# APRENDIZADO DE MÁQUINA APLICADO A DADOS DE GENÔMICA E TRANSCRIPTÔMICA

**Palavras-Chave: Machine Learning, Bioestatística, Redes Neurais**

**Autores:**

**João Pedro de Campos Formigari, IMECC – UNICAMP**

**Profª. Drª. Samara Flamini Kiihl, IMECC - UNICAMP**

## INTRODUÇÃO:

Para avançar na biomedicina e na saúde pública, é essencial entender os complexos mecanismos biológicos em nível celular, algo possível graças ao uso de tecnologias avançadas. O scRNA-Seq (sequenciamento de RNA de célula única) desponta como uma ferramenta inovadora, permitindo a análise da expressão gênica em células individuais. Essa técnica altamente detalhada revela subpopulações celulares, estágios do ciclo celular e padrões específicos de expressão gênica, oferecendo insights inéditos na biologia.

No entanto, a grande quantidade e a complexidade dos dados de scRNA-Seq representam um grande desafio. É nesse contexto que as técnicas de aprendizado de máquina se tornam valiosas. Utilizando métodos estatísticos de aprendizado de máquina, é possível entender melhor as interações nos sistemas biológicos e os padrões de expressão gênica em organismos complexos, como os humanos. Esses métodos, especialmente os supervisionados, são essenciais para identificar padrões de expressão e agrupar células em categorias específicas.

Para realizar essas análises, o ambiente R/Bioconductor se destaca como uma plataforma líder. O projeto Bioconductor desenvolve softwares especializados na análise de dados genômicos, incluindo scRNA-Seq, oferecendo uma infraestrutura robusta e avançada para que os pesquisadores possam explorar a complexidade dos dados biológicos em nível celular. Essa combinação entre tecnologia de ponta e análise estatística avançada é crucial para promover avanços na compreensão da biologia celular e na medicina personalizada.

## METODOLOGIA:

Este estudo se foca na análise de dados obtidos a partir da técnica de scRNA-Seq, que permite o sequenciamento de RNA em células individuais dos doadores. A análise inicia-se com a obtenção de uma matriz de contagem, onde cada linha representa um gene e cada coluna representa uma célula, contendo a quantidade de vezes que cada gene foi detectado em cada célula. Devido ao tamanho reduzido das unidades experimentais, essa matriz tende a ser esparsa, apresentando muitos valores zero. A preparação dos dados envolve a remoção de ruídos e a correção de efeitos de lote (Batch Effects), que são variações introduzidas por diferenças nos procedimentos experimentais.

Após essa preparação, os dados ainda podem ser muito volumosos, necessitando a aplicação de técnicas de redução de dimensionalidade, como PCA e UMAP, para facilitar a visualização e o agrupamento. Essas técnicas auxiliam na manipulação de dados de alta dimensionalidade, permitindo uma análise exploratória mais eficaz. Posteriormente, métodos de agrupamento são aplicados para identificar subpopulações celulares e diferenciar células com base nos padrões gênéticos.

Além disso, técnicas de aprendizado de máquina são empregadas para extrair *insights* dos dados. Entre essas técnicas, as redes neurais se destacam por sua eficácia em modelar padrões complexos e não lineares nos dados de *scRNA-Seq*. Inspiradas na estrutura neural do cérebro humano, essas redes consistem em camadas de neurônios interconectados, capazes de aprender padrões complexos e realizar tarefas como classificação e previsão a partir dos dados de entrada.

O modelo de rede neural utilizado foi um modelo sequencial construído utilizando a biblioteca Keras. A arquitetura é composta pelas seguintes camadas:

1. **Camada Densa (Dense Layer)**
    - **Unidades**: 128
    - **Função de Ativação**: tanh
    - **Input Shape**: Número de colunas do conjunto de treinamento amostral (genes)
2. **Camada de Dropout (Dropout Layer)**
    - **Taxa de Dropout**: 0,5
3. **Camada Densa (Dense Layer)**
    - **Unidades**: 64
    - **Função de Ativação**: tanh
4. **Camada de Dropout (Dropout Layer)**
    - **Taxa de Dropout**: 0.5
5. **Camada de Saída (Output Layer)**
    - **Unidades**: Número de classes (clusters)
    - **Função de Ativação**: softmax

O modelo foi compilado utilizando a função de perda *sparse_categorical_crossentropy* e o otimizador adam com uma taxa de aprendizado de 0.001. A métrica de desempenho utilizada foi *accuracy*.

### Função de Ativação tanh

A função de ativação *tanh* (hiperbólica tangente) é definida pela fórmula:

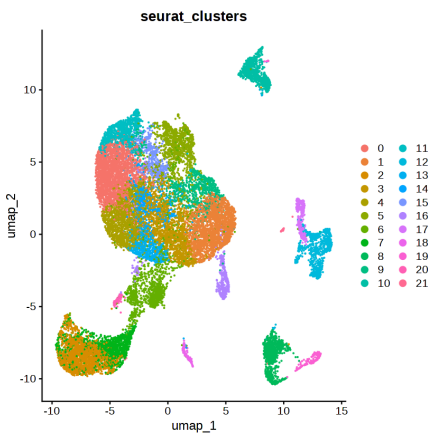
$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Ela transforma a entrada em valores no intervalo de -1 a 1, o que pode ser vantajoso em comparação com a função *sigmoid*, que mapeia valores no intervalo de 0 a 1. A função *tanh* é útil porque:

- **Centrada no Zero**: Ao contrário da função *sigmoid*, a saída da *tanh* é centrada em torno de zero, o que pode ajudar na convergência mais rápida do modelo durante o treinamento, pois as atualizações dos pesos podem ser feitas de forma mais eficiente.
- **Gradientes Mais Fortes**: Para entradas próximas de zero, a tanh fornece gradientes mais fortes, ajudando na aprendizagem dos parâmetros do modelo.

## RESULTADOS E DISCUSSÃO:

### Visualização dos Resultados da Clusterização




Para visualizar os resultados da clusterização, foi utilizado o método UMAP para reduzir a dimensionalidade dos dados e observar a distribuição das células em diferentes clusters. No gráfico UMAP apresentado, cada ponto representa uma célula, e a cor indica a qual cluster a célula pertence. A proximidade entre os pontos reflete a similaridade dos perfis de expressão gênica das células. Clusters bem definidos sugerem subpopulações de células com características distintas, enquanto a proximidade entre clusters pode indicar grupos de células com perfis similares.

*Figura 1 - Clusterização das Células*

Os clusters identificados pelo modelo indicam a presença de subpopulações celulares distintas com base nos perfis de expressão gênica. A separação clara entre os clusters sugere que o modelo conseguiu capturar as principais características que diferenciam essas subpopulações.

### Avaliação de Desempenho do Modelo

Para avaliar o desempenho do modelo de rede neural, utilizamos as métricas de acurácia e loss. A tabela a seguir resume os resultados obtidos:

*Tabela 1 - Desempenho do Modelo de Redes Neurais*

| Métrica | Valor |
|---|---|
| Acurácia | 0.7845 |
| Loss | 0.7042 |

- **Acurácia (Accuracy):** A acurácia de 0,7845 indica que o modelo classifica corretamente aproximadamente 78,45% dos exemplos no conjunto de teste. Isso demonstra que o modelo tem uma boa performance na tarefa de classificação dos clusters baseados nos genes.
- **Loss:** O valor do loss de 0,7042 é a medida da função de perda *sparse_categorical_crossentropy*. Esse valor quantifica o quão bem o modelo está ajustando os dados de treinamento. Valores menores indicam um melhor ajuste, mas é importante considerar também o risco de overfitting.

Os resultados do estudo mostram que a combinação de técnicas de redução de dimensionalidade e algoritmos de aprendizado de máquina, como redes neurais, pode ser eficaz na análise de dados de scRNA-Seq. A acurácia de 78,45% é um indicativo positivo da capacidade do modelo em classificar corretamente as células em seus respectivos clusters. No entanto, o valor do loss de 0,7042 sugere que há espaço para melhorias no ajuste do modelo.

Um dos desafios encontrados foi a esparsidade dos dados, que é uma característica comum em experimentos de scRNA-Seq. A remoção de ruídos e a correção de efeitos de lote foram etapas importantes para melhorar a qualidade dos dados antes da aplicação dos métodos de aprendizado de máquina.

Além disso, a escolha dos parâmetros para a redução de dimensionalidade, como UMAP, influenciou significativamente a qualidade dos clusters identificados. Ajustes pequenos nesses parâmetros podem levar a uma melhor separação entre os clusters e, consequentemente, a uma melhor classificação.

## CONCLUSÕES:

Este estudo demonstra que o uso de técnicas avançadas de aprendizado de máquina, combinadas com métodos de análise de dados de alto rendimento, pode proporcionar insights valiosos sobre a heterogeneidade celular. A rede neural utilizada conseguiu identificar subpopulações celulares com uma boa acurácia, evidenciando a complexidade biológica em nível celular.

Futuros trabalhos podem se beneficiar da exploração de diferentes arquiteturas de redes neurais e da aplicação de outras técnicas de pré-processamento de dados para melhorar ainda mais os resultados. A aplicação de métodos de aprendizado de máquina em dados de scRNA-Seq continua a ser uma área promissora para a biomedicina e a saúde pública.

## BIBLIOGRAFIA

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An Introduction to Statistical Learning with Applications in R. Second Edition, 2023.

Liao, M., Liu, Y., Yuan, J. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med 26, 842–844 (2020).

AWAN, Abid. Neural Network Models in R. DataCamp, https://www.datacamp.com/tutorial/neural-network-models-r, 2023.