Evento	Salão UFRGS 2020: SIC - XXXII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2020
Local	Virtual
Título	Predição de genes essenciais com redes neurais para grafos
Autor	JOAO GABRIEL SCHAPKE DE OLIVEIRA
Orientador	MARIANA RECAMONDE MENDOZA GUERREIRO

Predição de genes essenciais com *redes neurais para grafos*

João Gabriel Schapke
Orientado por:
Mariana Recamonde-Mendoza
Anderson Tavares

O termo genes essenciais é usado para caracterizar o subconjunto de genes (ou proteínas) do genoma de um organismo que são essenciais à vida. A identificação de genes essenciais é um passo crítico para uma melhor compreensão da biologia e patologia humana. Métodos experimentais classificam genes como essenciais com um bom nível de certeza, porém, são custosos e demorados. Abordagens computacionais ajudaram a mitigar estas restrições, explorando métodos de aprendizado de máquina (AM) e a correlação de essencialidade com informações biológicas, especialmente redes de interação proteína-proteína (PPI), para prever genes essenciais de forma mais rápida e barata. No entanto, seu desempenho ainda é limitado, pois as centralidades de rede não são indicativos exclusivos da essencialidade, e os métodos tradicionais de AM são incapazes de aprender com domínios não euclidianos, como grafos. *Graph neural networks* (GNNs) são algoritmos de aprendizado profundo que visam desenvolver modelos preditivos através de conhecimento representado por grafos. Neste trabalho, hipotetizamos que usando redes PPI como grafos base para GNNs podemos alcançar um desempenho melhor do que os métodos comumente adotados nesta tarefa. Assim, propomos um modelo baseado em *Graph Attention Networks (GAT)*, uma GNN estado-da-arte, que aprende os padrões de essencialidade diretamente de redes PPI, integrando evidências adicionais de dados multiômicos codificados como atributos de nós. Nosso modelo foi treinado e avaliado para quatro organismos, incluindo humanos, atingindo área sob a curva ROC de 0.78 a 0.97. Nosso modelo superou significativamente os métodos baseados em rede e algoritmos tradicionais de AM, e alcançou um desempenho muito competitivo em relação ao método de *embedding node2vec*, estado-da-arte neste contexto. Notavelmente, o modelo GAT foi mais robusto em cenários com dados de treinamento limitados e desbalanceados. Assim, o método proposto oferece uma maneira eficaz e promissora para identificar genes e proteínas essenciais a partir de dados multi-ômicos.