CONSTRUÇÃO DE UMA FERRAMENTA COMPUTACIONAL BASEADA EM ALGORITMOS DE APRENDIZADO PROFUNDO PARA IDENTIFICAÇÃO DE PADRÕES EM DADOS GENÔMICOS

José Pires de Oliveira Neto[1]; Gilberto Nerino Souza Junior [2]; Fabricio Almeida Araújo[3]; Marcus de Barros Braga[4].

1. José Pires de Oliveira Neto, PIBIC, Graduando em Sistemas de Informação, Paragominas, e-mail: jose88pires@gmail.com; 2. Gilberto Nerino Sousa Junior; 3. Fabricio Almeida Araújo; 4. Marcus de Barros Braga, Paragominas, UFRA, e-mail: marcus.braga@ufra.edu.br

RESUMO:

A análise *downstream* de dados de sequências ômicas é frequentemente feita por meio da anotação funcional que pode ser realizada por várias ferramentas de bioinformática. Uma etapa fundamental da anotação funcional é o alinhamento entre uma sequência ainda não anotadas e um banco de dados de sequências já conhecidas. O padrão ouro para este alinhamento é feito por meio da ferramenta BLAST. Entretanto, o BLAST tem um custo computacional elevado, especialmente quando o banco de dados é grande. É o caso da utilização do BLAST com o UNIPROT, referência em sequências conhecidas, que contém quase 250 milhões de sequências. Neste sentido, novas abordagens computacionais têm sido desenvolvidas para mitigar este problema. É o caso da utilização de técnicas de Aprendizado Profundo - AP (*Deep Learning*). As técnicas de AP oferecem um importante conjunto de métodos para analisar sinais como áudio e fala, e texto, por meio de Redes Neurais Profundas (*Deep Neural Networks - DNN)*, em particular as Redes Neurais Convolucionais (*Convolutional Neural Networks - CNN*). É o caso das ferramentas SAdLSA, pLM-BLAST e BetaAlign, por exemplo. O trabalho proposto implementou uma CNN na ferramenta GOFEAT. Esta ferramenta é amplamente utilizada pela comunidade acadêmica para anotação funcional. GOFEAT utiliza como estratégia de alinhamento de sequências o BLAST. Os outros procedimentos da ferramenta GOFEAT não foram alterados. A nova ferramenta, CNN-GOFEAT, ainda em testes, apresentou resultados similares aos resultados do GOFEAT (~75%), demonstrando que a nova abordagem para um problema fundamental da bioinformática pode ser utilizada. Para um número de sequências pequeno (<100), o tempo de execução do GOFEAT foi consideravelmente menor que o CNN-GOFEAT. A diferença de tempo de execução entre as ferramentas vai caindo à medida que o número de sequências vai aumentando. Somente a partir de um número de sequências muito grande (>1000000) é o que tempo de execução entre as duas ferramentas se assemelha. Os testes foram executados em um computador que não possui GPU e este é, possivelmente, um dos motivos pelo qual a CNN não tenha tido uma performance tão boa, uma vez que modelos de IA, em geral, tem uma performance muito superior quando GPUs são utilizadas. Espera-se que este trabalho possa continuar e melhorar a performance do modelo proposto por meio da utilização de GPUs e outros modelos de Aprendizado Profundo.

PALAVRAS-CHAVE: BIOINFORMÁTICA; APRENDIZADO PROFUNDO; ANOTAÇÃO FUNCIONAL.