Redes Neurais Multicamadas *Perceptron* na Identificação de Proteínas Efetoras

Gustavo José da Silva[1], Ramon Gustavo Teodoro Marques da Silva[1], Ricardo Vasconcellos de Carvalho Remédio[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – IFSULDEMINAS (campus Muzambinho) – Caixa Postal 02 –37.890-000 – Muzambinho – MG – Brasil

gustavo.jose@muz.ifsuldeminas.edu.br, ramon.silva@ifsuldeminas.edu.br, ricardo.vcr2@gmail.com

Abstract. *The study of the proteome of bacteria is important in the discovery of targets for the development of drugs for the treatment of bacterial diseases. Bacteria secrete effector proteins that alter cellular processes causing disease. The sequences of effector proteins make it possible to extract characteristics that determine their behavior. Thus, the objective of this work was to develop a multi-layer perceptron neural network capable of predicting the effector potential of proteins from 8 biological characteristics extracted from their sequences, related to hydropathy, apoptosis and presence of double helix. We used 392 protein sequences, divided into training and post-training sets, for training and evaluation of the neural network. The identification index was 97.5% in the training set and 80.7% in the post-training set.*

Resumo. *O estudo do proteoma de bactérias é importante na descoberta de alvos para o desenvolvimento de fármacos para o tratamento de doenças bacterianas. As bactérias secretam proteínas efetoras que alteram os processos celulares causando doenças. As sequências de proteínas efetoras possibilitam extrair características que determinam seus comportamentos. Assim, o objetivo deste trabalho foi desenvolver uma rede neural multicamadas perceptron capaz de predizer o potencial efetor de proteínas a partir de 8 características biológicas extraídas de suas sequências, relacionadas com a hidropatia, apoptose e presença de dupla hélice. Foram utilizadas 392 sequências de proteínas, divididas em conjuntos de treino e pós-treino, para treinamento e avaliação da rede neural. Chegou-se ao índice de identificação de 97,5% no conjunto de treinamento e 80,7% no conjunto de pós-treinamento.*

1. Introdução

A análise de dados biológicos por meio de recursos computacionais auxilia no entendimento de diversos processos celulares, principalmente na interação entre genes, proteínas e doenças. As sequências genéticas de nucleotídeos e aminoácidos oferecem informações importantes acerca do funcionamento da maquinaria celular. Nas últimas décadas os avanços tecnológicos viabilizaram o mapeamento de milhares de sequências de proteínas, o que impulsionou a utilização de técnicas computacionais nas análises de proteoma. Dentre as técnicas utilizadas, o aprendizado de máquina ganhou grande

destaque, devido a sua grande aplicabilidade e flexibilidade na extração de regras e padrões de conjuntos de dados (LIN et al., 2017).

As proteínas exercem um papel fundamental no organismo humano, estando relacionadas com diversas patologias. Em muitas delas, as bactérias secretam proteínas durante a interação com a célula no organismo hospedeiro, podendo alterar os processos celulares e causar doenças no mesmo. Essas proteínas são chamadas de proteínas efetoras (ALVAREZ-MARTINEZ; CHRISTIE, 2009).

De acordo com Meyer et al. (2013), as proteínas possuem características biológicas relacionadas à hidropatia e inibição da apoptose que podem ser extraídas por análises de sequências. Em relação a hidropatia pode-se citar: hidropatia total, hidropatia média, hidropatia do C- terminal, carga do C-terminal e aminoácidos polares básicos na parte C-terminal. Já em relação à inibição da apoptose pode-se citar: presença de sinal nuclear e presença de sinal mitocondrial. Essas características podem apresentar padrões para proteínas efetoras (WANG et. al., 2017). Outra característica de destaque é a dupla hélice, uma característica de espécies eucariontes, que pode ser encontrada em proteínas efetoras, pois ao interagir com a célula do hospedeiro as proteínas efetoras tendem a imitar o comportamento do mesmo (LOCKWOOD et. al., 2011).

Desta forma, o objetivo deste trabalho é desenvolver uma rede neural multicamadas *perceptron,* que a partir de características biológicas de proteínas, relacionadas a hidropatia, apoptose e estrutura de dupla hélice, identificará padrões e auxiliará na predição de proteínas efetoras.

2. Desenvolvimento da Rede Neural

Inicialmente foi realizado o download das sequências *FASTA* de 392 proteínas efetoras e não efetoras de bactérias no *Genbank*[1]. Após um levantamento junto a literatura, as sequências foram divididas em dois grupos, um contendo as proteínas efetoras (144 proteínas) e outro as não efetoras (248 proteínas). Em seguida, foram extraídas as características sobre hidropatia, inibição de apoptose e presença de dupla hélice dos dois grupos de proteínas, totalizando **8** características para cada proteína. As características são representadas por valores numéricos, e foram extraídas utilizando as seguintes ferramentas de bioinformática: *Hydrocalc Proteome*[2], *NLStradamus*[3], *TargetP*[4], e *Coiled Coil Prediction*[5].

A ferramenta *Hydrocalc Proteome* possibilitou a extração de 5 características relacionadas a hidropatia das proteínas: hidropatia total, hidropatia média, hidropatia do C- terminal, carga do C-terminal e aminoácidos polares básicos no C-terminal. A ferramenta *NLStradamus* permitiu extrair dados sobre a presença do sinal de localização nuclear. A ferramenta *TargetP* gerou dados relacionados ao sinal de localização mitocondrial. E a ferramenta *Coiled Coil Prediction* foi utilizada para a identificação da estrutura de dupla hélice.

---

[1] https://www.ncbi.nlm.nih.gov/genbank/
[2] http://www.gmb.bio.br/hydrocalc/
[3] http://www.moseslab.csb.utoronto.ca/NLStradamus/
[4] http://www.cbs.dtu.dk/services/TargetP/
[5] https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_lupas.html

Os dados foram reorganizados em mais dois grupos, um para o treinamento da rede neural e outro para o pós-treinamento. No documento de treinamento foram armazenadas 100 proteínas efetoras e 100 proteínas não efetoras, totalizando 200 proteínas no documento. Enquanto no documento de pós-treinamento foram armazenadas as proteínas restantes, 44 proteínas efetoras e 148 não efetoras. Os documentos foram transformados para o formato de arquivo *Comma Separated Values* (csv), para a entrada no software *WEKA*[6], utilizado na implementação da rede neural.

As análises no *WEKA* foram realizadas em 2 etapas: a primeira com 5 neurônios, e a segunda com 2 camadas intermediárias e 5 neurônios em cada, totalizando 10 neurônios. Na figura 1 é apresentado um esboço do funcionamento da rede neural para 10 neurônios.

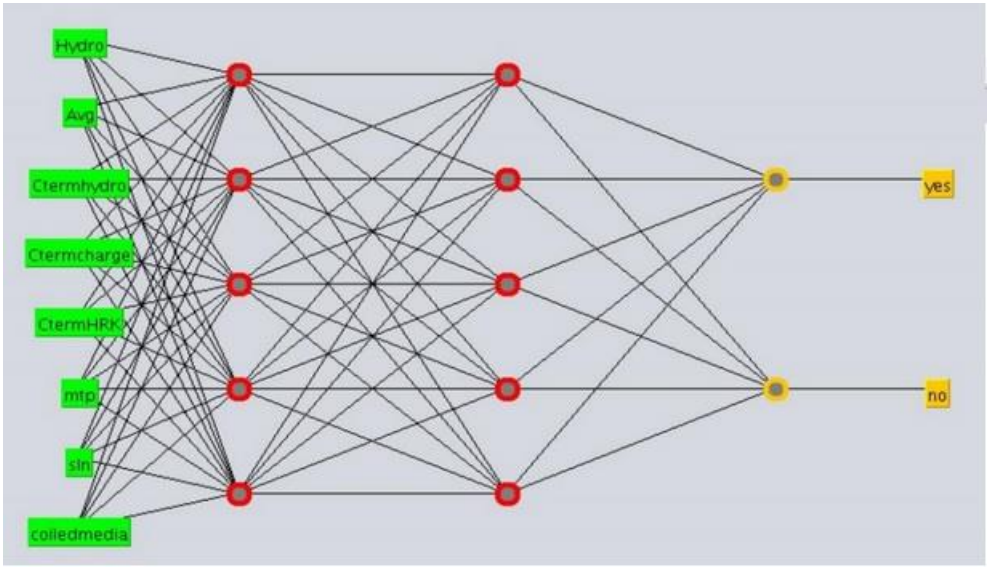

Figura 1. Esboço da rede neural multicamadas *perceptron* para 10 neurônios

No treinamento foi utilizada a rede neural multicamadas *perceptron*, que possibilita a configuração de limites de épocas. O limite de época é a quantidade de vezes que é feito o ajuste dos pesos em uma rede. Foram realizados treinamentos com as seguintes quantidades de limite de época: 1.000, 5.000, 10.000 e 5.000.000 de épocas.

3. Resultados e Discussão

A rede neural foi desenvolvida a partir dos dados das características biológicas descritas no tópico 2. Foram utilizadas diferentes quantidades para número de épocas nas análises do conjunto de treinamento. Não há uma regra específica que indica o número de limite de épocas ideal para a rede, porém, diante de cada treinamento realizado, foi possível observar que o número de ajustes dos pesos na rede possibilitou melhores resultados, tanto nas análises com 5 neurônios, quantos nas análises com 10 neurônios. A melhor taxa de acertos foi de 97,5%, obtida pela rede neural com 5 neurônios e 5 milhões de épocas. Com 5 neurônios a rede conseguiu uma melhor generalização, pois uma maior complexidade não refletiu melhor desempenho. Isso é explicado como *overfiting*, ou seja, complexidade maior do que o necessário pode afetar o desempenho. Na tabela 1

[6] https://www.cs.waikato.ac.nz/ml/weka/

são apresentados os resultados das análises no conjunto de treinamento para redes com 5 e 10 neurônios, com as taxas de erro e acerto por número de épocas.

Tabela 1. Taxas de erro e acerto por número de épocas para redes neurais com 5 e 10 neurônios

Número de épocas	Rede neural com 5 neurônios		Rede neural com 10 neurônios	
	Erro por época	Taxa de acerto %	Erro por época	Taxa de acerto %
1.000	0.0403747	96	0.0586401	93.5
5.000	0.0304651	97	0.0330315	96
10.000	0.029665	97	0.0312143	96
5.000.000	0.0250004	97.5	0.0259416	96.5

As análises no conjunto de pós-treinamento geraram 80,7% de acertos e 19,3% de erros. Utilizou-se 5 milhões de épocas e 5 neurônios, visto que estes valores apresentaram as melhores taxas de acerto nas análises do conjunto de treinamento, conforme especificado na tabela 1.

4. Considerações Finais

O trabalho propôs o desenvolvimento de uma rede neural multicamadas *perceptron* para a predição de proteínas efetoras. A rede neural alcançou uma taxa de acerto 97.5% no conjunto de treinamento e 80,7% no conjunto de pós-treinamento, o que mostra que o método foi eficaz para pequenos conjuntos de sequências, porém novos testes podem ser realizados para avaliar a eficácia com grandes conjuntos de sequências. Outro fator a ponderar é que novas características biológicas relacionadas a eucariotos, como a presença de padrões anquirina, podem auxiliar na evolução da rede. Um conjunto de treinamento maior também poderá contribuir para melhorar a rede neural.

Referências

Alvarez-Martinez, C.E.; Christie, P.J. Biological Diversity of Prokaryotic Type IV Secretion Systems. Microbiology And Molecular Biology Reviews. Houston, v. 73 (4), p. 775-808, 2009.

Lin, C., Jain, S., Kim, H., Bar-Joseph, Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Research, v. 45(17), p. 1-11, 2017.

Lockwood, S. et al. Identification of Anaplasma marginale Type IV Secretion System Effector Proteins. Plos One, v. 6(11), p. 1-8, 2011.

Meyer, D.F. et al. Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. Nucleic Acids Research, v. 41 (20), p. 1-12, 2013

Wang, Y., Guo, Y., Pu, X., Li, M. Effective prediction of bacterial type IV secreted effectors by combined features of both C-termini and N-termini. Journal of Computer-Aided Molecular Design, v. 31 (11), p. 1029 – 1038, 2017.