Título: Desenvolvimento e otimização de modelos de *deep learning* para classificação de enzimas ativas de carboidratos (CAZymes)

Pesquisadora Responsável: Dra. Gabriela Félix Persinoti

Unidade: Laboratório Nacional de Biorrenováveis (LNBR)

1. Introdução

Ambientes naturais, como solos ricos em matéria orgânica, resíduos agrícolas e o trato digestivo de animais herbívoros, são ecossistemas complexos que abrigam uma imensa diversidade microbiana (Cabral et al 2022, Santos et al 2025). Esses microrganismos são fontes valiosas de genes e enzimas com alto potencial biotecnológico, como as CAZymes (Carbohydrate-Active Enzymes), enzimas envolvidas na degradação e modificação de carboidratos (Santos et al 2025).

Essas enzimas têm papel fundamental na despolimerização da biomassa lignocelulósica, transformando-a em açúcares fermentescíveis utilizados na produção de biocombustíveis, bioplásticos e outros bioprodutos. A descoberta de novas CAZymes é, portanto, estratégica para o avanço da bioeconomia e a transição para fontes renováveis de energia. No entanto, a alta diversidade genética e funcional dos microrganismos produtores dessas enzimas, em especial os não-cultiváveis, dificulta sua identificação por métodos tradicionais.

2. Estado da Arte

A anotação funcional de proteínas ainda depende, majoritariamente, de métodos baseados em alinhamento de sequências ou na identificação de domínios conservados, como os modelos ocultos de Markov (HMMs) (Drula el. al. 2022). Embora úteis, essas abordagens são limitadas na detecção de proteínas altamente divergentes ou pertencentes a novas famílias funcionais.

Recentemente, avanços em inteligência artificial, sobretudo em aprendizado profundo (deep learning), têm permitido a análise de grandes volumes de dados biológicos, revelando padrões complexos e possibilitando predições mais sensíveis e abrangentes (Bileschi el. al 2022). Modelos de linguagem treinados com sequências de proteínas, como os inspirados em NLP (Natural Language Processing), vêm se destacando por sua capacidade de inferir propriedades funcionais com base apenas na sequência primária (Heinzinger et al 2024).

A aplicação desses modelos à bioprospecção de CAZymes representa uma abordagem inovadora e promissora, com potencial para acelerar significativamente a descoberta de enzimas com novas atividades catalíticas.

3. Objetivos

Objetivo Geral

Desenvolver e aplicar modelos de deep learning para a identificação e classificação de enzimas ativas sobre carboidratos (CAZymes), a partir de dados metagenômicos.

Objetivos Específicos

- Implementar e treinar modelos de linguagem (protein language models) baseados em redes neurais profundas para classificação funcional de proteínas.
- Avaliar a capacidade dos modelos em identificar CAZymes conhecidas e detectar possíveis novas famílias com baixa similaridade com as já catalogadas.
- Comparar os resultados obtidos com ferramentas tradicionais como dbCAN e PFAM.
- Identificar e priorizar sequências candidatas com potencial biotecnológico para análise funcional futura.

4. Metodologia

- **Coleta e curadoria de dados:** Utilização de bancos públicos como CAZy (www.cazy.org) e UniProt (www.uniprot.org) para extração de sequências de CAZymes anotadas e de dados metagenômicos não anotados.
- **Pré-processamento:** Remoção de sequências redundantes, normalização de comprimentos e conversão para formatos compatíveis com entrada de modelos de IA.
- **Modelagem:** Treinamento de modelos de linguagem (e.g., ESM, ProtBERT, ProtT5) utilizando bibliotecas como PyTorch e Hugging Face Transformers.
- **Classificação:** Implementação de camadas finais de classificação supervisionada para rotular as proteínas como CAZymes e atribuir classes funcionais (GH, GT, PL, etc.).
- **Avaliação:** Métricas como acurácia, precisão, recall e F1-score serão utilizadas, além da comparação com ferramentas tradicionais de anotação.
- **Análise de novas sequências:** Identificação de clusters de proteínas com baixa similaridade às conhecidas, visando possível descoberta de novas funções enzimáticas.

5. Resultados Esperados

- Um modelo computacional capaz de classificar CAZymes com alta acurácia, mesmo em casos de baixa similaridade com sequências conhecidas.
- Um pipeline reprodutível para anotação funcional automatizada de proteínas de origem metagenômica.
- Identificação de candidatos enzimáticos com potencial para compor novas famílias ou subfamílias de CAZymes.

6. Cronograma

	Meses			
Atividade	**1-3**	**4-6**	**7-9**	**10-12**
Revisão bibliográfica	X			
Preparo do conjunto de treinamento	X			
Implementação dos primeiros modelos de linguagem		X		
Treinamento dos modelos com sequências anotadas		X		
Análise comparativa com métodos tradicionais			X	
Refinamento dos modelos desenvolvidos			X	
Aplicação dos modelos em dados metagenômicos para detecção de possíveis novas CAZymes				X
Relatório e apresentação em eventos				X

7. Referências

BILESCHI, M. L. et al. Using deep learning to annotate the protein universe. *Nature Biotechnology*, v. 40, p. 932–937, 2022.

CABRAL L. et. al. Gut microbiome of the largest living rodent harbors unprecedented enzymatic systems to degrade plant polysaccharides. *Nature Communications* 13, 629, 2022.

HEINZINGER, M. et. al. Bilingual language model for protein sequence and structure, *NAR Genomics and Bioinformatics*, V. 6, Issue 4, 2024.

Santos, C.A., Morais, M.A.B., Mandelli, F. *et al.* A metagenomic 'dark matter' enzyme catalyses oxidative cellulose conversion. *Nature* **639**, 1076–1083 (2025).

DRULA, E. et al. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research*, v. 50, D571–D577, 2022.