UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE COMPUTAÇÃO

Exame de Qualificação de Mestrado

17 de Outubro de 2019

Classificação de Estruturas Secundárias e de Funções de Proteínas Utilizando Aprendizado de Máquina

Candidato: Gabriel Bianchin de Oliveira

Orientador: Prof. Dr. Zanoni Dias

Coorientador: Prof. Dr. Hélio Pedrini

Sumário

Resumo

Com avanços na área da biotecnologia, o sequenciamento de aminoácidos que formam as proteínas vem se tornando cada vez mais simples. Porém, diferente desse avanço no sequenciamento, a identificação de características das proteínas, como estruturas tridimensionais (estruturas secundárias e estruturas terciárias) e funções, ainda é complexa. Devido ao elevado custo de análises por métodos laboratoriais, abordagens que utilizam tecnologias da informação vêm se tornando cada vez mais presentes na análise de processos biológicos. Embora abordagens utilizadas na literatura apresentem bons resultados, a predição de estruturas secundárias e de funções de proteínas continuam sendo problemas desafiadores. Neste projeto, nós propomos três classificadores, sendo dois classificadores para estruturas secundárias (um classificador para classificação Q3 e outro classificador para classificação Q8), e um classificador para funções de proteínas. Para os classificadores de estruturas secundárias, utilizaremos redes neurais convolucionais para extrair informações locais e redes neurais recorrentes para extrair informações globais das proteínas. Para o classificador de funções de proteínas, utilizaremos redes neurais densas para analisar a sequência de aminoácidos e estruturas tridimensionais das proteínas. Para treinamento, validação e teste, usaremos a base de dados PDB. Como resultados iniciais, treinamos redes neurais recorrentes para a predição de estruturas secundárias e obtivemos 0,71 de acurácia balanceada, 0,73 de precisão balanceada, 0,73 de revocação balanceada e 0,73 de taxa F1 balanceada para a classificação Q3 e 0,38 de acurácia balanceada, 0,57 de precisão balanceada, 0,60 de revocação balanceada e 0,57 de taxa F1 balanceada para a classificação Q8.

Capítulo 1

Introdução

Este capítulo caracteriza o problema a ser investigado, apresenta os principais objetivos e contribuições do trabalho, bem como a organização do texto.

1.1 Caracterização do Problema

As proteínas são fundamentais em vários processos biológicos dos seres vivos, como regulação de reações químicas e resposta imunológica. As proteínas são formadas por cadeias de aminoácidos unidos por ligações peptídicas, que formam estruturas tridimensionais devido às interações químicas de atração entre os aminoácidos.

Com os avanços na tecnologia, principalmente na área biológica, tornou-se simples sequenciar os aminoácidos que formam a proteína, porém determinar as características da proteína, como as estruturas secundárias, estrutura terciária e funções, ainda requer muito esforço.

A análise das estruturas secundárias e terciárias da proteína é crucial para entender as funções e possíveis aplicações, como fabricação de remédios e biossensores [19, 48]. Algumas doenças, como fibrose cística, Alzheimer e outras doenças neurodegenerativas são atribuídas ao enovelamento errado da proteína [27], o que compromete sua função.

Para determinar as estruturas tridimensionais de uma proteína, métodos experimentais são necessários, como cristolografia por raio-X e ressonância magnética. Uma possível maneira de predizer a estrutura tridimensional completa de uma proteína desconhecida é determinar a sua estrutura secundária [15].

Devido ao custo de determinar as estruturas e funções das proteínas por métodos laboratoriais, diversas pesquisas têm sido feitas para encontrar técnicas capazes de obter bons resultados na classificação das estruturas secundárias das proteínas. Dentre as abordagens utilizadas, as tecnologias de informação vêm ganhando espaço na resolução de processos biológicos [14].

As abordagens utilizadas para classificação de estruturas secundárias são baseadas na interação entre aminoácidos próximos ao aminoácido analisado. A classificação de funções de proteínas utiliza a similaridade entre proteínas com funções desconhecidas com proteínas com funções conhecidas.

A base de dados online PDB[1] possui mais de 150.000 macromoléculas, como proteínas, ácidos desoxirribonucleicos (DNA) e ácidos ribonucleicos (RNA), e para cada uma delas a base possui as estruturas tridimensionais determinadas experimentalmente [36]. Para as proteínas da base, existem informações como sequência de aminoácidos, estruturas tridimensionais e funções.

1.2 Objetivos e Contribuições

O objetivo deste trabalho é desenvolver uma abordagem capaz de predizer estruturas secundárias e funções de proteínas e generalizar o aprendizado adquirido para novas entradas de dados.

Para desenvolver a metodologia proposta, alguns objetivos específicos devem ser alcançados:

- levantamento bibliográfico e estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas;
- aquisição da base de dados PDB;
- proposição de métodos para a classificação de estruturas secundárias e de funções de proteínas;
- realização de experimentos;
- avaliação e comparação do método proposto com outras abordagens disponíveis;
- documentação e publicação dos resultados.

Este projeto visa contribuir com o desenvolvimento de uma metodologia para a classificação de estruturas secundárias e funções de proteínas da base de dados PDB, bem como a generalização para a classificação de outras proteínas, não necessariamente depositadas no PDB.

1.3 Organização do Texto

O Capítulo 2 descreve os conceitos e técnicas relevantes relacionados ao tema sob investigação. O Capítulo 3 descreve a metodologia proposta, a base de dados, as métricas de avaliação e os recursos computacionais que serão empregados no desenvolvimento do projeto. O Capítulo 4 apresenta alguns resultados preliminares. O Capítulo 5 apresenta o plano de trabalho e o cronograma de execução das atividades.

[1] `https://www.wwpdb.org`

Capítulo 2

Revisão Bibliográfica

Este capítulo descreve conceitos e técnicas relevantes relacionados ao tema sob investigação.

2.1 Conceitos Biológicos

Esta seção apresenta conceitos biológicos utilizados para a predição de estrutura secundária e de funções de proteínas.

2.1.1 Proteínas

Proteínas são macromoléculas que estão presentes em todos os organismos vivos. As proteínas são formadas por sequências de aminoácidos unidos por ligações peptídicas. Elas formam a base de vida molecular e celular e possuem diversas funções, como proteção, regulação e transporte [27].

As proteínas possuem 4 estruturas. A estrutura primária consiste na sequência linear dos aminoácidos que formam a proteína [35]. A estrutura secundária da proteína ocorre devido às interações entre as ligações de hidrogênio dos aminoácidos, formando padrões (*motifs*) tridimensionais [27]. A estrutura terciária da proteína é representada pela estrutura tridimensional de todos os átomos dos aminoácidos da proteína [7]. A estrutura quaternária da proteína consiste em duas ou mais proteínas juntas que formam um complexo [35].

2.1.2 Estruturas Secundárias

As estruturas secundárias das proteínas são estruturas tridimensionais locais da proteína que ocorrem devido às ligações de hidrogênio. O enovelamento tridimensional formado é energeticamente eficaz [13].

As estruturas locais formadas se dividem em classes distintas. A classificação Q3 considera que existam 3 classes diferentes de estruturas e a classificação Q8 considera que há 8 classes de estruturas diferentes.

Na classificação Q8, existem as classes de estruturas H (*alpha helix*), G (*3-helix*), B (*residue in isolated beta bridge*), E (*extended strand*), I (*5-helix*), T (*hydrogenbonded turn*), S (*bend*) e L (*loop*) [24]. Na classificação Q3, há 3 classes, H (*helix*), E (*strand*) e C (*coil*) [10].

A classificação Q8 é mais desafiadora que a classificação Q3 [44]. A classificação Q8 pode ser vista como uma subclassificação da categoria Q3 [24]. A Tabela 2.1 mostra a classificação Q8 e a respectiva classe na classificação Q3.

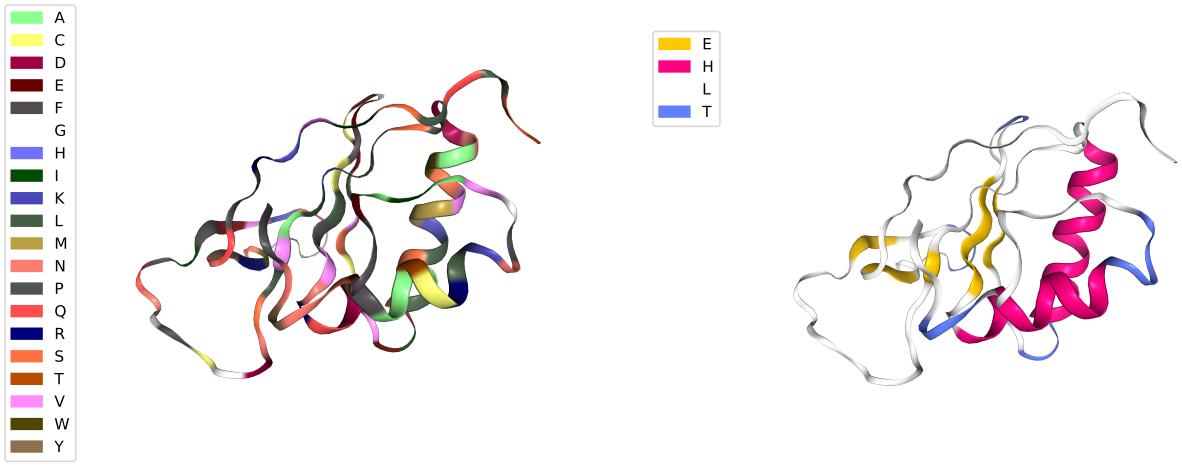
Q8	**Q3**
G	H
H	H
B	E
E	E
I	C
L	C
S	C
T	C

Tabela 2.1: Classificações Q8 e Q3.

A Figura 2.1 mostra a sequência de aminoácidos na forma tridimensional e estruturas secundárias formadas da proteína PDB ID: 6BI6. Cada aminoácido e estrutura secundária está representado por uma cor diferente. Na Figura 2.1, é possível perceber que a estrutura secundária formada por cada aminoácido depende dos aminoácidos que estão próximos a ele.

```
>6BI6:A|PDBID|CHAIN|SEQUENCE
GPTSLQLSIVHRLPQNYRWSAGFAGSKVEPIPQNGPCGDNSLVALKLLSPDGDNAWSVMYKLSQALSDIEVPCSVL
ECEGEPCLFVNRQDEFAATCRLKNFGVAIAEPFSNYNPF
```

(a) Sequência de aminoácidos.



(b) Sequência de aminoácidos na forma tridimensional.

(c) Estruturas secundárias.

Figura 2.1: Sequência de aminoácidos e estruturas secundárias da proteína PDB ID: 6BI6.

2.1.3 Funções das Proteínas

As funções das proteínas são definidas pelas estruturas tridimensionais, principalmente das estruturas secundária [48] e terciária [11]. Uma única proteína pode desempenhar funções diferentes em ambientes diferentes. Isso implica que cada proteína pode possuir mais de uma função [9]. Várias funções nas células, como suporte estrutural, mobilidade, proteção, regulação e transporte, são desenvolvidas pelas proteínas [27].

A classificação de funções das proteínas utiliza a ideia de que proteínas com origem evolutiva comum devem possuir as mesmas funções. Para realizar a predição de funções de proteínas, pode-se utilizar informações de similaridade de sequência de aminoácidos, similaridade de estruturas tridimensionais e interações com proteínas com funções conhecidas [38].

A ontologia genética (*Gene Ontology*[1]) é um projeto que visa criar e manter termos sobre funções e relações de genes. O modelo de ontologia genética apresenta três aspectos das funções: função molecular, processo biológico e componente celular [46]. Atualmente, existem 10.417 termos para funções moleculares, 4.022 termos para componentes celulares e 29.146 termos para processos biológicos [12].

2.2 Conceitos Computacionais

Esta seção apresenta conceitos computacionais utilizados nas abordagens de predição de estrutura secundária e de funções de proteínas encontradas na literatura.

2.2.1 Aprendizado de Máquina

O aprendizado de máquina consiste no reconhecimento de padrões por uma máquina. A máquina deve possuir um algoritmo que é capaz de aprender com resultados anteriores e fazer previsões para os próximos dados que irá analisar.

Muitas aplicações na sociedade utilizam aprendizado de máquina. Recomendações de itens em lojas eletrônicas, identificação de objetos em imagens e buscas filtradas em navegadores da Internet são alguns exemplos [21].

O aprendizado de máquina está presente na comunidade científica. Alguns exemplos de trabalhos recentes envolvendo aprendizado de máquina que podem ser citados são reconhecimento automático de fala e transformação de texto em áudio [34], reconhecimento de imagens [31] e predição de estruturas de proteínas [10].

2.2.2 Aprendizado Profundo

Com a dependência da extração de características para realizar o aprendizado de máquina, novas abordagens foram criadas para realizar a representação dos dados de forma automática [3]. Dentre os métodos para a representação de dados, o aprendizado profundo (*deep learning*) é capaz de representar dados em multiníveis de forma não linear, transformando-os

[1] `http://geneontology.org`

em representações no mais alto e abstrato nível. O aprendizado profundo está trazendo grandes avanços em muitos problemas da comunidade de inteligência artificial que não possuíam soluções [21].

Classificadores lineares podem dividir os dados em regiões muito simples, separadas por um hiperplano, porém a maioria dos problemas não é linearmente separável, criando dependência da extração de características importantes. Com isso, o uso de aprendizado profundo pode auxiliar na representação em alto nível dos dados.

2.2.3 Redes Neurais Profundas

A rede neural foi criada com base no cérebro. As unidades básicas do cérebro, conhecidas como neurônios, realizam ligações entre eles, formando uma rede de neurônios. Um sinal do neurônio anterior chega nos dendritos do neurônio atual, passa pelo corpo celular e sai pelo axônio para o próximo neurônio. Na rede neural, cada neurônio possui uma função de ativação pré-definida e cada ligação entre neurônios possui um peso.

Na rede neural, os neurônios são separados em camadas. A primeira camada, chamada de camada de entrada, recebe os valores dos dados em questão e passa a saída dos neurônios para a entrada dos neurônios da próxima camada. As camadas intermediárias da rede são chamadas de camadas ocultas. A última camada é chamada de camada de saída. A Figura 2.2 mostra uma rede neural com 5 neurônios na camada de entrada, 7 neurônios na camada oculta e 3 neurônios na camada de saída.

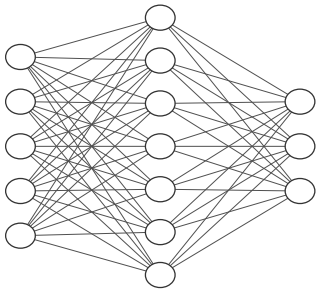


Figura 2.2: Rede neural com 3 camadas.

Durante o treinamento da rede, os pesos, que são os valores de ligação entre os neurônios, devem ser otimizados para reduzir a função de perda. Essa otimização é chamada de descida do gradiente [40]. O algoritmo de retropropagação do erro (*backpropagation*) calcula quanto que o erro é afetado para cada um dos pesos e realiza a otimização, realizando o percurso contrário da etapa de treinamento, ou seja, da camada de saída para a camada de entrada.

Um importante aspecto que deve ser evitado é o sobreajuste (*overfitting*) da rede, onde esta se ajusta bem ao conjunto de dados do treinamento, porém não é capaz de classificar novos dados durante os testes. Para evitar o sobreajuste, pode-se utilizar a técnica de *dropout*, que faz com que certa quantidade de neurônios sejam desligados durante o processo de treinamento [39].

Para ser considerada uma rede neural densa, que utiliza aprendizado profundo, uma rede neural deve possuir mais do que uma camada oculta, ou seja, no mínimo quatro camadas.

Hoje em dia, uma rede neural profunda típica usada em aprendizado profundo tem entre cinco até centenas de camadas [40].

Algumas variações da rede neural profunda que estão ligadas ao problema em questão serão apresentadas a seguir.

2.2.4 Redes Neurais Recorrentes

Redes neurais possuem alto poder computacional para resolver problemas. Porém, existem algumas limitações nesse método, como dados que dependem de tempo ou sequência [25]. Alguns exemplos de dados que atingem essa limitação são quadros de um vídeo, sequência de textos e sequências de notas de uma música.

Para resolver essa limitação, foi criada a rede neural recorrente, que é uma rede capaz de analisar dados de uma sequência temporal ou espacial. Para isso, é necessária uma memória interna em cada nó da rede para ter acesso às informações históricas da rede.

Um dos métodos utilizados para guardar a informação é a unidade LSTM (*Long Short-Term Memory* [16]), onde as células de memória são capazes de lembrar informações passadas. O LSTM aprimorou as redes recorrentes com informações locais e globais [43]. Outro neurônio capaz de guardar informações da rede é a unidade GRU(*Gated Recurrent Unit* [8]). A unidade GRU consegue resultados próximos dos resultados da unidade LSTM, porém utiliza menos parâmetros [23].

Além de ligações com os neurônios da camada seguinte, existem ligações entre os neurônios da mesma camada, fazendo com que a informação da rede em um momento anterior continue naquela camada. Os neurônios de uma camada podem possuir ligações em uma única direção (esquerda para direita ou direita para esquerda) ou possuir ligações bidirecionais. Na análise de sequências de aminoácidos para a definição de estruturas, deve-se analisar aminoácidos tanto anteriores ao aminoácido que está sendo analisado quanto os próximos aminoácidos, portanto, ligações bidirecionais são mais adequadas a esse problema [15].

2.2.5 Redes Neurais Convolucionais

As redes neurais convolucionais foram inspiradas no córtex visual dos animais. Este tipo de rede é invariante à translação, escala e distorção [22]. As primeiras camadas da rede neural convolucional são capazes de reconhecer características locais e as camadas mais distantes, ou seja, camadas mais próximas do final da rede, reconhecem características globais.

Redes neurais convolucionais possuem camadas de convolução, que são responsáveis pela aplicação de filtros nos dados, camadas de *pooling*, que são responsáveis pela diminuição da dimensão dos dados, e camadas totalmente conectadas, como as camadas ocultas da rede neural.

Normalmente, as redes neurais convolucionais são utilizadas em problemas de visão computacional, porém estudos recentes em outras áreas, como classificação de estruturas secundárias de proteínas [24, 26, 48], mostraram resultados promissores.

2.3 Trabalhos Correlatos

Nesta seção, são apresentados trabalhos que buscam resolver a classificação de estruturas secundárias e de funções de proteínas, respectivamente.

2.3.1 Classificação de Estruturas Secundárias

Lin et al. [24] construíram uma rede densa para a classificação Q3, Q8, acessibilidade relativa de solvente (*Relative Solvent Accessibility* - SAR) e acessibilidade absoluta de solvente (*Absolute solvent accessibility* - SAA). A rede criada utiliza camadas de convolução, *pooling* e *dropout*, além de aplicar a técnica de *shift-and-stitch*. As bases de dados utilizadas para experimentos foram CullPDB (base de dados utilizada para avaliar algoritmos de predição de estruturas [48]), CB513 (base de dados com 513 proteínas com estruturas tridimensionais e que não possuem relação entre elas [32]) e 4prot (base de dados derivada da anterior [32]). O treinamento foi realizado na base CullPDB e testes na base CB513, além de treinamento e testes na base 4prot. Os resultados obtidos foram:

- Classificação Q3: 89,6% de acurácia na base 4prot.
- Classificação Q8: 76,7% de acurácia na base 4prot e 68,4% de acurácia na base CB513.
- Classificação SAR: 84,9% de acurácia na base 4prot.
- Classificação SAA: 86,1% de acurácia na base 4prot.

Hattori et al. [15] construíram uma rede para classificar a categoria Q8. A arquitetura consiste em uma rede neural recorrente bidirecional densa com LSTM e camadas de *dropout*. As bases de dados utilizadas para experimentos foram CullPDB para treinamento e CB513 para testes. Como resultado, a arquitetura proposta atingiu 68,0% de acurácia na base CB513.

Fang et al. [10] classificaram estruturas secundárias nas categorias Q3 e Q8 utilizando uma rede densa. A arquitetura da rede criada consiste em *inceptions* de camadas convolucionais em paralelo, com uma camada de concatenação ao final dos *inceptions*, seguida por camadas densas. Os experimentos foram realizados nas bases de dados CullPDB, JPRED (que é uma base de dados que todas as proteínas pertencem a super-famílias diferentes, permitindo resultados mais objetivos [10]), CASP10, CASP11, CASP12 (que são bases de uma organização que conduz experimentos para mensurar o estado da arte em modelar estruturas tridimensionais a partir de sequências de proteínas [29]), CB513 e nos arquivos disponibilizados de 1 de Julho de 2017 até 15 de Agosto de 2017 no PDB. Os arquivos do PDB foram divididos em duas classes, casos fáceis, que são proteínas que possuem o *e-value* menor ou igual a 0,5 quando comparadas com proteínas do CullPDB, e casos difíceis, quando o valor do *e-value* é maior que 0,5. O treinamento da rede foi feito com 9.000 proteínas da base CullPDB. Os resultados obtidos nas bases de dados foram:

- Classificação Q3: 85,98% de acurácia na base CASP10, 83,59% de acurácia na base CASP11, 80,59% de acurácia na base CASP12, 88,20% de acurácia nos casos fáceis do PDB, 83,37% de acurácia nos casos difíceis do PDB.

- Classificação Q8: 70,63% de acurácia na base CB513, 77,10% de acurácia na base CASP10, 73,92% de acurácia na base CASP11, 70,48% de acurácia na base CASP12, 78,65% de acurácia nos casos fáceis do PDB, 72,84% de acurácia nos casos difíceis do PDB.

2.3.2 Classificação de Funções

Nadzirin e Firdaus-Raih [30] classificaram proteínas com funções desconhecidas da base de dados PDB. Para isso, foram realizadas pesquisas na base de dados UniProtKB (base de dados que contém a sequência de aminoácidos das proteínas [41]), e consultas nas ferramentas BLAST (ferramenta que pesquisa similaridade entre sequências de proteínas [1]) e DALI, (ferramenta que pesquisa similaridade entre estruturas tridimensionais de proteínas [17]), visando encontrar proteínas com funções já citadas na literatura ou homólogas às proteínas com funções conhecidas. Inicialmente, foram adquiridas proteínas do PDB com função desconhecida (*Unknown Function*), totalizando 2.549 proteínas. Após pesquisa na base de dados UniProtKB e consultas nas ferramentas BLAST e DALI, 1.465 das 2.549 possuíam anotações de funções ou similaridade com alguma proteína com função conhecida. Portanto, das 2.549 proteínas iniciais, apenas 1.084 realmente eram proteínas com funções desconhecidas.

Roy et al. [37] criaram uma plataforma para predizer estrutura tridimensional e funções de proteínas a partir da sequência de aminoácidos. Inicialmente, a plataforma encontra *templates* de proteínas que possuam estruturas similares. Após esta etapa, são criadas simulações de enovelamento utilizando informações do PDB, como correlações entre sequências e hidrofobicidade, além de outras informações, como restrições espaciais dos modelos de enovelamento. Após criadas as simulações de enovelamento, é realizada a identificação de enovelamentos com pouca energia livre, criando centroides das médias das coordenadas 3D da estrutura. Para predizer a função da proteína da consulta, é realizada a comparação entre a estrutura tridimensional encontrada com proteínas com estruturas tridimensionais e funções conhecidas do PDB. Para isso, os autores construíram três conjuntos de termos das proteínas analisadas, sendo um conjunto com os números EC (classificação numérica de enzimas [2]), um conjunto com ontologia genética e um conjunto com *ligand-binding sites*.

Zhang et al. [47] construíram uma arquitetura com aprendizado profundo para a predição de função de proteínas utilizando sequência de aminoácidos de proteínas similares e interações entre proteínas. A arquitetura proposta recebe a sequência de aminoácidos e as características selecionadas pela ferramenta InterPro (ferramenta para classificar sequências de proteínas em famílias e informações como domínio proteico [28]) e as interações das proteínas. As informações são concatenadas em uma camada densa de características. A base de dados utilizada para treinamento é derivada do trabalho de Kulmanov et al. [20], e a base de dados para teste é a CAFA3 (banco de dados criado para o desafio CAFA, onde são analisados e avaliados os métodos para predição de função de proteínas [33]). Como resultado, a arquitetura atingiu 67,0% de precisão média e 48,0% de revocação média, obtendo melhores resultados que o método de predição de função de proteínas que utiliza aprendizado profundo DeepGO [20].

Capítulo 3

Material e Métodos

Este capítulo descreve a metodologia, a base de dados, as métricas de avaliação e os recursos computacionais que serão utilizados no desenvolvimento do projeto.

3.1 Metodologia

Esta seção apresenta a metodologia proposta dividida em duas partes, predição de estruturas secundárias e predição de funções de proteínas.

3.1.1 Predição das Estruturas Secundárias

As subseções a seguir apresentam as etapas para a realização da predição de estruturas secundárias de proteínas. A metodologia proposta é apresentada na Figura 3.1.

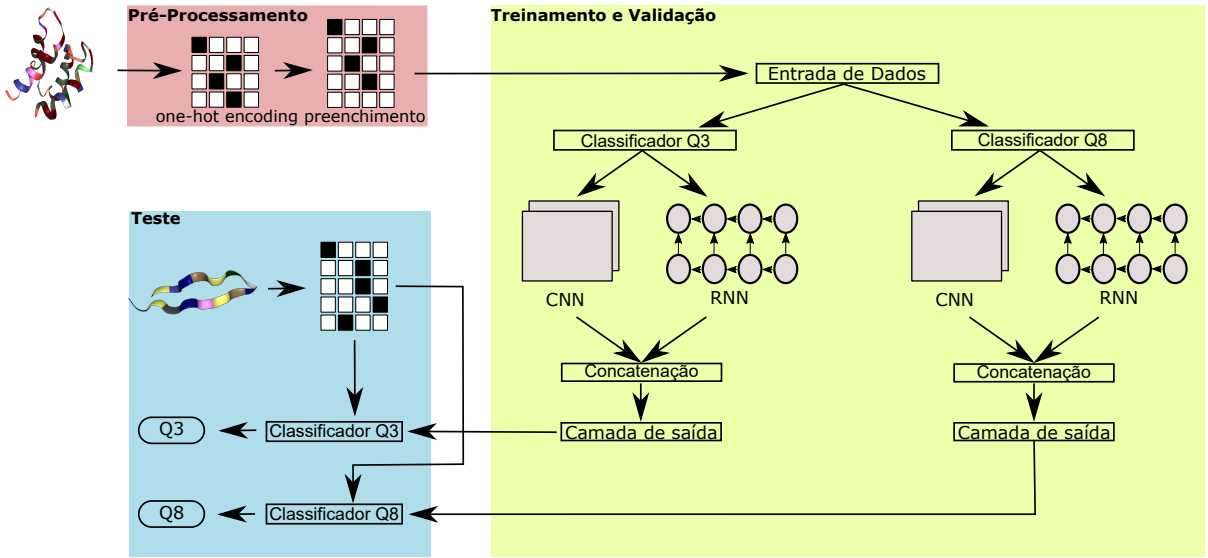


Figura 3.1: Diagrama das etapas da metodologia para predição de estruturas secundárias.

Pré-Processamento

Durante a etapa de pré-processamento das proteínas para a predição de estruturas secundárias, selecionaremos as proteínas do PDB com até 700 aminoácidos, já que 96,73% das proteínas do PDB possuem 700 aminoácidos ou menos. Proteínas que possuem menos que 700 aminoácidos receberão um preenchimento (*padding*), como utilizado nos trabalhos de Fang et al. [10] e Zhou e Troyanskaya [48].

Ao todo, existem 20 aminoácidos diferentes que podem formar uma proteína e cada um dos aminoácidos é representado por uma única letra. Alguns aminoácidos possuem casos especiais, como Alanina, que é representada pelas letras "A" e "X", Asparagina, que é representada pelas letras "N" e "B" e Glutamina, que é representada pelas letras "Q" e "Z". Nestes casos, trataremos o aminoácido "X" como "A", o aminoácido "B" como "N" e o aminoácido "Z" como "Q", como foi feito no trabalho de Fang et al. [10]. Caso alguma proteína apresente algum aminoácido que não possua alguma letra conhecida, esta proteína será desconsiderada.

A geração de características das proteínas para a classificação consiste em transformar os aminoácidos em vetores de características de tamanho igual a 20, utilizando o *one-hot encoding*. O vetor de características do preenchimento possuirá todos os valores iguais a 0. Outras características serão avaliadas para aumentar o vetor de características dos aminoácidos da base de dados. As estruturas secundárias serão obtidas dos arquivos no PDB utilizando a ferramenta DSSP [18, 42].

Treinamento e Validação

A base será dividida em duas partes, 80% para treinamento/validação e 20% para teste. Para o treinamento e a validação, utilizaremos o método de validação cruzada *k-fold*. Os atributos da rede serão avaliados para que possamos encontrar os melhores parâmetros.

As arquiteturas dos classificadores consistem em camadas convolucionais, responsáveis pela obtenção de informações locais das sequências de aminoácidos, e camadas recorrentes, responsáveis pelas informações globais das sequências. Serão avaliados diversos filtros unidimensionais das camadas convolucionais e diversas unidades das redes recorrentes, como LSTM, GRU e *Neural Turing Machines*.

Teste

Durante a fase de teste, iremos submeter o vetor de características de uma proteína que não foi apresentada durante o treinamento e validação para os classificadores Q3 e Q8. Os resultados obtidos serão avaliados pelas métricas apresentadas na Seção 3.3.

3.1.2 Predição das Funções

As subseções a seguir apresentam as etapas para a realização da predição de funções de proteínas. As etapas da metodologia proposta são apresentadas na Figura 3.2.

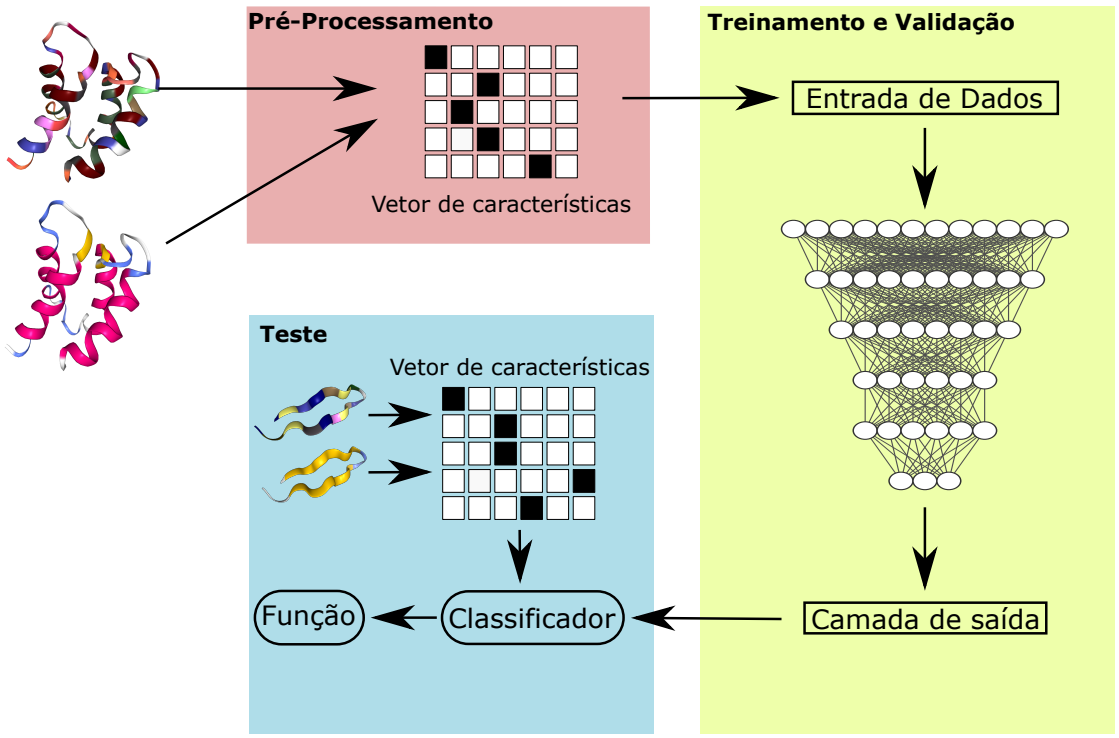


Figura 3.2: Diagrama das etapas da metodologia para predição de funções.

Pré-Processamento

Inicialmente, construiremos a base de treinamento utilizando proteínas que possuem os termos da função molecular da ontologia genética. As proteínas utilizadas serão selecionadas da base de dados PDB. O vetor de características das proteínas conterá informações da sequência de aminoácidos e da estrutura tridimensional da proteína.

Treinamento e Validação

Dividiremos a base em 80% para treinamento/validação e 20% para teste. Durante o treinamento, utilizaremos o método de validação cruzada *k-fold*. Os atributos da rede serão avaliados para que possamos encontrar os melhores parâmetros.

A arquitetura proposta do classificador de funções consiste em camadas densas onde serão analisadas as sequências de aminoácidos e estruturas tridimensionais das proteínas. Escolhemos camadas densas pois são capazes de representar as características dos dados de forma não linear.

Teste

Durante a etapa de testes, iremos submeter o vetor de características de uma proteína que não foi utilizada para treinamento e validação. As métricas de avaliação utilizadas para avaliar o classificador são apresentadas na Seção 3.3.

3.2 Base PDB

O PDB (*Protein Data Bank*) é um repositório criado em 1971 que reúne informações de estruturas 3D de proteínas, ácidos nucleicos e macromoléculas complexas. Inicialmente, o PDB foi hospedado no *Brookhaven National Laboratory* (BNL) [5].

Em Outubro de 1998, o PDB passou para a responsabilidade da *Research Collaboratory for Structural Bioinformatics* (RCSB) [6]. Em 2003, foi criada a fundação de colaboradores wwPDB (*WorldWide Protein Data Bank*)[1], que é responsável pela manutenção do PDB atualmente [4].

O wwPDB possui parceiros na América (RCSB PDB)[2], na Europa (PDBe)[3], no Japão (PDBj)[4], além do BMRB (*Biological Magnetic Resonance Data Bank*)[5]. Todos os parceiros colaboram com questões de arquivamento, políticas de depósito e anotação, formatos, padrões e atualizações semanais e cada um deles possui um site próprio, de modo que os dados são apresentados para diversas comunidades [5].

O início do PDB foi caracterizado por buscar interesse da comunidade em depositar as estruturas das proteínas [4]. A partir das décadas de 1980 e 1990, houve um enorme crescimento de depósitos de dados no PDB, conforme mostra a Figura 3.3. Atualmente, o PDB conta com mais de 150.000 estruturas, totalizando cerca de 1 TB de dados e continua crescendo a cada semana. Toda Quarta-Feira, os novos dados ficam disponíveis para *download*.

Há três formatos disponíveis no PDB. O primeiro utilizado foi o formato PDB, sendo simples de ler por humanos e utilizado em muitas aplicações computacionais [4]. Devido à limitação do formato PDB para aplicações mais modernas, foi criado o formato mmCIF (*macromolecular Crystallographic Information File*). Esse formato é complementamente legível computacionalmente e pode ser utilizado em bancos de dados relacionais [4]. O dicionário de palavras criado para o mmCIF foi utilizado para a criação do formato PDBML, com base na linguagem de marcação XML [45].

A versão do PDB utilizada para este projeto foi adquirida em 03/05/2019. Essa versão possui 150.736 estruturas ao total, sendo que 147.169 são proteínas e 3.852 delas possuem *Unknown Function* dentre as suas classificações.

3.3 Métricas de Avaliação

O desempenho da metodologia proposta será avaliado utilizando as métricas quantitativas citadas a seguir.

A precisão é representada pela Equação 3.1, onde VP é o número de verdadeiros positivos e FP é o número de falsos positivos. Pela Equação 3.1, é possível perceber que a precisão é

[1] `https://www.wwpdb.org`
[2] `https://www.rcsb.org`
[3] `https://www.ebi.ac.uk/pdbe`
[4] `https://pdbj.org`
[5] `http://www.bmrb.wisc.edu`

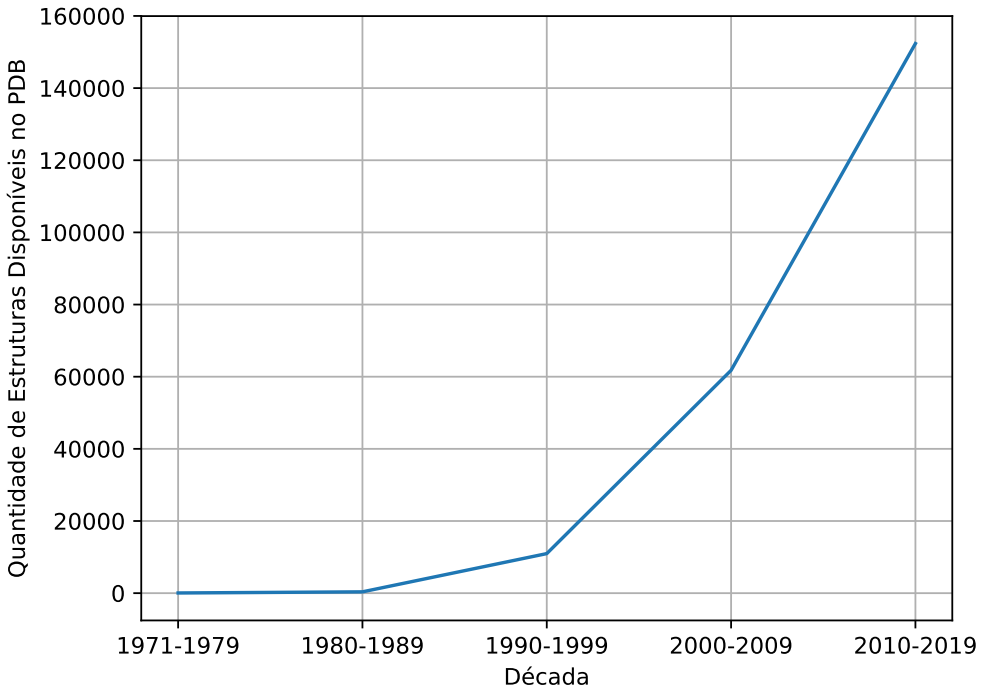


Figura 3.3: Evolução do número de proteínas disponíveis no PDB.

a capacidade de classificar como positivo um dado que realmente é positivo.

$$\text{Precisão} = \frac{\text{VP}}{\text{VP+FP}} \tag{3.1}$$

A taxa de revocação é dada pela Equação 3.2, onde VP é o número de verdadeiros positivos e FN é o número de falsos negativos. Essa medida verifica a capacidade do classificador de classificar corretamente dados positivos.

$$\text{Revocação} = \frac{\text{VP}}{\text{VP+FN}} \tag{3.2}$$

A taxa F1 é mostrada pela Equação 3.3. Pela Equação 3.3, a taxa F1 corresponde à média harmônica da precisão e da revocação.

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.3}$$

A acurácia é representada pela Equação 3.4, onde VP é o número de verdadeiros positivos, FP é o número de falsos positivos, P é o número de casos positivos e N é o número de casos negativos. Essa métrica obtém a porcentagem de decisões corretas atingidas pelo classificador.

$$\text{Acurácia} = \frac{\text{VP+VN}}{\text{P+N}} \tag{3.4}$$

No caso de dados desbalanceados, deve-se utilizar as métricas de avaliação de forma balanceada. A Equação 3.5 representa a forma balanceada das métricas de avaliação utilizadas

(precisão, revocação, taxa F1 e acurácia), onde y_i é a métrica avaliada da i-ésima classe e n é a quantidade de classes.

$$\text{Métrica Balanceada} = \frac{\sum_{i}^{n} y_i}{n} \tag{3.5}$$

3.4 Recursos Computacionais

A implementação deste projeto será feita em linguagem de programação Python, devido ao grande número de bibliotecas disponíveis e com boa documentação. O projeto utilizará bibliotecas de aprendizado de máquina, aprendizado profundo, funções científicas e numéricas e apresentação de gráficos. Algumas bibliotecas que podem ser destacadas são: `NumPy`[6], `scikit-learn`[7], `TensorFlow`[8], `Keras`[9] e `Matplotlib`[10].

Os experimentos deste projeto serão realizados em uma máquina equipada com processador Intel i5-6400 com 2.70 GHz e uma GPU GeForce GTX 1060, com 1280 núcleos CUDA, memória padrão DDR5 de 6 GB.

[6] `https://www.numpy.org`
[7] `https://scikit-learn.org`
[8] `https://www.tensorflow.org`
[9] `https://keras.io`
[10] `https://matplotlib.org`

Capítulo 4

Resultados Preliminares

Este capítulo apresenta os resultados preliminares da predição de estruturas secundárias da proteína.

4.1 Experimentos Iniciais

A base usada nos experimentos consiste em uma amostragem do PDB com 2.000 proteínas com tamanho menor ou igual a 700 aminoácidos. Esse tamanho foi escolhido pois representa cerca de 96,73% das proteínas do PDB. A frequência de aminoácidos da amostragem segue próximo a frequência de proteínas com até 700 aminoácidos do PDB, como mostra a Tabela 4.1.

	Amostragem		**PDB**	
Tamanho	**Quantidade**	**Frequência (%)**	**Quantidade**	**Frequência (%)**
[0, 100]	245	12,25	17.585	11,95
(100, 200]	512	25,60	38.449	26,12
(200, 300]	545	27,25	38.577	26,21
(300, 400]	369	18,45	26.073	17,71
(400, 500]	183	9,15	12.756	8,67
(500, 600]	106	5,30	6.095	4,14
(600, 700]	40	2,00	2.800	1,93
(700, 1000]	-	-	3.639	2,47
(1000, 1200]	-	-	640	0,43
(1200, 4000]	-	-	555	0,37

Tabela 4.1: Quantidade e frequência de proteínas na amostragem e no PDB.

Para o vetor de características, transformamos as siglas correspondentes a cada aminoácido na forma *one-hot encoding*. Dessa forma, cada aminoácido possui um vetor de características com tamanho 20, com 19 valores iguais a 0 e 1 valor igual a 1. Para as proteínas com tamanho menor que 700 aminoácidos, foi utilizado o preenchimento (*padding*) até atingir 700 aminoácidos. O vetor de características do preenchimento consiste em 20 valores iguais a 0. A classificação de estruturas secundárias foi feita para cada aminoácido das proteínas.

A amostragem do PDB utilizada é desbalanceada, tanto na classificação Q3, como mostra a Tabela 4.2, quanto na classificação Q8, conforme a Tabela 4.3. Em comparação com os dados do PDB, a proporção de estruturas secundárias utilizadas na amostragem atinge frequências próximas tanto na classificação Q3 quanto na categoria Q8.

	Amostragem		**PDB**	
Classe	**Quantidade**	**Frequência (%)**	**Quantidade**	**Frequência (%)**
C	211.598	41,02	16.638.946	41,21
E	115.196	22,33	8.901.535	22,05
H	189.014	36,65	14.833.903	36,74

Tabela 4.2: Quantidade e frequência de cada classe da categoria Q3 na amostragem e no PDB.

	Amostragem		**PDB**	
Classe	**Quantidade**	**Frequência (%)**	**Quantidade**	**Frequência (%)**
B	6.264	1,21	482.448	1,19
E	108.932	21,12	8.419.087	20,85
G	19.781	3,85	1.561.805	3,87
H	169.143	32,79	13.272.098	32,87
I	3.517	0,68	259.370	0,64
L	104.085	20,18	8.172.641	20,24
S	45.588	8,84	3.634.659	9,00
T	58.408	11,33	4.572.276	11,34

Tabela 4.3: Quantidade e frequência de cada classe da categoria Q8 na amostragem e no PDB.

Para treinamento e validação, utilizamos o método de validação cruzada 5-*fold*, separando o conjunto em 80% para treinamento e 20% para validação. Cada *fold* foi treinado por 100 épocas, com *batch size* igual a 32.

Para as classificações Q3 e Q8, testes com a rede recorrente foram realizados com camadas unidirecionais e bidirecionais, tanto com unidades LSTM quanto GRU. Em cada camada, utilizamos 100 unidades e *dropout* igual a 0,5. A quantidade de camadas variaram de 1 até 10. A camada de saída contém unidades com função de ativação *softmax*. Na predição Q3, a última camada possui 3 neurônios, enquanto na classificação Q8, a última camada contém 8 unidades. Em ambas, foi utilizado o *one-hot encoding* para transformar as estruturas secundárias de forma categórica. A Figura 4.1 apresenta a estrutura geral da rede utilizada.

Durante o treinamento, validação e teste com camadas unidirecionais, as proteínas eram apresentadas no sentido correto, ou seja, no sentido em que a proteína é encontrada no arquivo do PDB, e de forma invertida, onde os últimos aminoácidos de uma proteína se tornam os primeiros aminoácidos e os primeiros aminoácidos se tornam os últimos. Os resultados obtidos na etapa de teste da classificação Q3 são apresentados na Tabela 4.4. Os resultados obtidos na etapa de teste da classificação Q8 são apresentados na Tabela 4.5. Vale ressaltar que as métricas utilizadas estão balanceadas, já que a amostragem é desbalanceada.

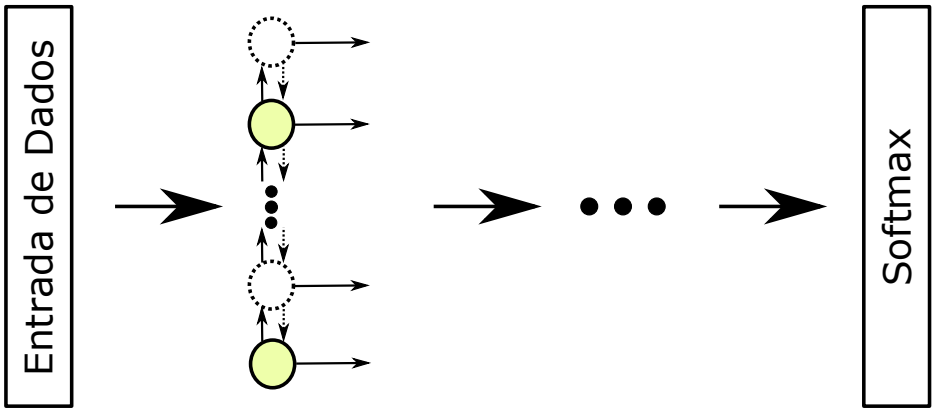


Figura 4.1: Rede recorrente utilizada nos experimentos.

	LSTM				**GRU**			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,56	0,58	0,58	0,58	0,57	0,58	0,59	0,58
2	0,57	0,59	0,59	0,58	0,57	0,59	0,59	0,59
3	0,58	0,59	0,59	0,59	0,58	0,60	0,60	0,60
4	0,58	0,59	0,60	0,59	0,58	0,60	0,60	0,60
5	0,58	0,53	0,56	0,55	0,58	0,60	0,60	0,60
6	0,52	0,53	0,55	0,52	0,58	0,59	0,60	0,59
7	0,52	0,53	0,55	0,52	0,58	0,59	0,60	0,59
8	0,47	0,44	0,51	0,45	0,58	0,59	0,59	0,59
9	0,48	0,45	0,51	0,46	0,58	0,59	0,59	0,59
10	0,49	0,47	0,52	0,47	0,58	0,59	0,59	0,59

(a) Redes unidirecionais com proteínas no sentido encontrado no PDB.

	LSTM				**GRU**			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,60	0,62	0,62	0,62	0,60	0,62	0,63	0,62
2	0,60	0,62	0,63	0,62	0,61	0,63	0,63	0,63
3	0,60	0,63	0,63	0,63	0,61	0,63	0,64	0,63
4	0,61	0,63	0,63	0,63	0,61	0,63	0,64	0,63
5	0,61	0,60	0,63	0,63	0,61	0,63	0,64	0,63
6	0,60	0,63	0,63	0,62	0,60	0,63	0,63	0,63
7	0,60	0,63	0,63	0,63	0,61	0,63	0,63	0,63
8	0,60	0,62	0,63	0,62	0,61	0,63	0,63	0,63
9	0,54	0,53	0,58	0,54	0,61	0,63	0,63	0,63
10	0,57	0,53	0,53	0,46	0,61	0,63	0,63	0,63

(b) Redes unidirecionais com proteínas no sentido invertido.

Tabela 4.4: Resultados obtidos na classificação Q3 com as redes unidirecionais.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,23	0,42	0,46	0,41	0,23	0,41	0,46	0,41
2	0,24	0,44	0,46	0,41	0,24	0,43	0,47	0,42
3	0,25	0,45	0,47	0,42	0,24	0,44	0,47	0,43
4	0,25	0,45	0,47	0,42	0,24	0,44	0,47	0,42
5	0,24	0,42	0,45	0,40	0,24	0,44	0,47	0,42
6	0,21	0,37	0,43	0,36	0,24	0,44	0,47	0,42
7	0,15	0,32	0,36	0,25	0,24	0,43	0,47	0,42
8	0,13	0,31	0,33	0,19	0,24	0,39	0,47	0,41
9	0,14	0,30	0,35	0,22	0,24	0,40	0,47	0,42
10	0,13	0,30	0,34	0,20	0,24	0,40	0,47	0,42

(a) Redes unidirecionais com proteínas no sentido encontrado no PDB.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,25	0,44	0,49	0,44	0,25	0,45	0,49	0,44
2	0,27	0,47	0,49	0,45	0,27	0,46	0,50	0,45
3	0,27	0,46	0,47	0,44	0,27	0,48	0,50	0,46
4	0,28	0,48	0,50	0,46	0,27	0,46	0,50	0,46
5	0,27	0,47	0,50	0,45	0,26	0,46	0,50	0,46
6	0,26	0,47	0,49	0,45	0,26	0,45	0,50	0,46
7	0,15	0,16	0,36	0,24	0,26	0,46	0,50	0,46
8	0,15	0,19	0,36	0,22	0,26	0,45	0,50	0,45
9	0,13	0,13	0,33	0,17	0,26	0,45	0,50	0,45
10	0,12	0,10	0,33	0,16	0,26	0,45	0,50	0,45

(b) Redes unidirecionais com proteínas no sentido invertido.

Tabela 4.5: Resultados obtidos na classificação Q8 com as redes unidirecionais.

Nas redes bidirecionais, fizemos testes com as proteínas sendo apresentadas tanto no sentido em que aparecem no PDB, quanto no sentido inverso. Os resultados obtidos da categoria Q3 são mostrados na Tabela 4.6 e os resultados obtidos da classificação Q8 são apresentados na Tabela 4.7. Novamente, as métricas estão na forma balanceada.

Em seguida, combinamos as predições de estruturas secundárias com as proteínas apresentadas no sentido encontrado no PDB e no sentido invertido, tanto com redes unidirecionais quanto com redes bidirecionais. A classe com maior probabilidade da combinação dos classificadores era escolhida, independente da direção. Os resultados obtidos na classificação Q3 são apresentados na Tabela 4.8 e os resultados obtidos na classificação Q8 são mostrados na Tabela 4.9.

Pelos resultados obtidos, observamos que as unidades GRU atingiram resultados melhores que as unidades LSTM. Constatamos também que as redes bidirecionais tiveram uma performance melhor que as redes unidirecionais tanto na classificação Q3 quanto na classificação Q8. Além disso, a combinação de classificadores possibilitou melhorias nos resultados.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,66	0,68	0,68	0,68	0,66	0,68	0,68	0,68
2	0,67	0,68	0,68	0,68	0,69	0,71	0,71	0,71
3	0,67	0,69	0,69	0,68	0,69	0,71	0,71	0,71
4	0,67	0,69	0,69	0,69	0,70	0,71	0,71	0,71
5	0,67	0,69	0,69	0,69	0,70	0,71	0,71	0,71
6	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
7	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
8	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
9	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
10	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71

(a) Redes bidirecionais com proteínas no sentido encontrado no PDB.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,66	0,68	0,68	0,68	0,67	0,68	0,68	0,68
2	0,66	0,68	0,68	0,68	0,69	0,72	0,71	0,71
3	0,67	0,69	0,69	0,68	0,69	0,72	0,71	0,71
4	0,67	0,69	0,69	0,69	0,70	0,71	0,71	0,71
5	0,67	0,69	0,69	0,69	0,70	0,71	0,71	0,71
6	0,68	0,69	0,69	0,69	0,71	0,72	0,72	0,72
7	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
8	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
9	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71
10	0,68	0,69	0,69	0,69	0,70	0,71	0,71	0,71

(b) Redes bidirecionais com proteínas no sentido invertido.

Tabela 4.6: Resultados obtidos na classificação Q3 com as redes bidirecionais.

4.2 Dados das Estruturas Secundárias no PDB

Realizamos uma análise dos tamanhos das estruturas secundárias nas sequências depositadas no PDB. Para cada tipo de estrutura secundária, calculamos os valores de média, mediana, valor mínimo, valor máximo e desvio padrão das categorias Q3 e Q8, que podem ser observadas na Tabela 4.10. Vale ressaltar que estruturas isoladas, isto é, que apresentam estruturas diferentes tanto antes quanto depois da estrutura analisada, foram consideradas sequências de tamanho igual a 1.

Outra hipótese estudada consiste na possível existência de padrões de estruturas no início, meio e fim das proteínas. Para esta análise, consideramos os valores de início e fim iguais a 50 aminoácidos. Utilizamos proteínas com no mínimo 150 aminoácidos para que não houvesse casos onde uma mesma estrutura estivesse contida na categoria início e na categoria fim. Os resultados obtidos são apresentados na Tabela 4.11. Constatamos que a probabilidade de algumas estruturas aparecerem em certas regiões é maior que em outras, como a estrutura H na classificação Q3, que aparece com maior probabilidade no centro das proteínas, e a

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,29	0,51	0,55	0,50	0,30	0,51	0,55	0,51
2	0,35	0,53	0,55	0,52	0,34	0,55	0,58	0,55
3	0,36	0,53	0,56	0,53	0,37	0,56	0,59	0,56
4	0,36	0,55	0,56	0,54	0,37	0,56	0,59	0,56
5	0,37	0,54	0,56	0,54	0,38	0,56	0,59	0,57
6	0,36	0,54	0,56	0,53	0,38	0,56	0,59	0,56
7	0,37	0,54	0,56	0,53	0,38	0,56	0,59	0,56
8	0,37	0,55	0,56	0,54	0,38	0,56	0,59	0,56
9	0,37	0,55	0,56	0,54	0,38	0,55	0,59	0,56
10	0,37	0,55	0,56	0,54	0,38	0,56	0,59	0,56

(a) Redes bidirecionais com proteínas no sentido encontrado no PDB.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,29	0,51	0,55	0,50	0,30	0,51	0,55	0,51
2	0,34	0,53	0,55	0,52	0,35	0,55	0,58	0,55
3	0,36	0,54	0,56	0,53	0,37	0,56	0,59	0,56
4	0,36	0,55	0,56	0,54	0,38	0,56	0,59	0,56
5	0,37	0,55	0,56	0,54	0,38	0,56	0,59	0,56
6	0,36	0,54	0,56	0,53	0,38	0,56	0,59	0,56
7	0,37	0,54	0,56	0,53	0,38	0,56	0,59	0,56
8	0,37	0,55	0,56	0,54	0,38	0,55	0,59	0,56
9	0,37	0,55	0,56	0,54	0,37	0,55	0,59	0,56
10	0,37	0,55	0,56	0,53	0,38	0,56	0,59	0,56

(b) Redes bidirecionais com proteínas no sentido invertido.

Tabela 4.7: Resultados obtidos na classificação Q8 com as redes bidirecionais.

estrutura E na classificação Q8, que aparece com maior probabilidade no início das proteínas. Em alguns casos, a probabilidade de uma estrutura aparecer no início, meio e final das proteínas se mantém, como a estrutura C da classificação Q3 e a estrutura S da classificação Q8. Observamos também que todas as proteínas do PDB apresentam a estrutura secundária L (da categoria Q8, logo a estrutura C na categoria Q3) como a primeira e a última estrutura.

Estas análises poderão ser usadas em uma etapa de pós-treinamento para melhorar os resultados obtidos.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,63	0,66	0,66	0,65	0,64	0,66	0,66	0,66
2	0,64	0,66	0,66	0,66	0,64	0,67	0,67	0,67
3	0,65	0,67	0,67	0,67	0,65	0,67	0,68	0,67
4	0,64	0,67	0,67	0,67	0,65	0,68	0,68	0,67
5	0,65	0,63	0,66	0,65	0,65	0,67	0,67	0,67
6	0,63	0,66	0,66	0,65	0,64	0,67	0,67	0,66
7	0,63	0,66	0,66	0,65	0,65	0,67	0,67	0,67
8	0,62	0,65	0,65	0,64	0,65	0,67	0,67	0,67
9	0,57	0,62	0,61	0,60	0,65	0,67	0,68	0,66
10	0,61	0,54	0,58	0,54	0,65	0,67	0,67	0,67

(a) Combinação de redes unidirecionais.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,67	0,69	0,69	0,69	0,68	0,70	0,70	0,69
2	0,68	0,70	0,70	0,70	0,71	0,72	0,72	0,72
3	0,69	0,70	0,70	0,70	0,71	0,72	0,72	0,72
4	0,69	0,71	0,71	0,71	0,71	0,73	0,73	0,73
5	0,69	0,71	0,71	0,71	0,71	0,73	0,73	0,72
6	0,70	0,71	0,71	0,71	0,71	0,72	0,72	0,72
7	0,69	0,71	0,71	0,71	0,71	0,72	0,72	0,72
8	0,70	0,71	0,71	0,71	0,71	0,72	0,72	0,72
9	0,70	0,71	0,71	0,71	0,71	0,72	0,72	0,72
10	0,70	0,71	0,71	0,71	0,71	0,72	0,72	0,72

(b) Combinação de redes bidirecionais.

Tabela 4.8: Resultados obtidos na classificação Q3 na combinação de classificadores.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,26	0,50	0,53	0,46	0,26	0,48	0,53	0,46
2	0,28	0,53	0,53	0,47	0,27	0,50	0,53	0,47
3	0,28	0,53	0,53	0,47	0,28	0,53	0,54	0,48
4	0,29	0,53	0,54	0,48	0,28	0,52	0,54	0,48
5	0,28	0,52	0,53	0,47	0,27	0,50	0,54	0,48
6	0,26	0,47	0,52	0,46	0,27	0,50	0,54	0,48
7	0,17	0,34	0,39	0,27	0,27	0,47	0,54	0,48
8	0,15	0,33	0,36	0,23	0,27	0,46	0,53	0,47
9	0,14	0,30	0,35	0,21	0,27	0,46	0,53	0,47
10	0,13	0,28	0,33	0,18	0,27	0,49	0,53	0,47

(a) Combinação de redes unidirecionais.

	LSTM				GRU			
Camadas	**Acu**	**Pre**	**Rev**	**F1**	**Acu**	**Pre**	**Rev**	**F1**
1	0,30	0,52	0,56	0,51	0,30	0,52	0,56	0,52
2	0,35	0,55	0,57	0,54	0,35	0,56	0,59	0,56
3	0,37	0,55	0,58	0,55	0,37	0,57	0,60	0,57
4	0,37	0,57	0,58	0,55	0,38	0,57	0,60	0,57
5	0,38	0,56	0,58	0,55	0,38	0,57	0,60	0,57
6	0,37	0,56	0,58	0,55	0,38	0,57	0,60	0,57
7	0,37	0,56	0,58	0,55	0,38	0,57	0,60	0,57
8	0,38	0,57	0,58	0,55	0,38	0,56	0,60	0,57
9	0,38	0,56	0,58	0,55	0,38	0,56	0,60	0,57
10	0,38	0,56	0,58	0,55	0,38	0,56	0,60	0,57

(b) Combinação de redes bidirecionais.

Tabela 4.9: Resultados obtidos na classificação Q8 na combinação de classificadores.

	Q3			Q8							
	C	**E**	**H**	**B**	**E**	**G**	**H**	**I**	**L**	**S**	**T**
Média	4,69	4,42	9,47	1,02	5,32	3,37	11,12	5,38	1,92	1,56	2,09
Mediana	4	4	8	1	5	3	10	5	1	1	2
Valor Mínimo	1	1	1	1	1	3	1	5	1	1	1
Valor Máximo	185	92	164	4	92	17	164	15	79	16	20
Desvio Padrão	3,71	6,01	6,48	0,14	2,73	0,87	6,21	0,97	1,41	0,83	0,83

Tabela 4.10: Dados estatísticos das sequências de estruturas secundárias do PDB.

	Início (%)	**Meio (%)**	**Final (%)**
Q3			
C	42,56	40,88	41,19
E	32,24	21,55	39,01
H	25,20	37,57	19,80
Q8			
B	1,13	1,27	1,08
E	24,06	21,61	18,72
G	3,65	4,00	4,12
H	28,57	32,05	34,89
I	0,38	0,65	0,57
L	22,61	19,86	21,41
S	8,61	9,38	8,38
T	10,99	11,18	10,83

Tabela 4.11: Distribuição das estruturas secundárias nas proteínas do PDB.

Capítulo 5

Plano de Trabalho e Cronograma de Execução

O plano de trabalho é composto pelas seguintes atividades:

1. Estudo e análise das principais técnicas e abordagens disponíveis na literatura.
2. Preparação da base de dados.
3. Construção da rede para predizer estruturas secundárias das proteínas.
4. Realização de testes de estruturas secundárias das proteínas.
5. Construção da rede para predizer funções das proteínas.
6. Realização de testes de funções das proteínas.
7. Comparação dos resultados com outros trabalhos disponíveis na literatura.
8. Documentação e publicação dos resultados.
9. Escrita do documento da dissertação.
10. Defesa da dissertação.

O cronograma de execução das atividades propostas, divididas em 4 etapas, em um prazo de 24 meses, é apresentado na Tabela 5.1.

Atividades	**1º ano**						**2º ano**					
	1	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
Etapa 1 - Preparação												
Pesquisa bibliográfica	•	•	•	•	•	•	•	•	•			
Preparação da base de dados		•	•	•	•	•						
Etapa 2 - Estruturas Secundárias												
Construção da rede				•	•	•						
Realização de testes					•	•	•					
Comparação dos resultados com outros trabalhos						•	•					
Etapa 3 - Funções												
Construção da rede							•	•	•			
Realização de testes								•	•	•		
Comparação dos resultados com outros trabalhos									•	•		
Etapa 4 - Conclusão												
Publicação dos resultados								•			•	
Escrita da dissertação							•			•	•	•
Defesa da dissertação												•

Tabela 5.1: Cronograma de atividades dividido em bimestres.

Bibliografia

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[2] A. Barrett. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *European Journal of Biochemistry*, 250(1):1–6, 1997.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[4] H. M. Berman. The protein data bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):88–95, 2008.

[5] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley. The protein data bank at 40: reflecting on the past to prepare for the future. *Structure*, 20(3):391–396, 2012.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[7] S. Bhattacharya, C. Bhattacharyya, and N. Chandra. Structural alignment based kernels for protein structure classification. In *24th International Conference on Machine Learning (ICML)*, pages 73–80. ACM, 2007.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[9] Y. Cho and A. Zhang. Predicting protein function by frequent functional association pattern mining in protein interaction networks. *IEEE Transactions on Information Technology in Biomedicine*, 14(1):30–36, 2009.

[10] C. Fang, Y. Shang, and D. Xu. MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 86(5):592–598, 2018.

[11] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur. Protein interface prediction using graph convolutional networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6530–6539. Curran Associates, Inc., 2017.

[12] Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2016.

[13] V. Golkov, M. J. Skwark, A. Golkov, A. Dosovitskiy, T. Brox, J. Meiler, and D. Cremers. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4222–4230. Curran Associates, Inc., 2016.

[14] H. Hasic, E. Buza, and A. Akagic. A hybrid method for prediction of protein secondary structure based on multiple artificial neural networks. In *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1195–1200. IEEE, 2017.

[15] L. T. Hattori, C. M. V. Benitez, and H. S. Lopes. A deep bidirectional long short-term memory approach applied to the protein secondary structure prediction problem. In *4th IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE, 2017.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[17] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, 25(1):231–234, 1997.

[18] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[19] H. Kamisetty and C. J. Langmead. A bayesian approach to protein model quality assessment. In *26th Annual International Conference on Machine Learning (ICML)*, pages 481–488. ACM, 2009.

[20] M. Kulmanov, M. A. Khan, and R. Hoehndorf. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 2017.

[21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Z. Li and Y. Yu. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. In *25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2560–2567. AAAI Press, 2016.

[24] Z. Lin, J. Lanchantin, and Y. Qi. MUST-CNN: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. In *30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 27–34, 2016.

[25] Z. C. Lipton, J. Berkowitz, and C. Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, pages 1–38, 2015.

[26] Y. Liu, J. Cheng, Y. Ma, and Y. Chen. Protein secondary structure prediction based on two dimensional deep convolutional neural networks. In *3rd IEEE International Conference on Computer and Communications (ICCC)*, pages 1995–1999. IEEE, 2017.

[27] A. E. Márquez-Chamorro, G. Asencio-Cortés, C. E. Santiesteban-Toca, and J. S. Aguilar-Ruiz. Soft computing methods for the prediction of protein tertiary structures: A survey. *Applied Soft Computing*, 35:398–410, 2015.

[28] A. Mitchell, H. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas, and R. D. Finn. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, 43(D1):D213–D221, 2014.

[29] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano. Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function, and Bioinformatics*, 82:1–6, 2014.

[30] N. Nadzirin and M. Firdaus-Raih. Proteins of unknown function in the protein data bank (PDB): An inventory of true uncharacterized proteins and computational tools for their analysis. *International Journal of Molecular Sciences*, 13(10):12761–12772, 2012.

[31] E. Orhan. A simple cache model for image recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 10107–10116. Curran Associates, Inc., 2018.

[32] Y. Qi, M. Oja, J. Weston, and W. S. Noble. A unified multitask architecture for predicting local protein properties. *PLOS ONE*, 7(3):e32235, 2012.

[33] P. Radivojac, W. T. Clark, T. R. Oron, T. Schnoes, Alexandra Mand Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. A. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. E. Sternberg, N. Škunca, F. Supek, M. Bošnjak, P. Panov,

S. Džeroski, T. Šmuc, Y. A. I. Kourmpetis, A. D. J. van Dijk, C. J. F. t. Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, and I. Friedberg. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221, 2013.

[34] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Almost unsupervised text to speech and automatic speech recognition. In *36th International Conference on Machine Learning (ICML)*, pages 5410–5419, 2019.

[35] C. N. Rokde and M. Kshirsagar. Bioinformatics: Protein structure prediction. In *4th International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2013.

[36] P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman, and S. K. Burley. The RCSB protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic Acids Research*, 45(1):271–281.

[37] A. Roy, A. Kucukural, and Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010.

[38] R. D. Sleator and P. Walsh. An overview of in silico protein function prediction. *Archives of Microbiology*, 192(3):151–155, 2010.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[40] V. Sze, Y. Chen, T. Yang, and J. S. Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

[41] The UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2018.

[42] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, and G. Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1):D364–D368, 2014.

[43] Y. Wang, H. Leung, M. Gavrilova, O. Zatarain, D. Graves, J. Lu, N. Howard, S. Kwong, P. Sheu, and S. Patel. A survey and formal analyses on sequence learning methodologies and deep neural networks. In *17th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pages 6–15. IEEE, 2018.

[44] Y. Wang, H. Mao, and Z. Yi. Protein secondary structure prediction by using deep learning method. *Knowledge-Based Systems*, 118:115–123, 2017.

[45] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, and H. M. Berman. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, 21(7):988–992, 2004.

[46] J. M. Yunes and P. C. Babbitt. Effusion: prediction of protein function from sequence similarity networks. *Bioinformatics*, 35(3):442–451, 2018.

[47] F. Zhang, H. Song, M. Zeng, Y. Li, L. Kurgan, and M. Li. Deepfunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics*, pages 1–7, 2019.

[48] J. Zhou and O. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In *31st International Conference on Machine Learning (ICML)*, pages 745–753, 2014.