantes

O gene **MMP8** foi identificado como o biomarcador mais significativo, com uma importância de 0.584954, seguido por **RBM14** e **RAB1A**. Esses genes podem desempenhar um papel importante na diferenciação entre amostras infectadas e não infectadas.

Conclusão

A integração de técnicas de machine learning no trabalho proporcionou uma ferramenta importante para classificar amostras e identificar biomarcadores de expressão gênica. O modelo de árvore de decisão mostrou-se eficaz, com uma acurácia de 84%, destacando genes como MMP8, RBM14 e RAB1A como biomarcadores chave.

A rede de interação gênica gerada pelo GeneMANIA complementa essa análise, revelando interações funcionais importantes que podem fornecer mais insights sobre os mecanismos subjacentes à resposta à infecção por *Staphylococcus Aureus*. A identificação de genes centrais na rede sugere uma coordenação na resposta ao estresse, regulação metabólica e adaptação celular, todos cruciais para a resistência antimicrobiana.

Essas descobertas destacam a complexidade da resposta do hospedeiro à infecção e a importância de redes regulatórias coordenadas. A identificação de genes centrais e suas interações funcionais pode guiar futuras investigações para validar experimentalmente esses genes e explorar suas funções específicas em contextos clínicos, contribuindo para o desenvolvimento de estratégias mais eficazes no combate à resistência aos antibióticos.