Predição de Estrutura Secundária de Proteínas com Redes Neurais: Um Estudo Aplicado com o Conjunto de Dados RS126

José Fernando Carvalho de Oliveira[1], Lívia Maria Reis[2], Paloma Silva de Morais[3] e Wellington Luis da Cunha[4]

Departamento de Computação - DC
Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brasil

{josefernando354[1], liviamreis.10[2], wellington.cunha[3], paloma.morais5[4]}@gmail.com

1. Introdução

1.1. O problema da Predição de Estruturas Secundárias de Proteínas (PSSP)

A predição da estrutura secundária de proteínas (Protein Secondary Structure Prediction – PSSP) é uma das tarefas centrais na biologia computacional, com implicações importantes para a compreensão das funções biológicas das proteínas e para o avanço de áreas como a medicina personalizada, a engenharia de enzimas e o desenvolvimento de novos fármacos. As proteínas são macromoléculas orgânicas formadas pela união de vários aminoácidos através de ligações peptídicas. Cada proteína assume uma conformação estrutural necessária para desempenhar sua função dentro do ser vivo. Podemos classificar as proteínas de acordo com sua estrutura em quatro grupos: primárias, secundárias, terciárias e quaternárias (Saini e Hou, 2013; Pka et al., 2021).

Estruturas secundárias, como hélices alfa, folhas beta e regiões em espiral (coils), são elementos estruturais que emergem do enovelamento da cadeia polipeptídica. A determinação experimental dessas estruturas, por métodos como cristalografia de raios X ou espectroscopia de RMN, é dispendiosa e demorada como evidenciado em 2020 na pesquisa *A guide to membrane protein X-ray crystallography*. Nesse contexto, abordagens computacionais surgem como alternativas promissoras para prever essas estruturas com base apenas na sequência de aminoácidos da proteína. No entanto, métodos mais tradicionais ou menos convencionais, como predições baseadas apenas em regras heurísticas, alinhamentos diretos de sequências ou modelos simplificados de energia, frequentemente apresentam limitações. Eles podem falhar em capturar padrões complexos, ignorar interações não locais e gerar resultados pouco robustos quando aplicados a proteínas com baixa similaridade a sequências já conhecidas, o que compromete tanto a precisão quanto a aplicabilidade em larga escala.

1.2. Redes Neurais na Bioinformática

As redes neurais artificiais são modelos computacionais inspirados na estrutura e funcionamento dos neurônios biológicos. Elas têm capacidade de aprender relações complexas e não-lineares a partir de grandes dados, como por exemplo, milhares de proteínas com estrutura conhecida (Kovács, Z. L. , 2006).

No caso específico do PSSP, redes neurais oferecem a vantagem de capturar padrões espaciais e dependências contextuais entre resíduos de aminoácidos. O uso de redes

convolucionais (CNNs), originalmente desenvolvidas para processamento de imagens, tem se tornado cada vez mais comum na biologia computacional por sua capacidade de extrair características locais e invariantes de janelas deslizantes de sequência. Além disso, a combinação de CNNs com perfis evolutivos gerados por ferramentas como o PSI-BLAST[1] permite que o modelo se beneficie de informações de conservação evolutiva, enriquecendo significativamente a representação da sequência de entrada.

A tarefa do PSSP é, portanto, acelerar a anotação funcional de proteínas recém-descobertas e oferecer suporte estratégico a pesquisas em bioengenharia, farmacologia e design de fármacos, reduzindo custos e tempo em comparação a métodos experimentais. No presente trabalho, foi proposto o uso de redes neurais convolucionais (CNNs) para a predição da estrutura secundária de proteínas, explorando perfis evolutivos gerados pelo PSI-BLAST e utilizando o conjunto de dados RS126 para treinamento e validação do modelo, com a motivação de alcançar maior acurácia e generalização em relação a técnicas convencionais.

2. Metodologia

Este trabalho foi conduzido em três etapas principais: preparação dos dados, construção do conjunto de treinamento e teste, e desenvolvimento da rede neural convolucional (CNN) para a predição da estrutura secundária de proteínas, o link do repositório encontra-se no **Apêndice A**.

2.1. Preparação dos Dados

Inicialmente, o conjunto de dados RS126 foi convertido para o formato .FASTA. Um script em Python foi utilizado para ler o arquivo .txt contendo as sequências proteicas e suas respectivas anotações estruturais, separando cada sequência em arquivos individuais no formato .fasta, facilitando a análise subsequente.

2.2. Geração de perfis PSSM

Com os arquivos .FASTA prontos, foi utilizado um script de shell (.sh) automatizado para gerar os perfis de substituição de posição (PSSM) usando a ferramenta PSI-BLAST.

Esses perfis representam, para cada posição da sequência, as probabilidades de ocorrência de cada aminoácido, capturando informações de conservação evolutiva. Os arquivos .pssm foram normalizados usando z-score e preenchidos (padding) para um comprimento máximo de 754 resíduos, garantindo uniformidade na entrada da rede. As estruturas secundárias (H: hélice alfa, E: folha beta, C: coil) foram codificadas como vetores one-hot.

Essas informações foram agrupadas em arrays NumPy e salvas em um único arquivo compactado (pssm_dataset.npz), contendo os dados de entrada (X), os rótulos (y) e uma máscara de pesos (weights) para indicar quais posições das sequências são válidas (evitando considerar o padding durante o treinamento).

2.3. Arquitetura da Rede Neural

A rede neural proposta é uma CNN composta por duas camadas convolucionais 1D, seguidas por camadas de normalização por lote (BatchNormalization) e camadas de dropout

[1] https://www.ebi.ac.uk/jdispatcher/sss/psiblast

para regularização. Após as convoluções, foram adicionadas camadas densas com funções de ativação ReLU e tanh, finalizando com uma camada de saída softmax para classificar cada posição da sequência em uma das três classes estruturais.

A rede foi compilada com o otimizador Adam e a função de perda categorical_crossentropy. O treinamento utilizou 80% dos dados, com divisão adicional em validação e teste (10% cada). Foram utilizados mecanismos de early stopping e redução da taxa de aprendizado (ReduceLROnPlateau) para prevenir o overfitting. A máscara de pesos foi aplicada como sample_weight durante o treinamento para garantir que o modelo aprendesse apenas com posições válidas das sequências.

3. Resultados e Discussão

3.1. Desempenho da Rede

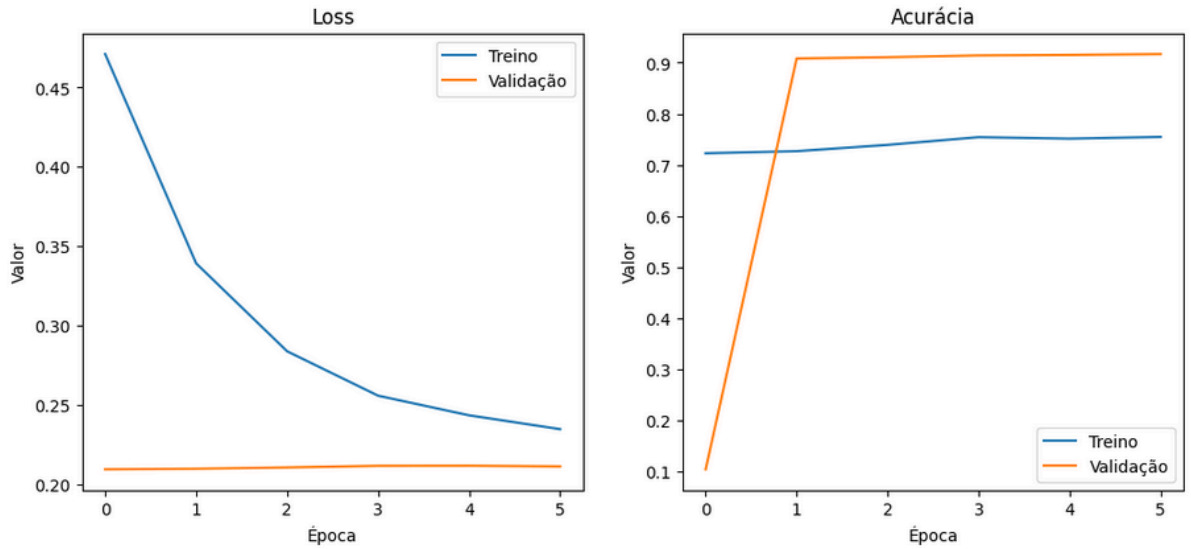
O modelo foi avaliado no conjunto de teste com base na métrica Q3, que representa a proporção de classificações corretas nas três classes estruturais. O valor obtido foi de Q3 = 0.7157. Além disso, o desempenho foi analisado em termos de precisão, recall e f1-score para cada classe:

Tabela 1

Classe	Precisão	Recall	F1-Score
H (Hélice alfa)	0.47	0.29	0.35
E (Folha beta)	0.39	0.52	0.45
C (Coil)	0.60	0.66	0.63

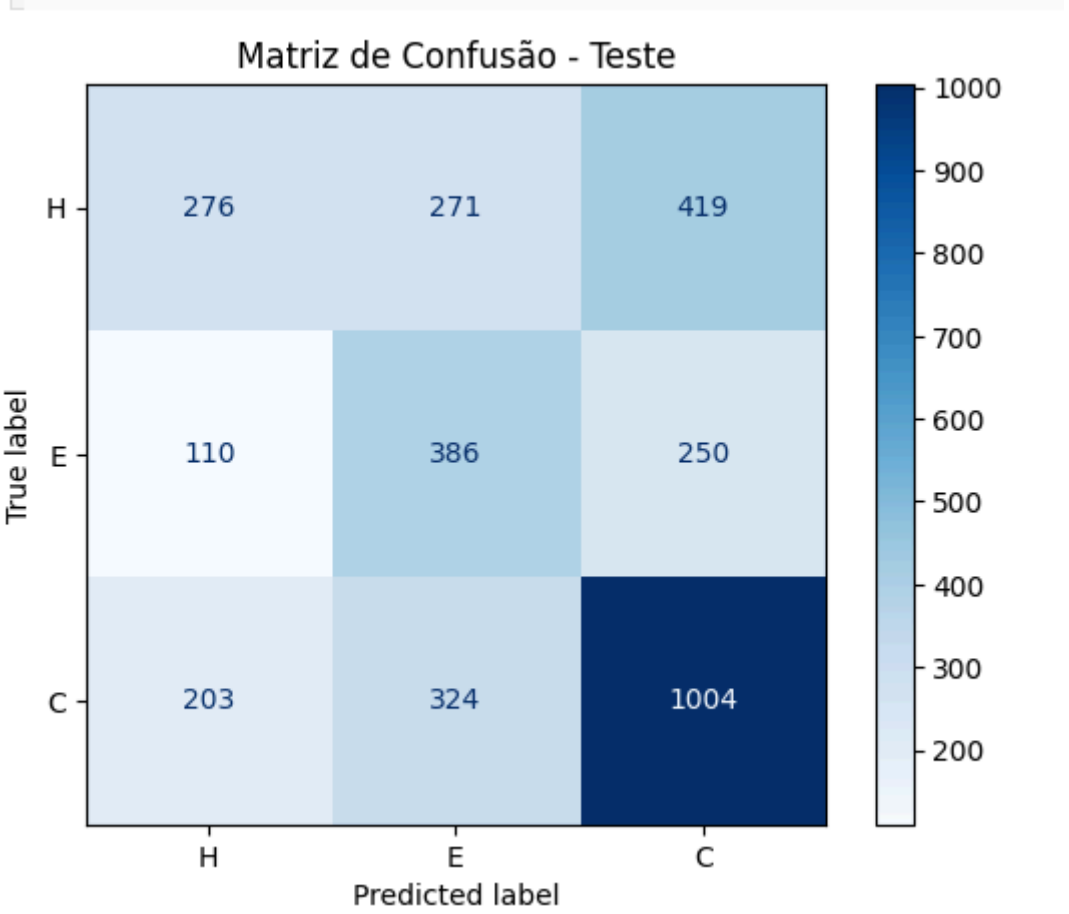
As curvas de loss e acurácia, apresentadas na Figura 1, mostram o comportamento do modelo ao longo das épocas de treinamento.

Figura 1. Curvas de Loss e Acurácia



fonte: autoral

Figura 2. Matriz de confusão



fonte: autoral

A Figura 2 apresenta a matriz de confusão sobre o conjunto de teste. É possível observar que há maior confusão entre as classes H (hélice alfa) e C (coil), com uma quantidade significativa de hélices previstas incorretamente como regiões de coil. Isso evidencia o desafio do modelo em distinguir estruturas secundárias com padrões similares.

3.2. Análise Geral

O modelo alcançou um Q3 de aproximadamente 71,6%, resultado consistente com valores relatados em trabalhos similares na literatura, indicando que a abordagem adotada apresenta desempenho competitivo dentro do estado da arte para predição de estrutura secundária de proteínas. Apesar de satisfatório, o índice sugere que ainda há espaço para aprimoramentos.

A análise por classe revelou um desempenho superior na classe "coil", possivelmente em função de sua maior representatividade no conjunto de dados, o que confere maior suporte estatístico durante o processo de aprendizado. Em contrapartida, as estruturas helicoidais apresentaram maior índice de erro, fato que pode estar associado tanto à menor

quantidade relativa de exemplos quanto à complexidade intrínseca na diferenciação de padrões helicoidais em comparação a regiões mais desordenadas. Esse comportamento reforça a importância de balanceamento adequado das classes e de técnicas que possam mitigar o viés introduzido por distribuições desiguais nos dados.

As curvas de aprendizado indicaram boa estabilidade ao longo do treinamento, sem sinais expressivos de sobreajuste, o que sugere que a arquitetura escolhida conseguiu capturar os padrões relevantes dos dados de forma consistente. Além disso, a matriz de confusão evidenciou padrões específicos de confusão entre as classes, apontando, por exemplo, que certos segmentos classificados como hélice foram erroneamente previstos como folhas beta ou coils. Esse resultado indica que o modelo consegue identificar as estruturas de forma geral, mas ainda enfrenta desafios para distinguir fronteiras sutis entre padrões estruturais mais semelhantes.

De forma geral, os resultados obtidos demonstram que a CNN implementada constitui uma base para a predição de estrutura secundária de proteínas, apresentando notável nível de acurácia e estabilidade, embora ainda existam desafios relacionados ao desbalanceamento das classes e à distinção entre conformações helicoidais e folhas beta, que podem ser explorados em trabalhos futuros por meio de técnicas de aumento de dados, arquiteturas mais profundas ou integração de informações adicionais, como perfis evolutivos mais ricos.

Referências

A INTELIGÊNCIA artificial na resolução de um dos maiores mistérios da biologia: conheça o AlphaFold 2. *Blog do Profissão Biotec*, v. 10, 14 dez. 2023. Disponível em: https://profissaobiotec.com.br/inteligencia-artificial-resolucao-maiores-misterios-biologia-conheca-alphafold2/. Acesso em: 26 jul. 2025.

DILL, K. A.; OZKAN, S. B.; SHELL, M. S.; WEIKL, T. R. The protein folding problem. Annual Review of Biophysics, v. 37, p. 289–316, 2008. Disponível em: https://doi.org/10.1146/annurev.biophys.37.092707.153558. Acesso em: 27 Jul. 2025.

HO, C. T.; HUANG, Y. W.; CHEN, T. R.; LO, C. H.; LO, W. C. Discovering the Ultimate Limits of Protein Secondary Structure Prediction. Biomolecules, v. 11, n. 11, p. 1627, 3 nov. 2021. Disponível em: https://doi.org/10.3390/biom11111627. PMID: 34827624; PMCID: PMC8615938. Acesso em: 27 Jul. 2025.

JUNQUEIRA, L. C. & CARNEIRO, J. 2012. Biologia Celular e Molecular. 9 ed. Rio de Janeiro: Guanabara Koogan. 376 p. Acesso em: 27 Jul. 2025.

KOVÁCS, Zsolt László. Redes neurais artificiais. Editora Livraria da Fisica, 2006. Disponível em: Acesso em: 27 Jul. 2025.

KERMANI, A. A. A guide to membrane protein X-ray crystallography. FEBS Journal, v. 288, p. 5788-5804, 2021. Disponível em: https://doi.org/10.1111/febs.15676. Acesso em: 27 Jul. 2025.

Apêndice A Repositório GitHub