Predição de estrutura secundaria de proteinas utilizando uma rede neural MLP otimizada pelo algoritmo GWO

Leonardo A. Monte[1], **Emília G. Oliveira**[2], **Igor R. G. Souza**[3]

[1]Departamento de Computação (DC)
Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

{leonardo.amonte,emilia.galdino,igor.rgsouza}@ufrpe.br

Abstract.

Resumo.

1. Introdução

A proteína é um dos objetos de estudo centrais na Biologia, sendo um dos elementos-chave na execução de diversos processos biológicos, tais como suporte estrutural celular, catálise bioquímica, transporte celular, imunização, entre outras funções. As proteínas possuem diversos níveis de estrutura, os quais variam da cadeia de aminoácidos, denominada estrutura primária, até uma conformação tridimensional, resultante da interação entre os aminoácidos e das cadeias polipeptídicas. A estrutura secundária, um nível intermediário a estes, é definida pela ocorrência de estruturas estáveis na cadeia de aminoácidos. A conformação tridimensional, dita estrutura terciária, por sua vez, reflete diretamente a função da proteína, sendo a Cristalografia de Raios-X e Espectroscopia de Ressonância Magnética Nuclear os métodos experimentais mais utilizados para obtê-las.

Atualmente, com a finalização do sequenciamento do genoma humano e de diversos outros organismos, deu-se início a uma nova fase para as pesquisas genéticas, denominada Proteômica. Este termo envolve a identificação de todas as proteínas expressas pelo genoma bem como a determinação de suas funções fisiológicas e patológicas. Tal conhecimento é essencial para o desenvolvimento de novos medicamentos e métodos de diagnóstico, por exemplo.

As proteínas além de desempenharem funções catalíticas como de controle e regulação do metabolismo celular, desempenham funções de defesa e transporte, dentre outras. As funções de uma proteína estão diretamente relacionadas com a sua estrutura nativa. O processo a partir do qual a proteína sai de uma conformação aleatória e alcança sua estrutura nativa é conhecido como enovelamento (folding). A importância de se conhecer e solucionar o problema de enovelamento de proteínas acaba refletindo-se de forma considerável sobre a biotecnologia. Em princípio, este conhecimento pode permitir o desenvolvimento de novas proteínas com aplicações abrangendo um vasto campo.

2. Problema

Ao se estudar o enovelamento de proteínas pretende-se descobrir quais as propriedades do polipeptídio que levam a cadeia a adotar estrutura única e estável e, também, investigar como a sequência de aminoácidos (estrutura primária) de uma proteína está relacionada

com essas propriedades. Esse processo objetiva não apenas elucidar o problema do enovelamento de proteínas, mas também produzir proteínas que possuam estrutura (secundária e terciária) desejada. Para isso, métodos como Algoritmos Genéticos, Redes Neurais Artificiais e Simulações com Monte Carlo vêm sendo utilizados, a fim de predizer estruturas primárias (sequências) de aminoácidos que levem às estruturas secundárias ou terciárias desejadas.

Devido a sua complexidade, o problema de determinar a função de uma proteína a partir de sua estrutura primária utilizando métodos computacionais vem sendo abordado de diversas formas. A análise pode ser puramente sequencial, através da busca por padrões pré-determinados ou não, ou pode usar informações de proteínas de estruturas conhecidas, através de métodos preditivos ou comparativos.

Uma abordagem recorrente no problema de obtenção da estrutura secundária é o uso de aprendizagem de máquina, onde, a partir da observação de um conjunto de treinamento dado como entrada, é feita a predição da estrutura secundária correspondente a uma determinada sequência de aminoácidos, com um certo grau de acurácia

3. Abordagem

Um levantamento bibliográfico dos métodos mais recentes de predição foi realizado com o objetivo de analisar o estado da arte, avaliando o progresso atingido pelos métodos mais recentes. Também foram analisados aspectos relacionados aos dados de entrada, como a quantidade e formatação. Adicionalmente, foi proposto o desenvolvimento de um preditor de estrutura secundária baseado em uma rede neural artificial direcionada com múltiplas camadas (MLP, do inglês multilayer perceptron) otimizada pelo algoritmo evolutivo GWO (Grey Wolf Optmizer). Os resultados deste método foram comparados com os de uma MLP sem otimização.

A base de dados escolhida como entrada para o modelo foi a CB513, o qual é composta por 513 proteínas não-homólogas. Foi feita a coleta dos dados referentes ao conjunto de aminoácidos referentes a cada proteína e sua respectiva classe. A formatação dos dados para utilização no modelo construída de forma em que cada aminoácido foi mapeado a um número, de modo que cada protéina tem um conjunto especifíco de mapeamento (números). A quantidade de classes utilizadas para predição foram 3, onde é construido um vetor binario de tamanho 3 para cada tipo de classe, para que a rede consiga predizer a saída.

O modelo foi construído utilizando a biblioteca Keras, na linguagem Python. O modelo contém 4 camadas, sendo elas 1 de entrada, 2 intermedíarias e uma de saída. A camada de entrada é variante segundo a janela escolhida para representação do vetor característica, a utilizada neste trabalho foi 5. As camadas intermedíarias possuem 120 e 60 neurônios respectivamente e a camada de saída possui 3. Gerando assim 7980 parâmetros de peso possíveis de serem otimizados pelo algoritmo evolutivo.

O algoritmo evolutivo escolhido para a tarefa foi o Grey Wolf optmizer (GWO) [Mirjalili et al. 2014]. O GWO é um meta-heurística de inteligência coletiva, onde são escolhidos indivíduos para uma população a ser otimizada aleatoriamente e a partir dessa população de indíviduos, operadores matemáticos são usados para percorrer o espaço de busca de modo em que encontre o melhor indivíduo possível. A modelagem utilizada para este projeto, foi a de otimização de pesos da rede neural, utilizada de modo em que

cada indivíduo da população é um paramêtro (peso) a ser otimizado. A função de fitness para escolha dos melhores indivíduos foi feita em relação a acurácia no conjunto de validação, onde quanto maior a acurácia melhor o modelo. Após uma quantidade n de Gerações, sendo 1 geração uma iteração do algoritmo, é salva a melhor rede encontrada para a realização do teste.

Foi utilizado como método de experimentação do modelo o 10Fold-crossvalidation, onde a cada fold o conjunto de treinamento gerado é dividido em treino e validação, gerando as seguintes proporções: 10Fold: 90/10 e desses 90 gerado para treino é divido em 80/20, sendo 80 para treino e 20 para validação. O modelo é testado utilizando os 10 da divisão do 10Fold. Após o treinamento, validação e teste, é feito o cálculo da média de acurácia de teste, desvio padrão de teste e mediana de teste. Foi gerado um gráfico referente a média de acerto, um boxplot e uma tabela que podem ser vistas na sessão de resultados.

4. Resultados

Os resultados obtidos mostram a superioridade da abordagem utilizando o algoritmo evolutivo. Mesmo tendo sido utilizadas uma ínfima quantidade de gerações e de tamanho de poupulação, ainda assim o algoritmo consegue ter um melhor ínicio e melhores resultados no conjunto de teste.

	MLP	GWO
Média	52,91	69,81
Mediana	52,84	69,78
STD	0,0041	0,0040
Melhor Score	53,65	70,59

Figura 1. Tabela contendo resultados.

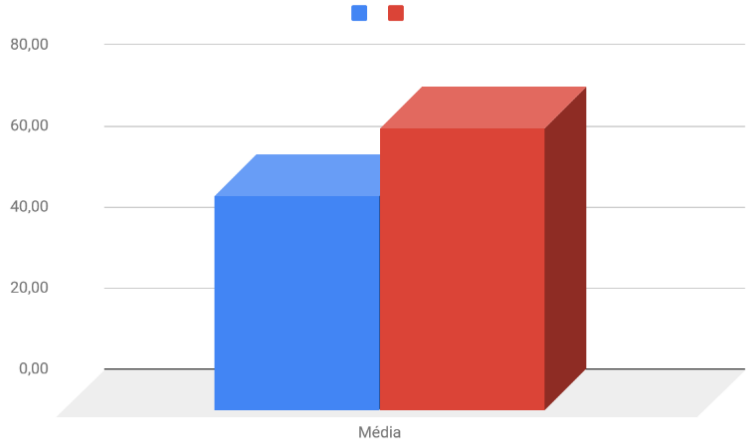

Figura 2. Gráfico comparando resultados de teste.

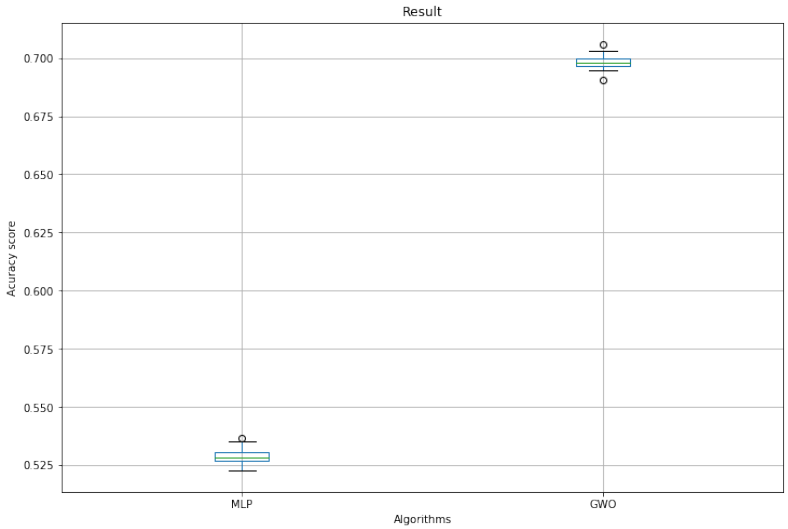

Figura 3. Boxplot do resultado de teste.

Referências

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69:46 – 61.