UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

PREDIÇÃO DE ESTRUTURA SECUNDÁRIA DE PROTEÍNAS ATRAVÉS DE MÉTODOS DE APRENDIZADO SUPERVISIONADO

JAMERSON FELIPE PEREIRA LIMA

RECIFE

JANEIRO/2015

PREDIÇÃO DE ESTRUTURA SECUNDÁRIA DE PROTEÍNAS ATRAVÉS DE MÉTODOS DE APRENDIZADO SUPERVISIONADO

Projeto apresentado como requisito para a nota da disciplina: Trabalho de conclusão de curso, ministrado pela Prof.ª Francielle Silva Santos no curso de Ciência da Computação na Universidade Federal Rural de Pernambuco - UFRPE

Orientadora: Prof.ª Jeane Cecília Bezerra de Melo

JAMERSON FELIPE PEREIRA LIMA

PREDIÇÃO DE ESTRUTURA SECUNDÁRIA DE PROTEÍNAS ATRAVÉS DE MÉTODOS DE APRENDIZADO SUPERVISIONADO

Projeto apresentado como requisito para a nota da disciplina: Trabalho de conclusão de curso, ministrado pela Prof.ª Francielle Silva Santos no curso de Ciência da Computação na Universidade Federal Rural de Pernambuco - UFRPE

29 de janeiro de 2015.

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por **Jamerson Felipe Pereira Lima** como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **Predição de Estrutura Secundária de Proteínas através de Métodos de Aprendizado Supervisionado** , orientado pelo Prof. Jeane Cecília Bezerra de Melo e aprovado pela seguinte banca examinadora:

Jeane Cecília Bezerra de Melo
DEINFO – UFRPE

Rodrigo de Souza
DEINFO-UFRPE

Obionor Nóbrega
DEINFO-UFRPE

Ás minhas mães,
a quem, sobretudo, dedico a minha vida.

AGRADECIMENTOS

Ao meu país, o Brasil, que tanto amo, e que garante a minha oportunidade de acesso à Educação. À Universidade Federal Rural de Pernambuco, minha segunda casa durante cinco anos.

Aos meus professores, mestres da minha formação.

À minha orientadora, Jeane, pela sensibilidade, paciência, inteligência e pela amizade, e que me acompanha desde o princípio deste percurso que este trabalho conclui.

Aos meus irmãos: Jonathan, Larissa, Letícia Marcelly, Letícia Helena, Geovanna, Renan, Guilherme, Raquel e (que ainda está por vir) Cecília, pelo convívio familiar e por fazer os meus dias melhores com a certeza do companheirismo. Também ao meu tio Jackson, convivido quase como irmão, e por quem nutro um imenso respeito.

Aos meus amigos, que estão e estiveram sempre comigo. A Thaís, em especial, pela injeção de ânimo na finalização deste trabalho. Também a ela e a Dayanne pela companhia que fizeram-me durante o Projeto de Conclusão. A Amaro, pela ajuda nas correções. A Alex, Cássia, Joana, amigos do coração. A Allyson e Renan, dois grandes amigos que ganhei neste percurso. Aos meus amigos portugueses Sofia e Gonçalo, que, por um feliz acaso, me acompanharam em meu início com a Biologia no IST. Aos amigos Inês, pela paciência e apoio que deu-me na conclusão das disciplinas; Thiago, parceiro de um dos melhores momentos da minha vida, o intercâmbio. Aos demais amigos e colegas, não menos importantes.

Ao meu namorado, Márcio, pelo amor, o apoio em todas as horas e as dicas de inglês.

Ao meu pai, Ribamar, pelo amor e simplicidade.

Por fim, e mais importante, às minhas mães: à avó, Iraci, pelo amor que sempre me confiou, mesmo não o sendo em sangue; à mãe-avó, Fátima, pelo amor que sempre é de sobra e pela companhia no dia-a-dia; e à minha mãe, Jaqueline, por sempre acreditar em mim e, acima de tudo, ensinar-me a coragem de mudar a minha própria realidade. Às três, pelo reconhecimento da importância da Educação e pela formação moral que tive. Estejam certas de que este é somente o meu primeiro grande passo.

RESUMO

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. Devido aos altos custos e tempo necessários a estas técnicas, métodos alternativos têm sido buscados. Portanto, o uso de métodos computacionais tornou-se uma alternativa para a obtenção da estrutura tridimensional da proteína a partir de sua estrutura primária. Esse processo, denominado Problema de Dobramento das Proteínas, é um problema NP-difícil, isto é, não possui nenhuma solução eficiente conhecida. Assim, abordagens computacionais que têm como objetivo obter estruturas intermediárias, como a estrutura secundária da proteína, que auxiliam na determinação da estrutura terciária, têm sido alvo de estudos. 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. 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 e a relação entre suas estruturas e seus resultados estatísticos, principalmente quanto ao modelo de aprendizado utilizado. 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 como múltiplas camadas (MLP, do inglês *multilayer perceptron*). Os resultados deste método foram comparados com os de um preditor baseado em máquina de vetores de suporte (do inglês *support vector machine*). A acurácia atingida foi 70,55% utilizando-se o conjunto de dados CB513 e *3-fold cross-validation*.

Palavras-chave: redes neurais artificiais, estrutura secundária da proteína, predição de estrutura da proteína.

ABSTRACT

Protein is one of the main objects of study in biology, being a key element in the implementation of several biological processes such as cell structural support, biochemical catalysis, cellular transport, immunization, among other functions. The proteins have different levels of structure, which may vary from amino acid chain, called primary structure, to a three dimensional conformation, resulting from the interaction of the amino acids and polypeptide chains. The secondary structure, an intermediate level, is defined by the occurrence of stable structures in the amino acid chain. The three-dimensional conformation, called tertiary structure, in turn, reflects directly the function of the protein, and X-Ray Crystallography and the Nuclear Magnetic Resonance Spectroscopy are the most used experimental methods to obtain them. Due to the high costs and time required in these techniques, alternative methods have been the subject of studies. Therefore, the use of computational methods has become an alternative to achieve the three-dimensional structure of the protein from its primary structure. This process, called Protein Folding Problem, was proved to be a NP-hard problem, that is, there is no known efficient solution. Thus, computational approaches that aim to obtain intermediate structures such as the secondary structure of the protein, which help in determining the tertiary structure, have been the subject of studies. A recurrent approach to obtain the secondary structure is the use of machine learning, which can predict the secondary structure corresponding to a particular sequence of amino acid based on the observation of a set of examples given as an input, with a given level of accuracy. A literature review of the latest methods of secondary structure prediction was carried out to analyze the state of the art, evaluating the progress achieved by current methods and the relation between their results and structure, with a special focus on the learning method. It was also analyzed the relation between some aspects and the amount and formatting of the input data. Additionally, a secondary structure predictor based on a multilayer perceptron (MLP) was proposed. The results of the developed method were compared with a predictor based on support vector machines. The resulting accuracy with the method proposed was 70.55% using CB513 data set and 3-fold cross-validation.

Keywords: artificial neural networks, protein secondary structure, protein structure prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

ANN – *Artificial Neural Network*, termo em inglês para Rede Neural Artificiai

CPM – *Compound Pyramid Model*, termo em inglês para Modelo de Pirâmide Composta

DNA – *Deoxirybonucleic Acid*, termo em inglês para Ácido Desoxirribonucleico.

ELM – *Extreme learning machine*, termo em inglês para Máquina de Aprendizado Extremo

HMM – *Hidden Markov model*, termo em inglês para Modelo oculto de Markov

KDD – *Knowledge Discovery in Databases*, termo em inglês para Descoberta de Conhecimento em Bases de Dados

MLP – *Multilayer Perceptron*, term em inglês para Redes Neurais Artificiais Direcionais de Camadas Múltiplas

MMBP – *Multimodal Backpropagation*, termo em inglês para Retropropagação Multimodal

NCBI – *National Center for Biotechnology Information*, termo em inglês para Centro Nacional de Informação Biológica

PSSM – *Position-Specific Scoring Matrix*, termo em inglês para Matriz de Pontuação de Posições Específicas

RBF – *Radial Basis Function*, Função de Base Radial

RNA[1] – Rede Neural Artificial

RNA – *Rybonucleic Acid*, termo em inglês para Ácido Ribonucleico.

RNAD – Rede Neural Artificial Direcional

RNR – Rede Neural Recorrente

RPROP – *Resilient Propagation*, termo em inglês para Propagação Resiliente

SVM – *Support Vector Machine*, termo em inglês para Máquinas de Vetores de Suporte

---

1 Ao longo do texto, será feito uso da sigla em inglês em detrimento ao homônimo em português devido à presença de outro termo com sigla igual.

SUMÁRIO

1. INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

O interesse pelo estudo de fatos biológicos, com o auxílio da computação, impulsionou o desenvolvimento de uma área de pesquisa: a Biologia Computacional. Seu objetivo central é utilizar-se dos algoritmos e métodos desenvolvidos na computação e aplicá-los ao contexto de dados biológicos. Neste trabalho, o objetivo é analisar e inferir sobre um ramo específico da Biologia Computacional, a predição de estrutura secundária de proteínas.

1.2 JUSTIFICATIVA

As proteínas compõem em torno de 20% do corpo humano e são essenciais na realização de funções vitais, tais como catálise bioquímica, suporte estrutural celular e imunização (NELSON; COX, 2008). Seu estudo é de fundamental importância, uma vez que os conhecimentos obtidos sobre a estrutura proteica podem ser utilizados no desenvolvimento de métodos de diagnóstico de patologias, como AIDS e câncer; fármacos, além de possuir outras aplicações nas áreas agrícola e de biodiversidade (IMMING; SINNING; MEYER, 2007); (ISMAEEL; ABLAHAD, 2013).

Para o estudo das funções biológica e patológica de uma proteína, é necessário o conhecimento sobre sua conformação tridimensional, denominada estrutura terciária, uma vez que função e estrutura proteica estão diretamente relacionados (NELSON; COX, 2008). A obtenção desta estrutura pode ser feita através de diversos métodos (COMPIANI; CAPRIOTTI, 2013). Os mais amplamente utilizados são a Cristalografia de Raios-X e Espectroscopia de Ressonância Magnética Nuclear, os quais, em geral, demonstram-se custosos e demorados, ou podem mesmo ser inaplicáveis ou inviáveis (ESWAR et al., 2008).

Sabe-se que toda a informação necessária à obtenção da estrutura terciária da proteína está contida em sua estrutura primária, representada pela sequência de aminoácidos que a compõem. A obtenção da estrutura primária é um processo de menor complexidade, de tal maneira que o número de sequências de aminoácidos depositadas no UniProtKB/Swiss-Prot em 2014 é de 546000 sequências, à medida que a quantidade de estruturas tridimensionais proteicas depositadas no *Research Collaboratory for Structural Bioinformatics Protein Data Bank* (RCSB PDB) (BERMAN et al., 2000) no mesmo período é

de 102364 sequências (RCSB, 2014); (UNIPROT, 2014). Desse modo, métodos preditivos, que facilitem a obtenção da estrutura tridimensional da proteína a partir de sua estrutura primária, têm sido alvo de pesquisas.

O grande volume de dados originados obtidos no estudo das proteínas incentivou o desenvolvimento de técnicas computacionais para tratá-los e analisá-los. A inferência da estrutura tridimensional da proteína através de métodos computacionais, conhecida como "Problema de Dobramento das Proteínas", foi demonstrada ser NP-difícil (HART; ISTRAIL, 1997). Devido a essa complexidade, normalmente essa classificação inclui passos intermediários. Um exemplo é a localização de elementos que ocorrem frequentemente na estrutura 3D, como hélices-α, folhas-β e *coils*, problema denominado *Predição de Estrutura Secundária de Proteínas* (SETUBAL; MEIDANIS, 1997), funcionando de maneira complementar às técnicas convencionais para a determinação da estrutura terciária da proteína.

Várias soluções computacionais já foram desenvolvidas para este problema, tais como as que utilizam Modelos Ocultos de Markov (*Hidden Markov Model* - HMM), *Support Vector Machines* (SVM), redes neurais e árvores de decisão (BETTELLA; RASINSKI; KNAPP, 2012). As técnicas de aprendizado de máquina foram as mais bem-sucedidas, uma vez que abordam o problema sem considerar as interações termodinâmicas que dão origem à estrutura terciária da proteína a partir da sequência de aminoácidos (COMPIANI; CAPRIOTTI, 2013). As estratégias que utilizam redes neurais, em particular, têm demonstrado bons resultados, com um nível de acurácia situado em volta dos 76% a 80%. Entretanto, em alguns casos o desvio padrão é de até 10%, à medida que resultados satisfatórios deveriam possuir desvio padrão menor que 5% (MONTGOMERY PETTITT, 2013). Um desvio padrão alto pode afetar a confiabilidade do preditor, uma vez que os resultados podem apresentar uma variação muito alta, acarretando a possibilidade de acurácia baixa em algumas situações.

O presenta trabalho apresenta uma atualização dos estudos descritos em (MELO, 2005), trazendo abordagens mais recentes para o problema da Predição da Estrutura Secundária de Proteínas. Adicionalmente, dada a importância do conhecimento preciso da estrutura da proteína, de maneira a não afetar estudos decorrentes, é fundamental que a predição tenha um bom resultado estatístico. Com base nesta necessidade, o preditor descrito em (MELO, 2005) foi modificado com o objetivo de desenvolver um novo preditor de

estrutura secundária que apresente uma performance ao menos comparável aos existentes quanto à acurácia e precisão, concomitantemente com uma redução do desvio padrão.

1.3 ESCOPO DO TRABALHO

Como alternativa aos métodos experimentais de obtenção da estrutura das proteínas, os modelos de aprendizado supervisionado têm sido usados para facilitar a solução do problema. Desde o trabalho pioneiro neste campo, que propôs o uso de redes neurais, realizado por Qian e Sejnowski (QIAN; SEJNOWSKI, 1988), o avanço da taxa de acerto, nos métodos mais recentes, chega a atingir valores próximos ao limite teórico de predição, que é de 88% (ROST, 2001).

A taxa de acerto em geral não é a mesma para cada classe de estrutura secundária. A classe mais afetada é a de folhas-β, uma vez que estão submetidas a interações de longa distância e estão em menor proporção em relação a hélices-α e *coils*. Alguns métodos tiveram algum sucesso com esse problema, entretanto, em alguns casos o desvio padrão é muito alto, o que reduz a confiabilidade da predição, uma vez que a acurácia pode variar muito (Melo, 2005).

Alguns métodos têm aplicado outras técnicas além das matrizes de alinhamento (PSSM) propostas por Jones (JONES, 1999), as quais representaram o maior avanço na taxa de acerto desde o trabalho de Qian e Sejnowski (QIAN; SEJNOWSKI, 1988). Algumas abordagens são o uso de funções matemáticas de pontuação que analisam a probabilidade de um aminoácido de pertencer a determinada classe, informações sobre as características termodinâmicas dos aminoácidos, entre outros.

Portanto, o escopo do trabalho compreende a atualização do estado da arte apresentado em (MELO, 2005), enfatizando os métodos de Inteligência Artificial para abordagem do Problema de Predição de Estrutura Secundária. Além disso, foi proposto também um preditor, baseado no modelo implementado por Melo, buscando melhorar o resultado estatístico do modelo de aprendizado supervisionado utilizando somente perfis de alinhamento de conjuntos de proteínas maiores, isto é, fornecendo mais exemplos ao modelo (AVDAGIC et al., 2009).

1.4 FORMULAÇÃO DE HIPÓTESES

É possível melhorar a performance estatística de um método de predição de estrutura secundária de proteínas baseado em redes neurais alterando a configuração da rede e utilizando-se somente de perfis de alinhamento de famílias de proteínas (PSSM).

1.5 ELABORAÇÃO DOS OBJETIVOS

1.5.1 Geral

Utilizar recursos de Inteligência Artificial para obter a estrutura secundária de uma proteína a partir de sua estrutura primária.

1.5.2 Específicos

- Investigar as técnicas de Inteligência Artificial para predição de estrutura secundária de proteínas
- Desenvolver um preditor de estrutura secundária de proteínas utilizando técnicas de Inteligência Artificial
- Analisar as saídas dadas pelo preditor e tratá-las buscando maximizar o desempenho estatístico, observando a acurácia e a precisão, bem como visando reduzir o desvio padrão.
- Comparar o desempenho do preditor desenvolvido com resultados de outros similares.

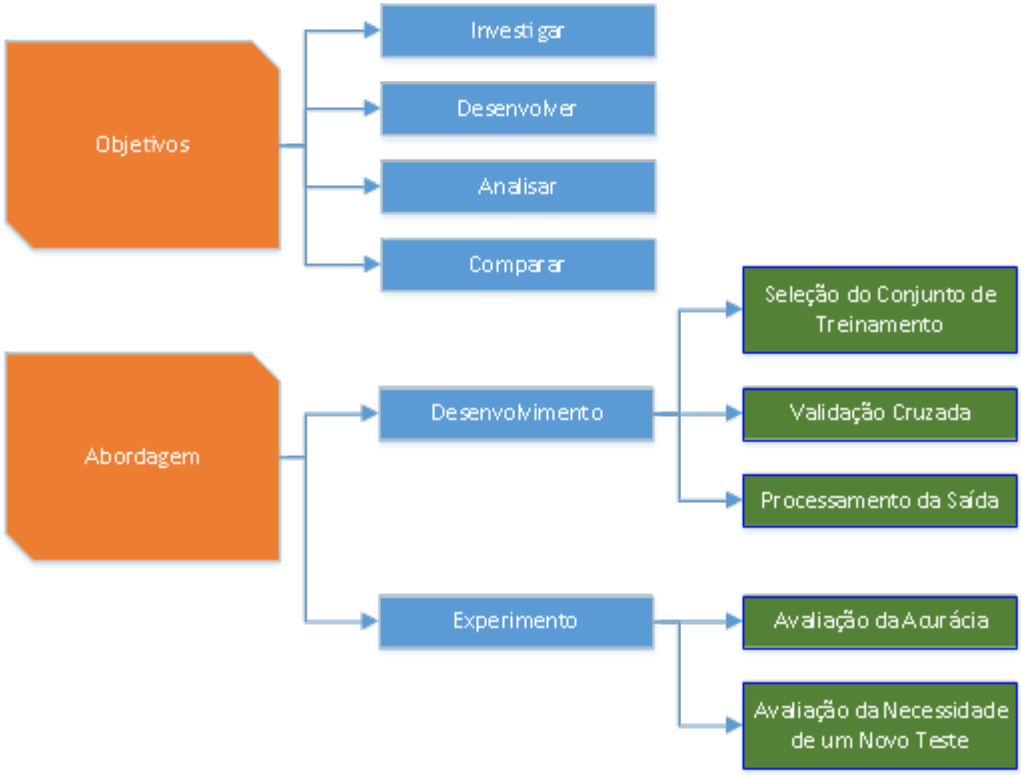
1.6 DEFINIÇÃO DA METODOLOGIA

Inicialmente, o objetivo foi determinar as técnicas de Inteligência Artificial a serem aplicadas no preditor a ser desenvolvido, bem como definir a arquitetura da rede. Ambas as etapas baseiam-se essencialmente no levantamento bibliográfico realizado, permitindo assim, definir escolhas em função dos melhores resultados já alcançados. A avaliação do desempenho foi feita através de sua implementação. O banco de dados a ser utilizado é o CB513 (CUFF; BARTON, 2000), sendo o treinamento realizado com validação cruzada, mais especificamente 3-*fold cross-validation* (KOHAVI, 1995). O processamento das saídas do preditor é realizado através de dois métodos, a somatório e máximo.

A partir do resultado da simulação feita com base de dados CB513, foi observado que, tomando como princípio o fato de que um conjunto de dados maior agrega mais

informação, o uso de conjuntos menores foi descartado. Além disso, o uso de bases de dados maiores tornaria o teste impraticável, dado o tempo de treinamento e o poder computacional necessários para sua realização. Um dos fatores que motivou a escolha da base CB513 foi para permitir a comparação que é feita com o preditor desenvolvido por Huang e Chen (HUANG; CHEN, 2013), a qual é objeto de discussão nos capítulos 4 e 5. A Figura 1 representa uma esquematização da metodologia do trabalho.

Figura 1 – Metodologia do trabalho.



1.7 ESTRUTURA DO TRABALHO

O capítulo 2 abordará os conceitos utilizados como base ao longo do trabalho, em que são listados conceitos que dizem respeito aos temas biológicos que concernem ao trabalho, como as proteínas e sua estrutura, assim como conceitos computacionais, de redes neurais, seu treinamento e validação. O capítulo 3 descreve alguns trabalhos que, assim como este, versam sobre a predição de estrutura secundária de proteínas. São comentados a metodologia, os dados e os resultados de cada método, com as respectivas vantagens e desvantagens em que podem incorrer. O capítulo 4 descreve o desenvolvimento do trabalho, o qual é resultante da metodologia proposta. Estão aí contidos como foram feitos o projeto e treinamento do modelo de aprendizado, a obtenção dos dados e a validação do preditor. No

capítulo 5 são discutidos alguns aspectos relativos aos resultados, com as devidas justificativas aos fatos ocorridos no desenvolvimento. O capítulo 6 contém as conclusões obtidas com os resultados do trabalho, bem como algumas sugestões a trabalhos futuros.

2. CONCEITOS BÁSICOS E TERMINOLOGIA

2.1 CONSIDERAÇÕES INICIAIS

A Biologia Computacional, como uma área de fronteira entre a Biologia e a Computação, envolve definições específicas de cada área. Este capítulo tem como motivação introduzir os conceitos básicos de Biologia e Computação necessários à compreensão dos tópicos abordados. A seção 2.2 discorre sobre a fundamentação biológica, na qual se inclui a ideia de estrutura proteica, envolvendo seus diversos níveis. A seção 2.3 enumera os conceitos computacionais de Inteligência Artificial utilizados, tais como o de aprendizado e redes neurais. Outros conceitos são apresentados oportunamente ao longo do trabalho.

2.2 FUNDAMENTOS BIOLÓGICOS

A vida pode ser definida como o ciclo contínuo de trocas que ocorrem entre um ente dito **ser vivo** e a natureza. Cada ser vivo, por sua vez, pode ter um grau de complexidade que varia de uma célula procariótica, a qual não possui um núcleo definido, até seres de complexidade mais elevada, tais como os mamíferos, classe que engloba os seres humanos. Um elemento comum a todos os seres vivos é o fato de que seus processos biológicos, em suma, são determinados e executados pelas **proteínas**, as quais têm sua codificação determinada a partir do código genético de cada ser vivo. Dessa forma, o conjunto das proteínas de um ser vivo pode determinar a maneira como ele lida com o meio ambiente.

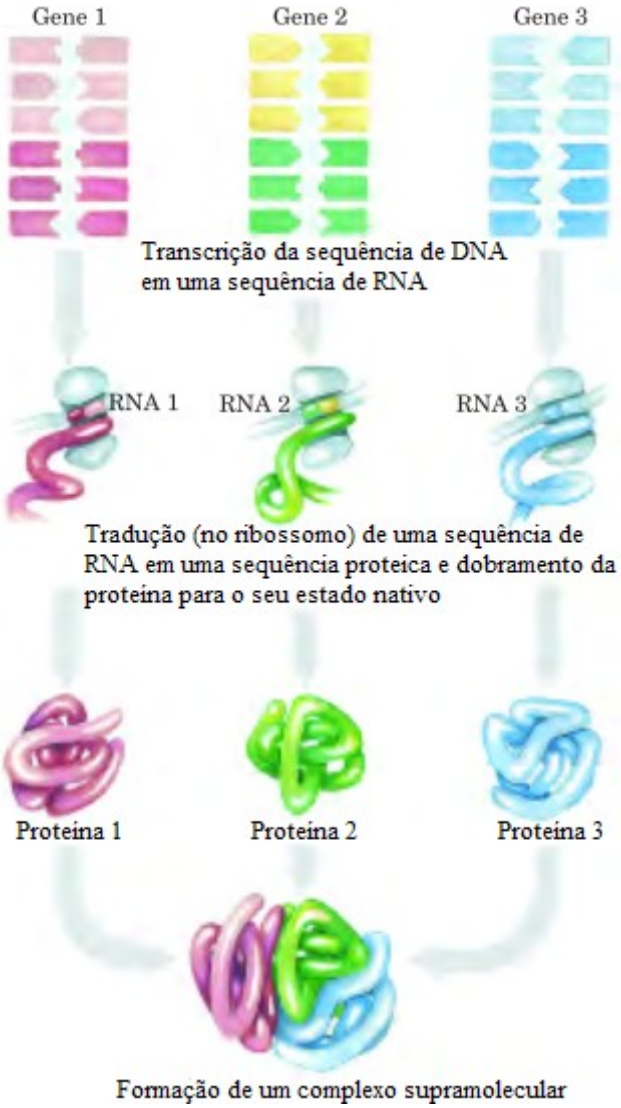
No corpo humano, especificamente, as proteínas perdem em proporção apenas para a água, que compõe cerca de 70%, à medida que as proteínas compõem cerca de 20%. As proteínas efetuam diversos papéis, entre os quais suporte estrutural celular, defesa do organismo, transporte celular, bem como catálise de reações químicas, entre outras variadas funções.

Uma proteína é composta por várias unidades menores, denominadas monoméricas: os aminoácidos. Através de ligações covalentes, chamadas **peptídicas**, os aminoácidos podem formar cadeias poliméricas, que dão origem às proteínas. Uma vez que na ligação entre dois aminoácidos ocorre a perda de uma molécula de água, o que encontra-se na cadeia polipeptídica é chamado **resíduo** do aminoácido original. Convencionalmente, determina-se que uma proteína é formada por pelo menos 20 aminoácidos e cadeias menores são

denominadas simplesmente cadeias polipeptídicas (BRANDEN; TOOZE, 1991). Tipicamente as proteínas podem conter entre 100 e 5000 aminoácidos.

As proteínas são resultantes da tradução do RNA (ácido ribonucleico), o qual é obtido a partir da transcrição da informação genética contida no DNA (ácido desoxirribonucleico). Além disso, pode-se formar complexos supramoleculares, como cromossomos, ribossomos ou membranas, a partir da interação de proteínas com outras proteínas, lipídios ou ácidos nucleicos. A Figura 2 apresenta um esquema da produção das proteínas. Uma determinada sequência linear de aminoácidos produz uma estrutura tridimensional única e não ambígua.

Figura 2 – Esquema de produção das proteínas.



Adaptado de (NELSON; COX, 2008).

Os aminoácidos que ocorrem com mais frequência nos seres vivos são 20, denominados aminoácidos padrão, no entanto, existem outros aminoácidos que aparecem

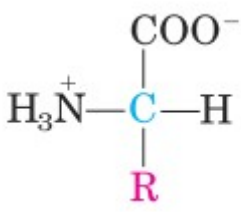
apenas em alguns tipos de proteínas. Os 20 aminoácidos padrão mencionados são listados no Quadro 1:

Quadro 1 – Aminoácidos que ocorrem naturalmente nas proteínas.

Código de uma letra	Código de três letras	Nome
A	ALA	Alanina
R	ARG	Arginina
D	ASP	Ácido Aspártico
N	ASN	Asparagina
C	CYS	Cisteína
E	GLU	Ácido glutâmico
Q	GLN	Glutamina
G	GLY	Glicina
H	HIS	Histidina
I	ILE	Isoleucina
L	LEU	Leucina
K	LYS	Lisina
M	MET	Metionina
F	PHE	Fenilalanina
P	PRO	Prolina
S	SER	Serina
T	THR	Treonina
W	TRP	Triptofan
Y	TYR	Tirosina
V	VAL	Valina

Um aminoácido é formado por um carbono central, o carbono alfa (C_α), como mostrado na Figura 3. A este átomo estão ligados um átomo de hidrogênio, um grupo amino (NH_2), um grupo carboxila (COOH) e uma cadeia lateral. A diferença entre os aminoácidos é feita pela cadeia lateral.

Figura 3 – Esquema geral de um aminoácido.



Extraído de (NELSON; COX, 2008).

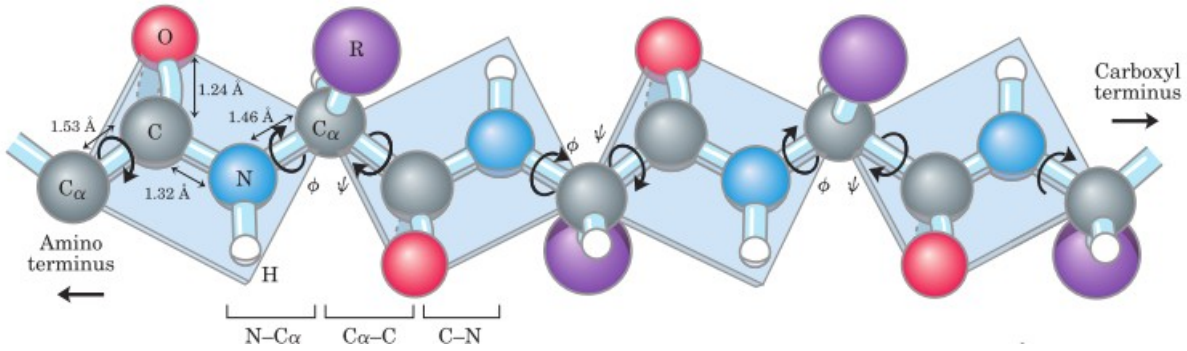
2.2.1 Estrutura da proteína

As proteínas, embora sejam formadas por uma cadeia de aminoácidos, não possuem uma estrutura linear, mas sim tridimensional, devido às rotações que podem ocorrer na cadeia de aminoácidos, as combinações entre esses ângulos podem gerar uma quantidade ilimitada de formas nas proteínas. A estrutura da proteína pode ser compreendida em diversos níveis: primária, secundária, terciária e quaternária.

2.2.1.1 Estrutura primária

A cadeia de aminoácidos forma uma estrutura denominada **esqueleto** *(*do inglês *backbone)* caracterizada pela repetição de um átomo de nitrogênio (N), o carbono central (C_α) e um grupo CO. A estrutura dessa cadeia é determinada pelos ângulos formados entre N e C_α; e C_α e C, nomeadamente φ e ψ, respectivamente, mostrados na Figura 4.

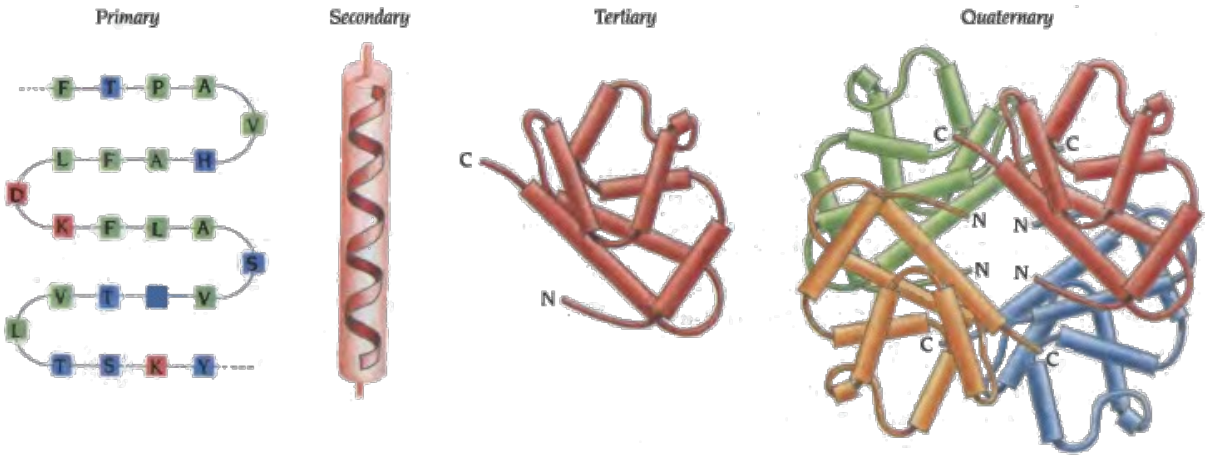
Figura 4 – Esquema das conexões entre os aminoácidos e ângulos φ e ψ.



São mostrados o C_α e as ligações N–C_α; e C_α–C. Extraído de (NELSON; COX, 2008).

Devido a esse tipo de interação, a estrutura nativa da proteína não é constituída por uma sequência linear de aminoácidos – denominada estrutura primária da proteína – mas em um formato tridimensional, apresentando as estruturas denominadas terciária e quaternária, além de conter alguns arranjos recorrentes estáveis, a estrutura secundária. A Figura 5 apresenta os níveis de estrutura citados.

Figura 5 – Estrutura das proteínas.

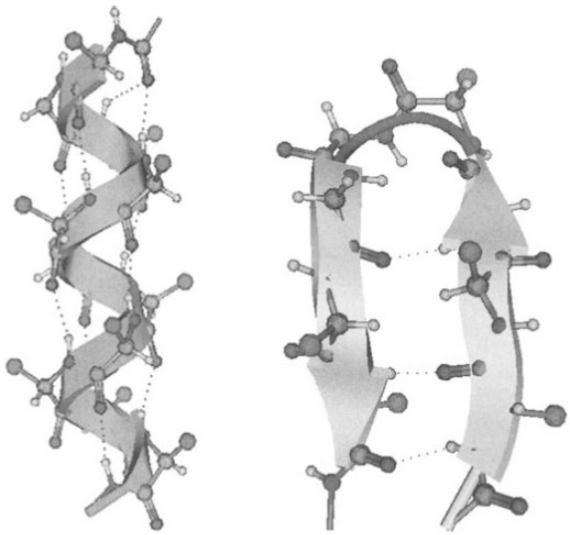


Extraído de (BRANDEN; TOOZE, 1991)

2.2.1.2 Estrutura secundária

A estrutura secundária da proteína é um conjunto de estruturas estáveis que resultam da interação entre os resíduos que compõem a proteína. Os mais comuns são as hélices-α e as folhas-β, mostradas na Figura 6. Além destas, existem ainda os *loops*, também chamados *coils*.

Figura 6 – Estrutura de uma hélice-α e um folha-β.



Extraído de (XU; XU; LIANG, 2007a)

As hélices são a estrutura secundária mais frequente em relação às demais. São sequências compostas por, em média, 10 aminoácidos, podendo atingir 40, que constituem um formato de hélice em torno de um eixo central imaginário. Cada giro da hélice é composto por cerca de 3,6 aminoácidos e é formado por uma ligação de hidrogênio entre o grupo C=O do resíduo *n* e o grupo N–H do resíduo *n+4* (XU; XU; LIANG, 2007a). Na formação da hélice, os ângulos ψ e φ correspondem a -45° a -50° e -60°, respectivamente. Experimentalmente,

sabe-se que alguns aminoácidos tem uma probabilidade mais alta de ocorrer em hélices, contudo, esse elemento não é determinante na predição desse tipo estrutura secundária. A estabilidade da estrutura é garantida por uma sequência de três a quatro ligações de hidrogênio.

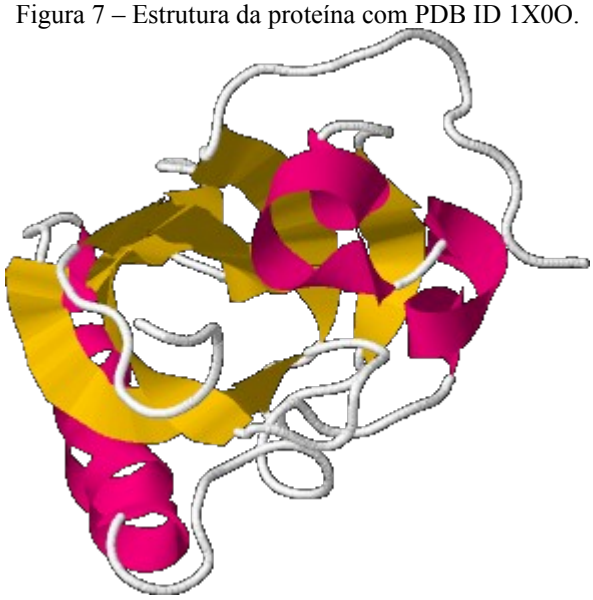
As folhas-β, por sua vez, são compostas pela conexão diversas sequências de aminoácidos, tipicamente formadas por 5 a 10 aminoácidos (NELSON; COX, 2008). Estes elementos são denominados fitas-β e tornam-se adjacente entre si, formando uma espécie de folha trançada. Similarmente ao que ocorre com as hélices-α, embora não seja um fator decisivo, alguns aminoácidos têm propensão a compor esse tipo de estrutura.

Conectando hélices-α e folhas-β, situam-se os *loops*, os quais não possuem uma forma definida. Usualmente, os *loops* situam-se no exterior da proteína, à medida que outras estruturas compõem o núcleo da proteína. Por este motivo, proteínas relacionadas evolutivamente possuem hélices e folhas bastantes semelhantes, mas podem ter *loops* bastante diferentes. Desse modo, diz-se que o núcleo da proteína é conservado (SETUBAL; MEIDANIS, 1997).

2.2.1.2.1 Estrutura supersecundária

A interação entre estruturas secundárias dá origem a estruturas maiores, chamadas motivos e domínios (NELSON; COX, 2008). Os motivos são formados pela combinação de estruturas secundárias que ocorrem em diversas proteínas. Podem efetuar alguma função, como o motivo hélice-*loop*-hélice, usado como sítio de ligação para átomos de cálcio. Outros motivos, entretanto, não têm qualquer papel no comportamento da proteína. Os domínios são estruturas mais complexas que realizam uma tarefa específica e, portanto, possuem um sítio ativo. Um sítio ativo é o local de uma proteína onde ocorrem interações com outras moléculas. Um ou mais domínios podem estar presentes em uma mesma proteína e a interação entre domínios e motivos resulta tem como resultado a estrutura terciária. A Figura 7 representa a estrutura da proteína com PDB ID 1X0O, a qual possui um domínio hélice-*loop*-hélice, situado na parte superior esquerda da figura.

Figura 7 – Estrutura da proteína com PDB ID 1X0O.



Adaptado do obtido no *Protein Data Bank* (BERMAN et al., 2000).

2.2.1.3 Estruturas terciária e quaternária

Enquanto diferentes estruturas secundárias surgem a partir de interações entre aminoácidos próximos, a estrutura terciária ocorre devido às interações de longa distância entre os aminoácidos da estrutura primária. A estrutura que a proteína atinge quando completamente dobrada é a estável do ponto de vista termodinâmico sob determinadas condições, o que resulta em uma ou algumas possibilidades. Uma proteína que encontra-se em uma conformação funcional e dobrada está em sua **conformação nativa** (NELSON; COX, 2008). A estrutura quaternária ocorre a partir da interação de diversas cadeias polipeptídicas iguais ou distintas, denominadas **subunidades**.

2.2.2 Alinhamento de Proteínas

A existência de diversos domínios e motivos recorrentes no núcleo das proteínas mostra que estas estruturas são conservadas de maneira mais confiável que a estrutura primária (SETUBAL; MEIDANIS, 1997), além disso a tendência é que, em caso de substituição de um aminoácido, o substituto tenha propriedades semelhantes às do aminoácido original (hidrofobicidade, carga). Desse modo, uma vez que aquelas estruturas têm relação com a função da proteína, a comparação de diversas proteínas pode fornecer informações sobre uma **família de proteínas**, a qual seria formada por um conjunto de proteínas que possuem semelhanças estruturais e funcionais (XU; XU; LIANG, 2007a). Normalmente, uma quantidade considerável de informação evolucionária está presente em uma família de proteínas. Conjuntos de famílias que possuem similaridade funcional e estrutural entre si, mas que não possuem similaridades entre suas estruturas primárias, são chamados de

superfamílias. As proteínas contidas em uma mesma família de proteínas são denominadas **homólogas**.

2.2.3 Problema de Dobramento das Proteínas

Sabe-se que, para a obtenção da estrutura terciária de uma proteína é necessária somente sua cadeia de aminoácidos. Isso é provado pelo **princípio termodinâmico de Anfinsen** (também chamado hipótese termodinâmica) (SELA; WHITE; ANFINSEN, 1957) (ANFINSEN, 1973). Esta hipótese foi formulada a partir da observação da que a cadeia de aminoácido de uma proteína dobra-se espontaneamente quando em condições biológicas, atingindo a sua estrutura nativa.

Uma vez que o objeto central no estudo das proteínas é a sua estrutura terciária, sua obtenção é um fator fundamental. Para isso, é necessário determinar a posição de cada aminoácido na estrutura secundária e como estas estruturas interagem de modo a determinar a estrutura tridimensional e, portanto, sua função. Esse problema é denominado **problema de dobramento da proteína**, do inglês *protein folding problem*.

Para tanto, é preciso que sejam obtidos os ângulos ψ e φ entre cada aminoácido, analisar as conformações possíveis e avaliar a quantidade de energia livre em cada uma dessas situações. Experimentalmente, sabe-se que esses ângulos em conjunto podem assumir alguns um número limitado de valores. Entretanto, o número de conformações possíveis pode tornar-se intratável. Como exemplo, supondo-se que cada ângulo pudesse assumir três valores, e considerando que cada par (ψ, φ) representa uma conformação diferente, seriam possíveis nove configurações. Para uma proteína com 100 aminoácidos, o número de conformações possíveis poderia atingir 9^{100} (SETUBAL; MEIDANIS, 1997). Neste caso, o tempo necessário para avaliar todos os estados e determinar o mais apropriado, caso este possa ser determinado como aquele que possui a menor quantidade de energia livre, seriam necessários 10^{26} segundos, o que representa mais que o tempo de existência conhecido do universo. O contraste que ocorre entre o fato que uma proteína dobra-se espontaneamente (normalmente menos de um segundo) e o tempo que seria necessário para analisar-se cada conformação possível de sua estrutura primária é denominado **paradoxo de Levinthal** (LEVINTHAL, 1969).

2.3 FUNDAMENTOS COMPUTACIONAIS

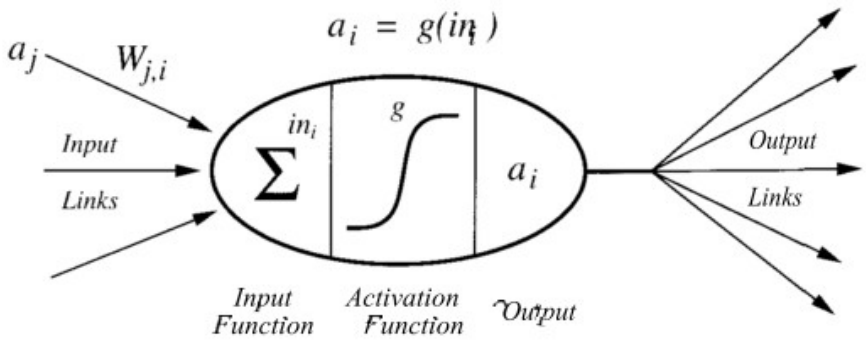
Problemas de maior complexidade motivam o desenvolvimento de diversas ferramentas computacionais para resolvê-los. Um exemplo é o aprendizado, quando é disponibilizada informação sobre o problema e deseja-se que uma ferramenta seja capaz de realizar inferências com base nos dados fornecidos, sejam conhecidas ou não as respostas esperadas para a informação fornecida. As redes neurais e as *support vector machines* são duas técnicas de Inteligência Artificial baseadas em aprendizado supervisionado, no qual são fornecidos ao modelo de aprendizado um exemplo e a saída desejada como resposta para aquele exemplo.

2.3.1 Redes neurais artificiais

Com o objetivo de se aproximar do raciocínio de seres vivos, notadamente os seres humanos, que têm sucesso na solução de problemas complexos da natureza e na absorção de conhecimento sobre elementos anteriormente desconhecidos, foi criado um modelo matemático que simula o comportamento do cérebro: as redes neurais artificiais (ANNs). Estes modelos são compostos por nós, também chamados neurônios ou unidades, interconectados entre si, e cada uma destas conexões possui um peso numérico associado.

Um modelo de neurônio foi inicialmente proposto por (MCCULLOCH; PITTS, 1943), o qual possui uma entrada e uma saída, como mostrado na Figura 8. Neste modelo, um valor de saída é devolvido caso a combinação linear dos valores da entrada seja maior que um determinado limite. Uma rede neural é um conjunto de neurônios e seu resultado depende diretamente da topologia na qual estes neurônios estão dispostos, bem como das características de cada um deles.

Figura 8 – Esquema de um neurônio.



Extraído de (RUSSEL; NORVIG, 2010)

Cada neurônio calcula uma média ponderada da entrada, dada pela equação 2.1.

$$¿_i = \sum_{j=0}^{n} W_{j,i} a_j \quad (2.1)$$

Para derivar a saída, é aplicada uma função *g*, representada pela equação 2.2, ao resultado da soma calculada.

$$a_i = g(¿_i) = g\left(\sum_{j=0}^{n} W_{j,i} a_j\right) \quad (2.2)$$

A análise do resultado realizada determina que a unidade está "ativa" quando a_i está próximo de 1, isto é, as entradas corretar foram recebidas, e "inativa" quando a_i está próximo de 0, quando entradas erradas foram recebidas. Para a função *g*, tipicamente são utilizadas as funções **limiar**, a qual devolve 1 para um resultado positivo ou 0, caso contrário; ou **sigmoide** (também chamada **função logística**).

Note-se que no cálculo da função de ativação é incluído um peso $W_{0,i}$, o qual é denominado **peso de desvio**. Sua função é definir o limite real da unidade, no sentido de que a unidade é ativada quando a soma ponderada das entradas "reais" dada por $\sum_{j=1}^{n} W_{j,i} a_j$ excede $W_{0,i}$.

Uma vez determinada a estrutura de um neurônio, o passo seguinte é a associação de um conjunto de neurônios de maneira que a estrutura do cérebro possa ser simulada. Essencialmente, existem duas formas distintas de implementação desse tipo de estrutura: as redes de alimentação direta (do inglês *feed-forward*) e as redes recorrentes (RUSSEL; NORVIG, 2010).

2.3.1.1 Redes de alimentação direta

Em uma rede neural artificial de alimentação direta (RNAD), as conexões seguem um sentido único, isto é, cada nó somente recebe valores de entrada provenientes de nós anteriores, sem a ocorrência de *loops*. Dessa maneira, esta estrutura pode ser visualizada como um grafo dirigido acíclico e sua saída é uma função do valor dado como entrada. Assim, o estado da rede é a própria rede, ou seja, o conjunto dos pesos de cada conexão.

As redes de alimentação direta podem ser organizadas em camadas de neurônios, em que um neurônio de determinada camada somente pode receber uma entrada proveniente de outro neurônio localizado na camada imediatamente antecedente. Uma RNAD na qual os nós da camada de entrada da rede estão conectados diretamente aos nós de saída são chamadas

redes *perceptron* (ROSENBLATT, 1958). De tal modo, a saída é uma função dos dados de entrada e a solução somente pode convergir se o conjunto de dados for linearmente separável. Uma RNAD que possua mais de uma camada é chamada *perceptron* de várias camadas (MLP, do inglês *multi-layer perceptron*). Neste modelo, os nós pertencentes Às camadas que não estão conectadas com as entradas ou saídas são chamados **nós escondidos** ou **camada escondida**.

2.3.1.1.1 Redes neurais de alimentação direta de uma única camada

Embora a aplicação desse modelo seja *a priori* limitada, sua vantagem é que uma função simples adaptará o *perceptron* a qualquer conjunto linearmente separável. O aprendizado é fundamentalmente o ajustamento de pesos de maneira a reduzir o erro da saída da rede, o que faz com que esse processo seja formulado como uma busca de otimização no espaço de pesos. A medida comum de erro é a soma dos erros quadráticos, dada pela equação 2.3.

$$E=\frac{1}{2}Err^2 \equiv \frac{1}{2}\left(y-h_w(x)\right)^2 \quad , \tag{2.3}$$

Para exemplo de treinamento com entrada **x** e saída verdadeira y e onde $h_W(x)$ é a saída do perceptron no exemplo e T é o valor de saída verdadeiro.

Usando o declínio de gradiente para reduzir o erro quadrático, calcula-se a derivada parcial de E em relação a cada peso:

$$\frac{\partial E}{\partial W_j}=Err\times\frac{\partial Err}{\partial W_j}$$

$$¿\ Err\times\frac{\partial}{\partial W_j}g\left(y-\sum_{j=0}^{n}W_j x_j\right)$$

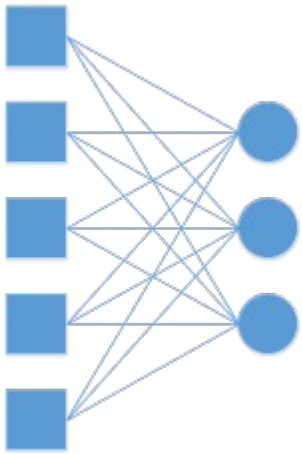
$$¿-Err\times g'(¿)\times x_j \tag{2.4}$$

No resultado, representado pela equação 2.4, *g'* é a derivada da função de ativação. Dessa maneira, utilizando-se o algoritmo de declínio de gradiente a atualização do peso é feita de acordo com a equação 2.5, onde α é a **taxa de aprendizagem**. Cada ciclo em que o erro é ajustado é chamado **época**, as quais são repetidas até que um determinado critério de parada seja atingido.

$$W_j \leftarrow W_j+\alpha\times Err\times g'(¿)\times x_j \tag{2.5}$$

A Figura 9 mostra uma representação de um *perceptron* com cinco nós na camada de entrada e três nós na camada de saída.

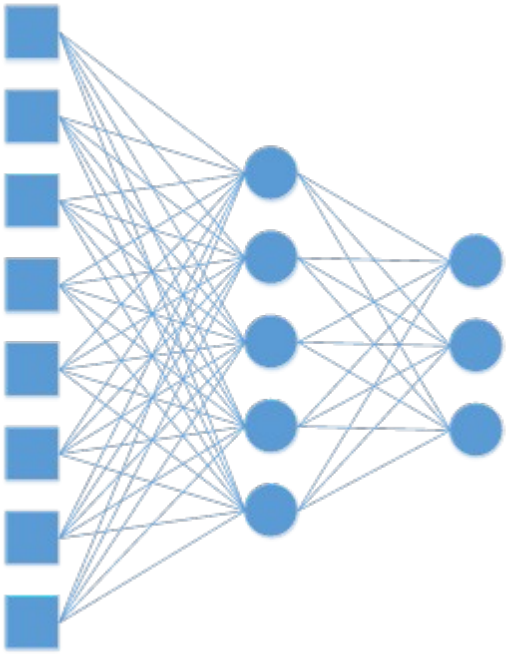
Figura 9 – Esquema de um *perceptron*.



2.3.1.1.2 Redes de alimentação direta de várias camadas

Os *perceptrons* de várias camadas (MLP) têm, em relação aos de uma única camada, a capacidade de solucionar problemas que possuam como resposta uma função contínua, à medida que os segundos em geral solucionam um espaço de entradas que seja linearmente separável. A Figura 10 mostra um esquema de um MLP com oito nós na camada de entrada, única camada escondida, com cinco nós, e três nós na camada de saída.

Figura 10 – Esquema de um MLP.



A vantagem em adicionar camadas intermediárias é motivada pelo aumento do espaço de hipóteses que cada camada pode representar. Não existe um método para determinar de maneira objetiva a quantidade de nós necessários nas camadas intermediárias da rede. O que se deve considerar é que, embora fosse potencialmente capaz de memorizar

todos os exemplos dados como entrada, uma rede com muitos nós não seria necessariamente capaz de criar generalização, isto é, responder corretamente a exemplos não conhecidos do mesmo domínio em que o aprendizado ocorreu (LORENA; CARVALHO, 2007).

Um problema associado ao uso de redes MLP é a atualização dos pesos dos nós intermediários, uma vez que os dados de treinamento não apontam o valor que seria correto para o dado nó, como ocorre com os nós da saída. Uma solução é utilizar um algoritmo de propagação de retorno ou retropropagação (do inglês *backpropagation*) (RUMELHART; HINTON; WILLIAMS, 1986) (WERBOS, 1974).

A propagação do retorno do erro de um dado nó para os seus antecedentes que pertencem às camadas intermediárias deve ser calculada de maneira proporcional, considerando que o erro obtido nas saídas é resultante da combinação dos erros nos nós intermediários. Assim, o erro da saída do nó a_i , dado pela equação 2.6 é propagado para o nó j da camada intermediária de acordo com a seguinte regra expressa pela equação 2.7 e a atualização dos pesos entre a camada de entrada e a camada intermediária é feita de acordo com a equação 2.8.

$$\Delta_i = Err_i \times g'(¿_i) \tag{2.6}$$

$$\Delta_j = g'(¿_j) \sum_i W_{j,i} \Delta \tag{2.7}$$

$$W_{k,j} \leftarrow W_{k,j} + \alpha \times a_k \times \Delta_j \quad . \tag{2.8}$$

Portanto, o processo de atualização pode ser resumido em dois passos:

- Calcular o erro a partir da observação do esperado para a saída.
- A partir da camada de saída até à camada conectada à camada de entrada, para cada camada, propagar o erro de volta até à camada seguinte e atualizar os pesos entre as duas camadas.

2.3.1.2 Redes recorrentes

Nas redes recorrentes, os valores de saída da rede são utilizados como entrada para a própria rede. Isso implica que essa rede pode atingir um estado estável ou mesmo caótico, além de ser capaz de representar um armazenamento de memória simples. Este modelo é mais próximo de um neurônio real, entretanto, a compreensão de sua estrutura é mais complexa.

2.3.1.3 Treinamento de redes neurais

A determinação da estrutura de uma rede neural mais adequada a dada situação ainda é um problema sem uma solução clara. Algo que, entretanto, pode-se tomar em consideração é

que um número muito grande de parâmetros na rede pode fazer com que a estrutura se especialize no conjunto de dados fornecidos no treinamento, contudo não necessariamente será capaz de generalizar o aprendizado a conjuntos de dados desconhecidos.

Uma abordagem comum no caso de redes completamente conectadas é a seleção dos números de nós em cada camada e o número de camadas intermediárias. A escolha pode ser feita através de validação cruzada (*cross-validation*) (KOHAVI, 1995), de modo a selecionar uma configuração de rede que atinja o melhor resultado.

A validação cruzada consiste da separação do conjunto de treinamento em diversas partições, entre os quais um é retirado e usado como conjunto de teste e os restantes são utilizados como conjunto de treinamento. O objetivo principal é analisar a capacidade de generalização do modelo testado. No caso em que o parâmetro em teste é a acurácia, a acurácia resultante é a média dos resultados de cada um dos conjuntos de testes distintos da validação cruzada. Numa validação cruzada de 7 partições (*7-fold cross-validation*, pela nomenclatura em inglês), por exemplo, em cada caso de teste seis conjuntos são utilizados como dados de treinamento e o conjunto restante é utilizado para avaliar a capacidade de acerto da estrutura treinada.

2.3.2 Support vector machines

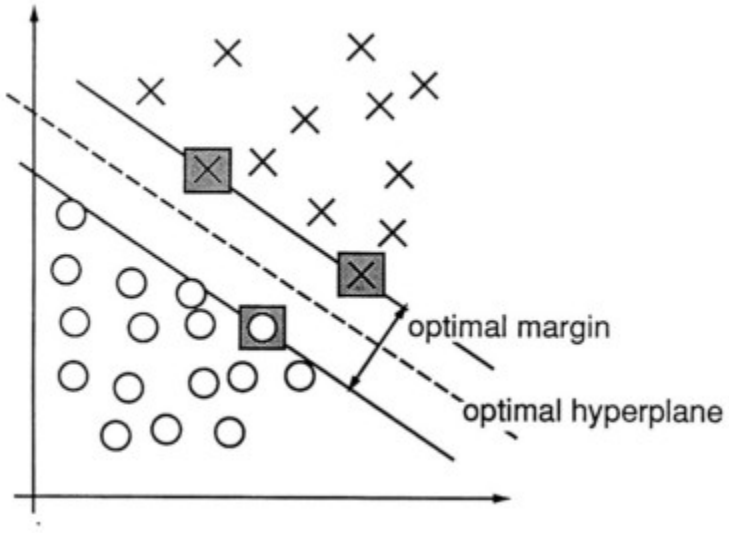
As redes neurais de propagação direta possuem alguns problemas. As RNADs de uma única camada possuem um treinamento eficiente, contudo, somente podem solucionar problemas que tenham o espaço de dados linearmente separável. As RNADs de múltiplas camadas conseguem realizar inferências sobre um problema no qual o espaço de dados seja qualquer função linear, entretanto, têm um treinamento mais complexo, uma vez que podem existir muito mínimos locais e o número de dimensões no espaço de pesos pode ser grande (RUSSEL; NORVIG, 2010).

As máquinas de vetores de suporte (do inglês *support vector machines*) são um modelo de aprendizado supervisionado pode unir o melhor dos dois casos de redes neurais, desempenhando um aprendizado mais eficiente, visto que, ao contrário do que ocorre com as RNADs, uma SVM possui um único mínimo global (LORENA; CARVALHO, 2007), o que facilita a convergência do modelo.

O conceito original de SVMs, desenvolvido por Cortes e Vapnik (CORTES; VAPNIK, 1995) consiste em um separador linear definido por hiperplano no espaço de dados de um problema de modo a determinar se os vetores de características fornecidos como

entrada fazer parte de uma entre duas classes, positiva ou negativa. Esse separador é gerado a partir da distância máxima entre os pontos que fazem parte de cada uma das classes. Na Figura 11, encontra-se uma representação da classificação feita por um hiperplano. Os pontos quadrados, a partir dos quais é determinada a posição do separador são chamados **vetores de suporte**.

Figura 11 – Hiperplano com um separador linear ótimo.



Extraído de (CORTES; VAPNIK, 1995).

Uma vez que, em geral, o número de vetores de suporte é muito menor que o número de pontos de dados, o número de efetivo de parâmetros necessário para a definição de um separador ótimo é muito menor que o número de pontos no espaço de dados. Caso o espaço de dados do problema a ser resolvido não seja linearmente separável, os vetores de entrada podem ser mapeados para um novo vetor de características em um espaço de dados suficientemente grande, o que faz com que eles sejam sempre linearmente separáveis (RUSSEL; NORVIG, 2010).

Em um ambiente de dados reais, no qual é possível a ocorrência de ruído ou *outliers*, é bastante provável que um classificador linear definido por um hiperplano não seja capaz de fazer a classificação dos dados. Com isso, o conceito desenvolvido por Cortes e Vapnik foi desenvolvido de maneira suavizar a curva do hiperplano para que os *outliers* e ruídos possam ser tolerados sem que a classificação seja prejudicada.

Caso o espaço de características no qual os vetores de entrada são mapeados tenham uma dimensão muito grande, ou mesmo infinita, a combinação linear que define o hiperplano pode ser substituída por uma **função de núcleo** (do inglês *kernel function*). Neste caso, o SVM é caracterizado fortemente pela sua função de núcleo (BURGES, 1998).

3. REVISÃO DE LITERATURA DOS PREDITORES DE ESTRUTURA SECUNDÁRIA DE PROTEÍNAS

3.1 INTRODUÇÃO

No desenvolvimento de preditores são feitas diversas abordagens, as quais englobam diferentes técnicas de Inteligência Artificial, conjuntos de entrada para treinamento e teste, diversas arquiteturas para os preditores, entre outros fatores, os quais podem influenciar fortemente a acurácia da predição de estrutura secundária.

Este capítulo tem como objetivo apresentar uma atualização do estado da arte apresentado em (MELO, 2005), enfatizando métodos que apresentaram os melhores resultados, buscando destacar o impacto de cada um dos fatores citados, na acurácia de cada preditor.

3.2 REFERENCIAL TEÓRICO

3.2.1 Considerações iniciais

Diversas técnicas vêm sendo utilizadas na abordagem do problema de predição da estrutura secundária da proteína (BETTELLA; RASINSKI; KNAPP, 2012). As redes neurais, de modo geral, apresentam os melhores resultados e, juntamente com outras técnicas de aprendizagem de máquina, como as *support vector machines* (SVM), têm sido usadas como método padrão para a predição de estrutura secundária (PAVLOPOULOU; MICHALOPOULOS, 2011).

Especificamente quanto às redes neurais, o trabalho de Qian e Sejnowski (QIAN; SEJNOWSKI, 1988) pode ser considerado um dos pioneiros, uma vez que propôs a utilização da técnica e realizou diversos testes quanto à formatação da entrada e a estrutura das redes neurais, por exemplo, de modo a maximizar o resultado do método. A acurácia alcançada foi de 64,3%, ainda um pouco distante do limite teórico de 88% (ROST, 2001). Qian e Sejnowski afirmam que o resultado alcançado é o limite de acurácia para a predição feita somente a com informação local das sequências de aminoácidos para proteínas não-homólogas. Em seu experimento, Qian e Sejnowski utilizaram o banco de dados RS126 (ROST; SANDER, 1994), concluindo também que as folhas-β são as estruturas mais difíceis de predizer, devido às

interações de longa distância, uma vez que não é possível codificá-las em uma janela de aminoácidos, e à sua reduzida proporção em relação às outras estruturas.

Para acrescentar informação aos dados de entrada na predição de estrutura secundária através de redes neurais, Jones (JONES, 1999) propôs uma abordagem que usa como entrada estruturas denominadas *Position-Specific Scoring Matrix* (PSSM), resultantes do alinhamento de uma família de proteína. Para cada proteína do conjunto de entrada, um perfil de alinhamento é construído a partir de uma base de dados de proteínas. Essas estruturas são capazes de fornecer informação evolucionária sobre as proteínas (XU; XU; LIANG, 2007b), permitindo, portanto, um aumento médio de 10% na acurácia da predição da estrutura secundária. A acurácia do preditor é de 76,5% a 78,3%, dependendo do método de atribuição da estrutura secundária, como o DSSP (KABSCH; SANDER, 1983), por exemplo. Assim como no método descrito em (QIAN; SEJNOWSKI, 1988), também são utilizadas redes neurais artificiais para classificar os dados. A técnica proposta por Jones ainda é considerada bastante eficiente quando comparada a outras, alcançando uma acurácia de pouco mais de 80%, em geral, uma vez que passa por atualizações desde que foi criada (BETTELLA; RASINSKI; KNAPP, 2012); (YASEEN; LI, 2014), tendo sido implementada no preditor PSIPRED.

3.2.2 Métodos de predição de estrutura secundária

Posteriormente aos avanços alcançados com o uso de perfis de alinhamento de proteínas, diversos métodos diferentes foram desenvolvidos e publicados. Entretanto, uma tendência clara é que grande parte daqueles estudos tem acurácia por volta de 76% a 80%, aproximadamente. Uma vez que as redes neurais têm gerando bons resultados, as implementações mais recentes têm apresentado alterações mais significativas no tratamento e formatação da entrada do que na estrutura do método de classificação. Além das redes neurais, outro conceito que se tornou bastante frequente foi a utilização de perfis de alinhamento em lugar de sequências de aminoácidos.

O método descrito por Bettella *et al* (BETTELLA; RASINSKI; KNAPP, 2012), denominado SPARROW, por exemplo, utiliza conjuntos de funções para avaliar a probabilidade de um aminoácido pertencer a determinada classe de estrutura secundária, utilizando uma abordagem de classificação binária (um-contra-um ou um-contra-o-resto). Dois conjuntos de funções são utilizados: o primeiro correlaciona a estrutura secundária do resíduo central de uma janela de resíduos de tamanho 15 no perfil de alinhamento dado como

entrada; o segundo é treinado para correlacionar a classificação desse resíduo central com as classificações dos outros resíduos da janela. Os resultados da segunda função são dados como entrada a uma rede neural artificial, que devolve como resultado a classificação para o aminoácido central de cada janela. A acurácia alcançada foi de 80,46%, com variação de ± 0,35%. No estudo comparativo apresentado, o PSIPRED consegue ter uma acurácia 0,6% melhor que o SPARROW, entretanto, o SPARROW costuma ter uma taxa de acerto maior para altos níveis de confiança da predição. O treinamento do SPARROW foi realizado com o conjunto ASTRAL40_1.73, com 9472 estruturas de domínios de proteínas. O tamanho dos dados é um ponto fraco para o experimento, uma vez que a geração dos perfis de alinhamento de famílias de proteínas, utilizados como entrada, é bastante custosa. O treinamento foi feito com uma abordagem 10-*fold cross-validation*. Possui resultados melhores que PROSPECT (XU; XU, 2000) e SSPro_4.03 (POLLASTRI et al., 2002), contudo, isso ocorre porque o SPARROW é uma implementação mais recente e tem acesso a conjuntos de dados maiores.

O preditor desenvolvido por Pollastri *et al* (POLLASTRI et al., 2002), chamado SSPro, é uma solução que propõe modificações na arquitetura do método de classificação utilizando redes neurais. Utiliza 11 redes neurais recorrentes bidirecionais, as quais podem solucionar limitações das redes unidirecionais e diferenciam-se destas pelo fato de serem capazes de identificar "contextos" da esquerda e da direita, relativamente ao resíduo central de uma janela de aminoácidos. Os melhores resultados para acurácia alcançados são de 78,13% no banco de dados RS126 (ROST; SANDER, 1994) e 77,67% no EVA (ROST; EYRICH, 2002), utilizando janelas laterais (esquerda/direita) de 3 ou 4 aminoácidos. O treinamento é tipicamente feito com uma abordagem híbrida, combinando treinamento *on-line* e em lotes; e chega ao fim após 8 iterações seguidas em que a taxa de aprendizado é reduzida. De acordo com o estudo comparativo realizado em (WEI; THOMPSON; FLOUDAS, 2011), o SSPro_4.0 apresenta os melhores resultados, juntamente com o PSIPRED, para o banco de dados PDBselect25, com acurácia de 82,57%, utilizando uma abordagem 6-*fold cross-validation*. Não consegue obter muito sucesso quanto ao problema relativo à predição das folhas-β.

O método descrito em (SARASWATHI et al., 2012), denominado FLOPRED (*Fast Learning Optimized Prediction Methodology*), propõe uma máquina de aprendizado extremo (*extreme learning machine* - ELM) baseada em redes neurais e utiliza otimização por enxame de partículas. O objetivo é determinar o número de neurônios intermediários de modo que ocorra a menor diferença possível entre os conjuntos de treinamento, teste e validação. Um

modelo que implemente essa característica é mais suscetível a alcançar uma melhor generalização sobre futuras amostras. O método utiliza como entrada potencial de informação baseado em conhecimento, calculado utilizando o algoritmo CABS. O algoritmo CABS é capaz de codificar interações de curta e longa distância na proteína, obtendo um conjunto de 27 características para representar cada proteína. Teoricamente, o método ELM acelera as computações, fornecendo uma capacidade de generalização maior quando comparado com outros métodos. Um grande fator limitante para FLOPRED é a geração dos perfis baseados em conhecimento com o algoritmo CABS, uma vez que o cálculo do perfil para uma proteína de pouco menos de 100 aminoácidos pode levar até dois dias, ou mesmo uma semana para uma proteína com 1500 aminoácidos. Os dados são compostos pelas proteínas do banco de dados CB513 (CUFF; BARTON, 2000) e pelos potenciais baseados em informações extraídos com o CABS, com uma abordagem 5-*fold cross-validation*. Não são gerados perfis de alinhamento de famílias de proteínas, embora os dados codifiquem internamente informação evolucionária. A acurácia de treinamento alcançada de 83% a 87% e de 81% a 84% com teste feito com validação cruzada.

De modo semelhante ao que é feito em (BETTELLA; RASINSKI; KNAPP, 2012), a técnica apresentada por Yaseen e Li (YASEEN; LI, 2014) é baseada em funções que devolvem algum resultado numérico que corresponde à sua classificação como estrutura secundária. No caso deste preditor, denominado SCORPION, as avaliações sobre a favorabilidade de determinado resíduo pertencer a dada classe são dadas como entrada a uma rede neural, juntamente com os perfis de alinhamento das proteínas (PSSM). A acurácia atingida utilizando o banco Cull7987, que contém 7987 proteínas com até 25% de semelhança, utilizando uma abordagem 7-*fold cross validation*, é de 82,74%. Um fator importante quanto ao método é que a melhora na acurácia dá-se por causa de melhora na predição de folhas-β. Uma redução significativa de erros na classificação é observada após a inserção de pontuação baseada no contexto na entrada da rede neural. Embora tenha havido melhora na predição de hélices-α e folhas-β, a acurácia na predição de *coils* é cerca de 5% menor em comparação com o PSIPRED (JONES, 1999). Em geral, entretanto, a melhora na acurácia em relação ao PSIPRED é de 0,5% a 2%.

Outro estudo que utiliza redes neurais é o apresentado por Qu *et al* (QU et al., 2011). O conceito envolvido é o uso de uma rede neural com *multi-modal back-propagation* (MMBP), o uso da Teoria de Descobrimento do Conhecimento baseado no Mecanismo Cognitivo Interior para construir um *Compound Pyramid Model* (CPM), que é composto por

três camadas que integram um MMBP, *mixed-model* SVM (MMS) e o processo modificado de *Knowledge Discovery in Databases* (KDDn). A primeira camada, denominada análise compreensiva, combina os resultados de predição do módulo MMBP e MMS; a segunda, o julgamento do núcleo, aplica o processo KDD e o algoritmo M para obter regras de classificação estrutural da entrada; a terceira; o julgamento assistente, aplica o KDD e o algoritmo M sobre os dados para obter regras de classificação de atributos. Para desenvolver o módulo MMBP, em específico, 480 proteínas do conjunto CB513 (CUFF; BARTON, 2000) foram selecionadas, tendo sido removidas todas as sequências menores que 30 aminoácidos ou famílias que continham poucas sequências que não foram válidas para gerar um alinhamento PSI-BLAST. Após isso, quatro perfis com informação evolucionária sobre as propriedades hidrofóbicas, ligações de oxigênio e carga físico-química são utilizados como entrada para a MMBP. Cada resíduo é codificado por um vetor de características com tamanho 24*w, onde w é o tamanho da janela de aminoácidos. O Sistema MMBP consiste de duas camadas: a primeira camada é uma rede classificando um trecho da cadeia para cada uma das estruturas secundárias. A camada de entrada compreende 360 nós, divididos em 15 grupos de 24 nós. A segunda camada é usada para filtrar saídas sucessivas da primeira rede. Esta rede tem uma camada de entrada compreendendo 60 nós, divididos em 15 grupos de 4, além de 60 nós intermediários. O processo de treinamento para quando o erro de validação aumenta durante um determinado número de iterações ou quando o número máximo de iterações é atingido. Para reduzir o viés da seleção de alguma classe, o treinamento utiliza uma abordagem 7-*fold cross-validation* e realiza treinamento estratificado. A acurácia alcançada é de 86,13%. Um detalhe importante a ser notado é que a atribuição de estrutura secundária utilizada é a mais restritiva, uma vez que, das oito classes fornecidas pelo DSSP (KABSCH; SANDER, 1983), somente H (hélice-α) é convertido em H, o E (folha-β) é convertido em E. Todas as outras classes são convertidas em C (coil). Este ripo de conversão de classes normalmente apresenta acurácias menores na predição.

O modelo apresentado em (COLE; BARBER; BARTON, 2008), denominado JPred, pode ser considerado híbrido, uma vez que combina dois conceitos com abordagens distintas: redes neurais e Modelos Ocultos de Markov. São utilizados PSSM e perfis de Modelos Ocultos de Markov (HMMER). Os resultados de ambos os métodos são fornecidos a uma rede neural do algoritmo JNet. O método foi desenvolvido realizando um treinamento com abordagem 7-*fold cross-validation* e um conjunto de dados não redundante obtido do compêndio ASTRAL dos dados do SCOP (release 1.71). O teste cego com 149 proteínas

produziu uma acurácia de aproximadamente 81,5%, cerca de 5% a mais que outras versões do JNet. O usuário pode fornecer o alinhamento das proteínas (PSSM), o que excluiria a fase correspondente no algoritmo do JPred, fazendo com que seja executado somente a etapa do HMMER. Entretanto, essa remoção representa uma queda de performance de 81,5% para 80,3%. Um benefício ao fornecimento do alinhamento pelo usuário é que, nessa situação, os resultados da predição são devolvidos em menos de 2 minutos, em média.

Diferentemente dos anteriores, o método descrito em (WEI; THOMPSON; FLOUDAS, 2011), denominado CONCORD, é baseado em consenso. O objetivo é utilizar as pontuações de confiança de cada resíduo dada por cada um dos preditores para determinar a probabilidade de o resíduo pertencer a determinada classe de estrutura secundária, utilizando a otimização da combinação linear de inteiros (mixed integer linear optimization - MILP). Sete métodos de predição de estrutura secundária foram selecionados, dentre eles DSC (KING; STERNBERG, 1996), PROF (OUALI; KING, 2000), PSIPRED (JONES, 1999), Predator (FRISHMAN; ARGOS, 1997) e SSpro (POLLASTRI et al., 2002). A combinação de diversos métodos pode permitir o tratamento adequado de erros sistemáticos, de maneira que não influenciem negativamente o resultado final. Outra motivação para o desenvolvimento de preditores baseados em MILP é que alguns preditores têm melhor performance em determinadas regiões da proteína, de modo que os pontos fortes de cada técnica podem ser associados. Entretanto, a simples combinação de resultados de outros preditores faz com que a implementação baseada em MILP seja fortemente dependente das técnicas que utiliza. O conjunto de treinamento é o PDBselect25, o mesmo utilizado em (POLLASTRI et al., 2002), e o treinamento tem uma abordagem 6-*fold cross-validation*, resultando uma acurácia de 83%.

Uma abordagem distinta das mais comuns baseadas em redes neurais é a relatada em (MARTIN; GIBRAT; RODOLPHE, 2006), que implementa a predição de estrutura secundária utilizando HMMs. Uma das motivações para o uso desta técnica é que as redes neurais não são modelos interpretáveis intuitivamente. É possível realizar a predição de estrutura a partir de perfis de múltiplas sequências. Esse processo dá-se em três passos: é realizada uma busca no PSI-BLAST, que devolve uma família de N sequências homólogas (PSSM); a estrutura secundária de cada sequência é predita independentemente utilizando HMM; e as N predições são combinadas em uma única. Este preditor, chamado OSS-HMM, foi treinado com 2019 sequências do ASTRAL_1.65, com uma abordagem 4-*fold cross-validation*, e testado com um conjunto de teste independente, contendo 505 sequências. O método, no entanto, ainda não

tem a capacidade de predizer com melhor qualidade as folhas-β, uma das principais barreiras ao aumento na acurácia do experimento, que alcançou 75,5% de acurácia.

A solução dada em (LIN et al., 2010), denominada SymPred, utiliza conceitos aplicados à língua natural de maneira a identificar regiões de semelhança nas proteínas. Similaridades locais em proteínas podem indicar uma região conservada, que pode ser utilizada para melhorar a acurácia da predição. São criados dicionários de sinônimos e a informação compartilhada pode ser usada para inferir superposições estruturais tridimensionais. O classificador é treinado com 8297 sequências pertencentes ao DSSPNR-25, com 10-*fold cross-validation*, com até 25% de semelhança, e alcançou acurácia de 81%. Assim como as HMM, este é um modelo interpretável, entretanto, a acurácia de métodos baseados em padrões chega a ser até 10% menor do que em outros métodos.

O método descrito por Huang e Chen (HUANG; CHEN, 2013), baseado em *Support Vector Machines* (SVM) e extração de características é dividido em três etapas. A primeira etapa, correspondente à extração de características, obtém informações sobre cinco características específicas: (1) parâmetros de conformação, (2) a matriz de alinhamento da família da proteína (PSSM), (3) a carga global, (4) hidrofobicidade e (5) massa da cadeia lateral. Assim como na maior parte dos outros métodos, utiliza janelas de aminoácidos para predizer a classe do resíduo central dessa janela de modo a observar a interferência dos vizinhos nesse resultado. O melhor tamanho de janela é obtido de maneira experimental. A base de dados selecionada é a CB513, que contém 513 proteínas não-homólogas. A função de núcleo (*kernel function*) utilizada no classificador da SVM é a RBF (*Radial Basis Function*). Os parâmetros ótimos e a acurácia são obtidos com uma abordagem *3-fold cross-validation*. O tamanho ideal encontrado para a janela de aminoácidos foi 13 e a acurácia alcançada, com filtragem de resultados, foi 79,52%.

3.3 RESUMO

No Quadro 2 são listadas alguns dos principais parâmetros que podem ser utilizados para comparar os diferentes métodos listados. Para facilitar a análise comparativa dos métodos, foram selecionados alguns parâmetros que caracterizam cada um dos métodos, são eles o método utilizado para codificar os dados, qual a entrada dada ao preditor, a acurácia alcançada e o número de proteínas utilizadas no treinamento e teste do método de predição. O objetivo central é avaliar a relação entre o método utilizado, a quantidade de informações

fornecidas pelo tipo de entrada e pelo número de proteínas que foram codificadas com o resultado obtido.

Quadro 2 – Resumo dos métodos de predição de estrutura secundária.

Referência	**Métodos**	**Entrada**	**Acurácia (Q_3)**	**Tamanho de conjunto de dados (em proteínas)**
(QIAN; SEJNOWSKI, 1988)	Redes neurais	Proteínas representadas ortogonalmente como sequências de aminoácidos	64,3%	126
(JONES, 1999)	Redes neurais	Perfis de alinhamento de famílias de proteínas	76,5% a 78,3%, dependendo da atribuição de estrutura secundária (relatado de até 81% em outros estudos)	187
(BETTELLA; RASINSKI; KNAPP, 2012)	Redes neurais	Pontuações sobre janelas de resíduos resultantes de funções de classificação binária aplicadas sobre perfis de alinhamento de famílias de proteínas	80,46% ± 0,35%	9472
(POLLASTRI et al., 2002)	Redes neurais recorrentes bidirecionais	Perfis de alinhamento de famílias de proteínas	78%	1180, para validação cruzada, e dois grupos para teste, o RS126
(SARASWATHI et al., 2012)	*Extreme learning machines* baseadas em redes neurais, utilizando otimização por	Perfis de frequência de criado com o algoritmo CABS sobre perfis de alinhamento de famílias de	Variação de 83% a 87% em treinamento e 81% a 84% utilizando validação	513

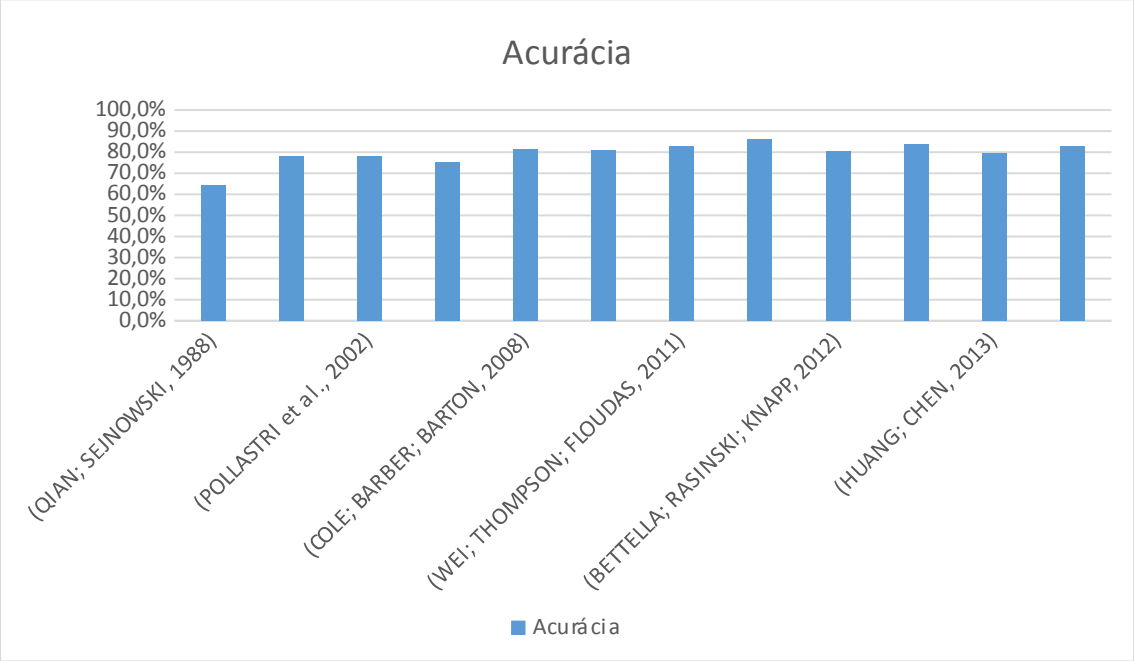
	enxame de partículas (PSO) para determinar o número de nós intermediários	proteínas	cruzada	
(YASEEN; LI, 2014)	Redes neurais	Pontuações feitas com função de classificação dos aminoácidos com base no contexto de uma janela. Aplicada sobre perfis de alinhamento de famílias de proteínas	82,74%	7987
(QU et al., 2011)	Combina redes neurais de retropropagação multi-modal (MMBP), *mixed-modal* SVM (MMS), o módulo classificador de associação estrutural (SAC) e o módulo classificador de associação de atributo.	Perfis de informação evolucionária obtidos a partir de perfis de alinhamento de famílias de proteínas	86,13%	256
(COLE; BARBER; BARTON, 2008)	Redes neurais	Perfis de alinhamento de famílias de proteínas e perfis de modelos ocultos de Markov	81,5%	7987
(WEI; THOMPSON; FLOUDAS, 2011)	Combinação de resultados de outros preditores de estrutura secundária	Pontuações de confiabilidade da predição realizada pelos métodos combinados	83%	3000
(MARTIN; GIBRAT;	Modelos ocultos de Markov	Perfis de alinhamento de	75,5%	2019 para validação e

RODOLPHE, 2006)	(HMM)	famílias de proteínas		505 para teste
(LIN et al., 2010)	Dicionários de sinônimos	Perfis de alinhamento de famílias de proteínas	81%	8297
(HUANG; CHEN, 2013)	*Support Vector Machines*	Parâmetros de conformação, a matriz de alinhamento da família da proteína (PSSM), a carga global, hidrofobicidade e massa da cadeia lateral	79,52%	513

3.3.1 Acurácia dos métodos citados

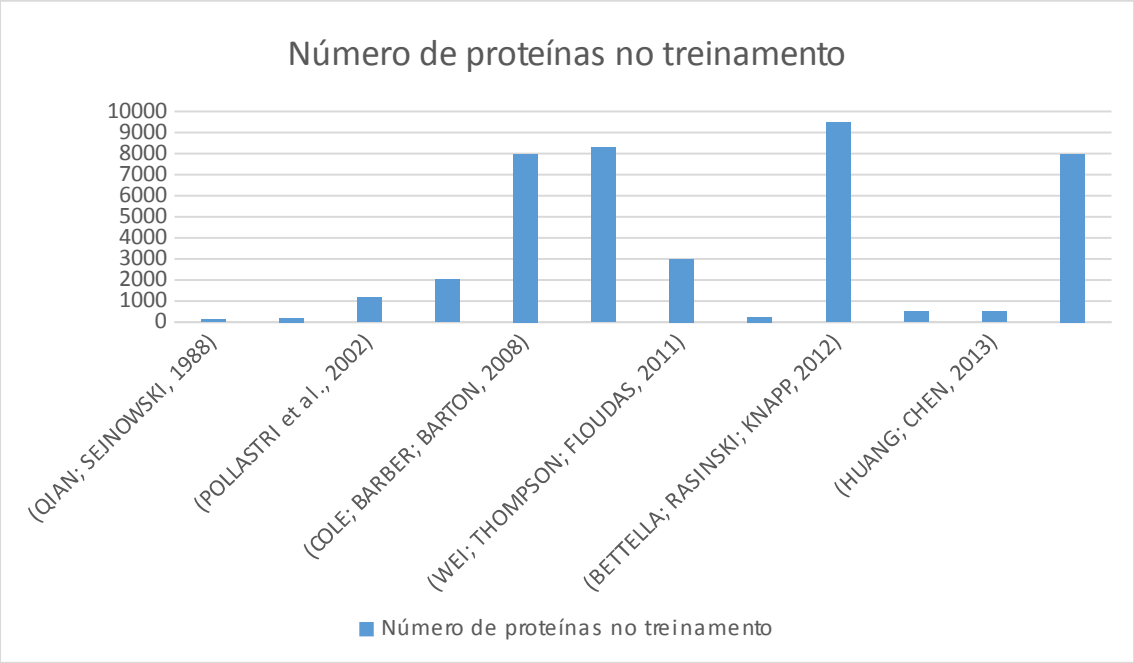
Nos Gráficos 1 e 2 estão graficamente representadas as acurácias dos métodos abordados e a quantidade de proteínas no conjunto de treinamento, respectivamente, de modo a verificar a ocorrência de algum relacionamento entre essas informações.

Gráfico 1 – Relação das acurácias dos métodos citados.



Os métodos mais recentes, quando não utilizam conjuntos mais comuns, como RS126, CB396 ou CB513, fazem uso de uma quantidade maior de informação no treinamento, da ordem de milhares de proteínas.

Gráfico 2 – Relação do número de proteínas nos conjuntos de treinamento.



3.4 DISCUSSÃO

Pode-se observar a partir do Quadro 2 e doas Gráficos 1 e 2 que, embora o conjunto de dados possa melhorar o resultado da predição, uma vez que mais exemplos são fornecidos ao método de aprendizagem, essa não é uma relação direta, uma vez que a metodologia aplicada no processo, os conjuntos de dados e características internas desses conjuntos, como a proporção de cada tipo de estrutura secundária, podem ser determinantes na melhora do desempenho.

Métodos baseados no aprendizado de redes neurais ainda são os mais utilizados ao tratar-se de predição de estrutura secundária. Alguns fatores que podem justificar isso são o número de possibilidades abertas de configurações de redes neurais, seja no uso de diferentes números de nós, seja na combinação de diversas redes. Entretanto, algo que pesa negativamente para as redes neurais artificiais é o fato de que são modelos de difícil interpretação, isto é, não é trivial inferir como uma rede neural consegue codificar uma determinada característica do conjunto de treinamento.

Em contrapartida, os modelos baseados em HMMs têm essa capacidade, o que torna possível o desenvolvimento de proteínas com um determinado conjunto de estrutura secundária (MARTIN; GIBRAT; RODOLPHE, 2006). Contudo, os métodos desenvolvidos até a atualidade têm apresentado um resultado abaixo da média obtida com as redes neurais

artificiais, entre 70% a 75% e 75% a 83%, respectivamente. Abordagens que utilizam combinações de ambos, como o JPred (COLE; BARBER; BARTON, 2008), o qual está entre os principais preditores de estrutura secundária, são boas alternativas.

O preditor FLOPRED (SARASWATHI et al., 2012), em específico, possui uma abordagem bastante quanto à obtenção da quantidades de nós em uma rede neural que devolve resultados ótimos para o conjunto de treinamento. O uso do algoritmo de otimização por enxame pode ser uma solução ao problema, evitando que a abordagem utilizada seja para isso seja puramente empírica, contudo, em compensação, a execução do algoritmo consome recursos em uma ordem muito maior em comparação aos métodos que não utilizam este algoritmo em seu desenvolvimento. Com o avanço constante de recursos de *hardware*, questões relacionadas à demanda de recursos físicos, embora também possam representar uma barreira, em geral são problemas menores quando contrapostos a problemas lógicos, como é o caso do aprendizado de máquina. Portanto, é uma alternativa que deve ser considerada e avaliada no desenvolvimento futuro de métodos de predição de estrutura secundária.

É possível também observar que, nos métodos mais recentes, as PSSMs, capazes de codificar informação evolucionária da proteína, são amplamente utilizadas, mesmo em métodos que não empregam redes neurais artificiais como modelo de aprendizado. O principal motivo é que a quantidade de informação codificada nas PSSMs é maior que aquela que se pode extrair somente da cadeia de aminoácidos (QIAN; SEJNOWSKI, 1988) (JONES, 1999). Em alguns casos, como no preditor SPARROW (BETTELLA; RASINSKI; KNAPP, 2012), por exemplo, a PSSM não é dada como entrada diretamente ao método de aprendizado, mas opera-se sobre as mesmas de modo a tentar extrair informação de maneira diferente. Neste caso, em específico, funções de avaliação da probabilidade de um dado resíduo pertencer a determinada classe são aplicadas sobre os perfis de alinhamento.

Uma exceção são métodos que realizam a combinação dos resultados de preditores, os quais são também uma abordagem interessante, uma vez que têm uma tendência a reduzir a influência dos pontos fracos de cada preditor utilizado e a potencializar seus pontos fortes de modo a tornar o resultado mais eficaz. Como exemplo, pode ser citada precisão da predição de hélices-α e folhas-β, a qual pode variar, dependendo da metodologia e dos dados de treinamento de um preditor.

O uso de SVMs para a predição de estrutura secundária é mais recente que o uso de redes neurais, e bons resultados têm sido alcançados. Contudo, alguns preditores baseados em

redes neurais ainda têm resultados estatísticos melhores e o número de métodos baseados em SVMs disponíveis ainda é pequeno em relação aos que usam ANNs.

O preditor implementado no trabalho de Huang e Chen (HUANG; CHEN, 2013), em particular, utiliza uma quantidade de informação biológica maior que a utilizada na maioria nos preditores, as quais normalmente têm como base informação obtida a partir de perfis de alinhamento de famílias de proteínas. A entrada fornecida por Huang e Chen ao modelo de aprendizado engloba cinco elementos: (1) parâmetros de conformação, (2) a matriz de alinhamento da família da proteína (PSSM), (3) a carga global, (4) hidrofobicidade e (5) massa da cadeia lateral

Parâmetros de conformação são as proporções que os aminoácidos tendem a pertencer a determinada classe de estrutura secundária; são calculados a partir da base de dados e obtidos através da fórmula $S_{ij} = a_{ij}/a_i$, onde i é um número de 1 a 20 que corresponde a algum dos aminoácidos e j é um número de 1 a 3 que corresponde a cada uma das classes de estrutura secundária. Os parâmetros de conformação são utilizados porque o dobramento de determinados aminoácidos tem relação direta com a formação de uma estrutura específica. A segunda característica utilizada como entrada é a matriz de alinhamento da proteína (PSSM).

Quanto à carga líquida, há cinco aminoácidos que a possuem: R, D, E, H e K. Aminoácidos com mesma carga elétrica repelem-se, tornando-se adversos à formação de hélices-α. Além disso, os resíduos contínuos de uma folha-β não podem possuir a mesma carga. A hidrofobicidade é importante ao conhecimento do dobramento da proteína uma vez que resíduos hidrofóbicos tendem a manter-se no núcleo da proteína, à medida que os hidrofílicos tendem a permanecer no exterior.

A aplicabilidade de modelos de aprendizado é ampla em grande parte devido à capacidade que estes têm de codificar informação sobre o dobramento das proteínas sem que seja necessário analisar todos os estados possíveis e escolher o que seria a conformação nativa da proteína. Além disso, a eficácia que alguns preditores têm atingido torna o uso de Inteligência Artificial uma ótima alternativa de análise da estrutura secundária das proteínas.

4. DESCRIÇÃO DO SISTEMA PROPOSTO

4.1 SELEÇÃO DO MÉTODO DE PREDIÇÃO

Com base nas necessidades de acurácia na predição da estrutura das proteínas e utilizando como referência o preditor descrito em (MELO, 2005), neste trabalho é proposto um método de predição baseado em redes neurais de alimentação para a frente de várias camadas (*multilayer perceptron*).

A seleção de redes neurais como método de predição é motivada principalmente pela popularidade do método, sendo de uso bastante comum para a esse tipo de aplicação. Uma vez que é um problema ainda sem solução definitiva, a seleção da configuração da rede, como o número de nós das camadas de entrada, intermediárias e de saída, o potencial de aprendizado das redes neurais ainda dá abertura a diversos testes com diversas configurações ainda inexploradas e diferentes conjuntos de dados. Além disso novas abordagem baseadas em conceitos recentes, como aprendizado profundo (do inglês *deep learning*) (AREL; ROSE; KARNOWSKI, 2010) ou redes neurais baseadas em computação em nuvem (IKRAM et al., 2013), entre outros.

Um segundo fator motivador da seleção deste método, que é uma consequência de sua popularidade, é a análise comparativa possível, dado o número de métodos já explorados e a disponibilidade de pacotes de *software* populares que implementam esse tipo de tecnologia.

4.2 METODOLOGIA DE TREINAMENTO

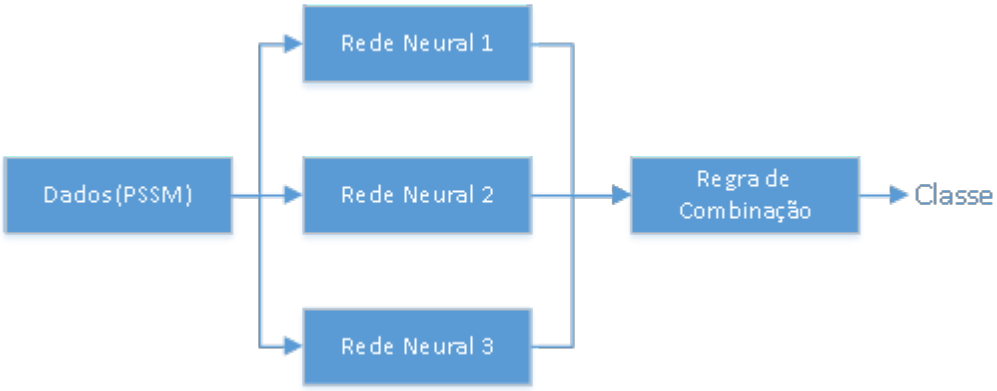
O treinamento do método tomará como ponto de partida a hipótese de que a comparação de dois métodos de predição de estrutura secundária somente pode ser feita de maneira precisa quando os dados de treinamento forem os mesmos, uma vez que, caso ocorra o contrário, qualquer dos métodos comparados pode sofrer impacto positivo ou negativo. A quantidade de informação fornecida pelos dados de treinamento, por exemplo, pode aumentar a capacidade de generalização de um dos métodos analisados e, consequentemente, melhorar o resultado da predição. Desse modo, e utilizando como referência de comparação o método desenvolvido por Huang e Chen (HUANG; CHEN, 2013), o qual utiliza máquinas de vetores de suporte (SVM) como modelo de aprendizado, neste trabalho foram utilizados no

treinamento o conjunto de 513 proteínas não-homólogas produzidas por Cuff e Barton (CUFF; BARTON, 2000), o CB513, e validação cruzada com 3 partições (*3-fold cross-validation*). O algoritmo de atualização dos pesos utilizados será o RPROP, uma variação mais eficiente do algoritmo de retropropagação, o qual requer um número menor de épocas para atingir a convergência (RIEDMILLER; BRAUN, 1992). Os resultados serão discutidos no capítulo 5, de modo a evidenciar hipóteses em relação ao comportamento de cada um dos métodos na predição de estrutura secundária de proteínas.

4.3 ESTRUTURA DO MÉTODO DESENVOLVIDO

O preditor desenvolvido por Melo (MELO, 2005), mostrado na Figura 12 consiste da combinação do resultado de três redes distintas. A proposta de preditor deste trabalho possui duas fases distintas: a primeira, igual ao preditor de Melo, denominada **sequência-estrutura**, é composta por três redes neurais multicamada (MLP), as quais recebem como entrada uma janela (intervalo) de aminoácidos das proteínas de entrada e devolve um vetor que corresponde a uma classe de estrutura secundária que corresponde ao aminoácido central da janela de entrada; a segunda, denominada **estrutura-estrutura**, composta por uma rede neural multicamada (MLP), recebe como entrada uma janela de classes de estrutura secundária correspondentes à saída fornecida pela primeira etapa e devolve uma classe de estrutura secundária que corresponde à classe central da janela de entrada. A inserção de uma segunda fase no preditor proposto por Melo tem como objetivo filtrar predições incorretas realizadas e, consequentemente, melhorar o resultado da predição. Esta fase possui uma estrutura semelhante à primeira, entretanto, é adaptada aos dados das classes de estrutura secundária que são fornecidos como entrada. A Figura 13 mostra um esquema da estrutura do preditor e as Figuras 14 e 15 apresentam a primeira e segunda fases, respectivamente, de forma mais detalhada.

Figura 12. Arquitetura do preditor desenvolvido por Melo.



Adaptado de (MELO, 2005).

Figura 13. Arquitetura do preditor desenvolvido.

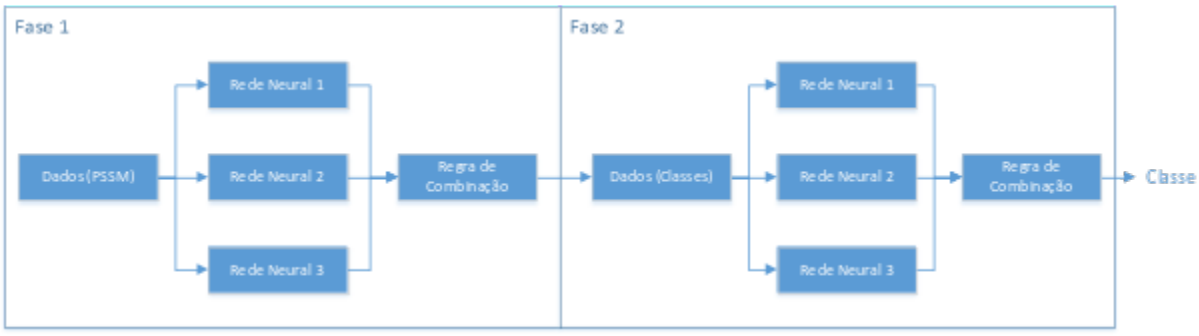


Figura 14 – Estrutura da fase sequência-estrutura do preditor proposto.

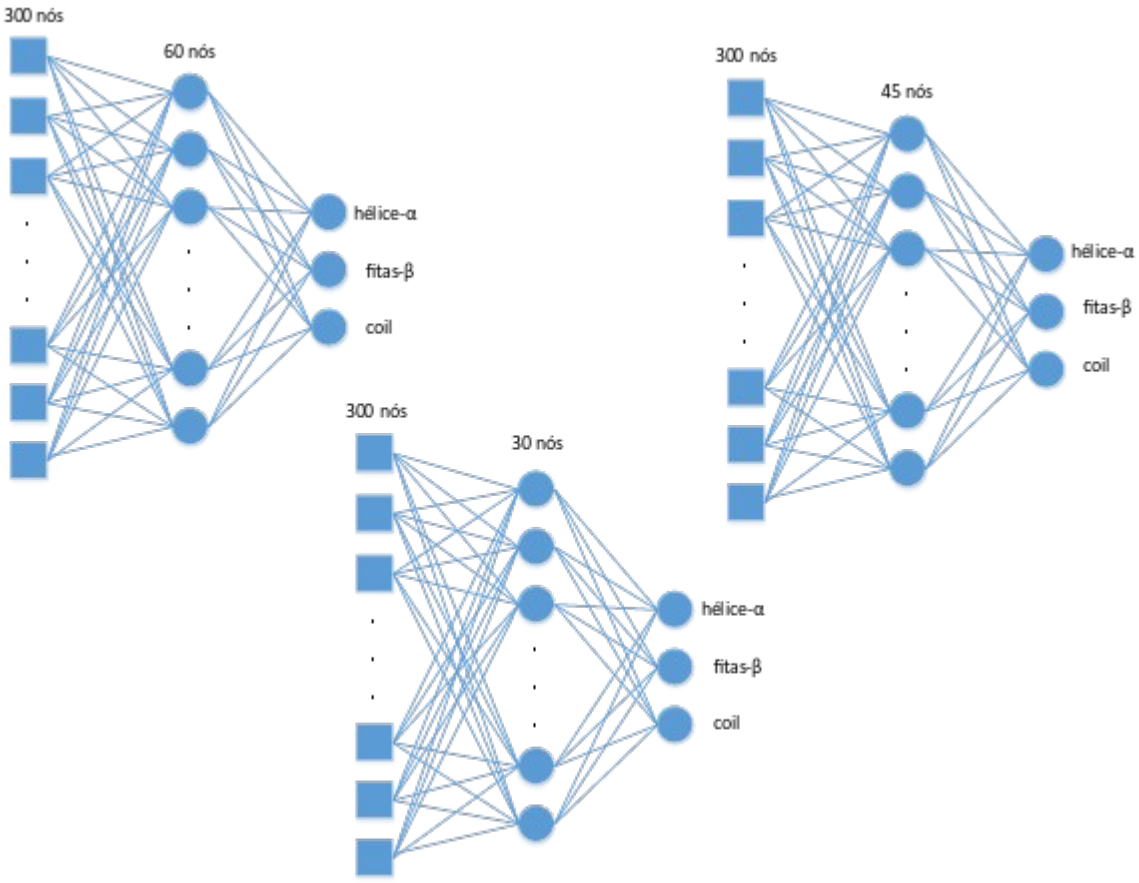
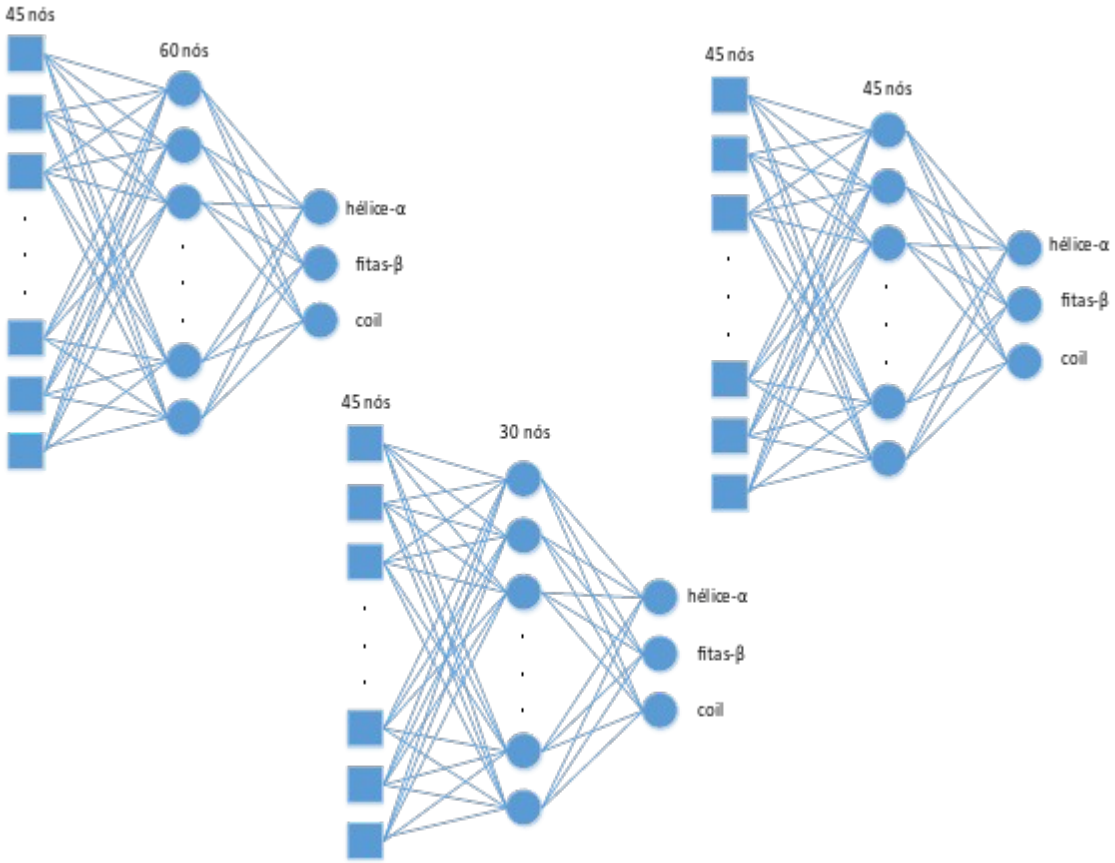


Figura 15 – Estrutura da fase estrutura-estrutura do preditor proposto.



Na primeira fase, as três redes possuem números distintos números de nós intermediários distintos e seus resultados são combinados de modo a criar um consenso entre diferentes treinamentos, haja vista que se espera que a confiabilidade da predição seja maior quando diversas predições são feitas de maneira independente. Os métodos de combinação dos resultados das três redes serão abordados oportunamente.

Embora a determinação do número de nós para a predição ótima tenha sido feita de maneira experimental, um elemento que influenciou fortemente a decisão foi o fato de que um número muito grande de nós poderia fazer com que ocorresse superajuste dos dados (do inglês *overfitting*), reduzindo a capacidade de generalização do preditor sobre dados desconhecidos (CHANDONIA; KARPLUS, 1996).

4.3.1 Camada de Entrada

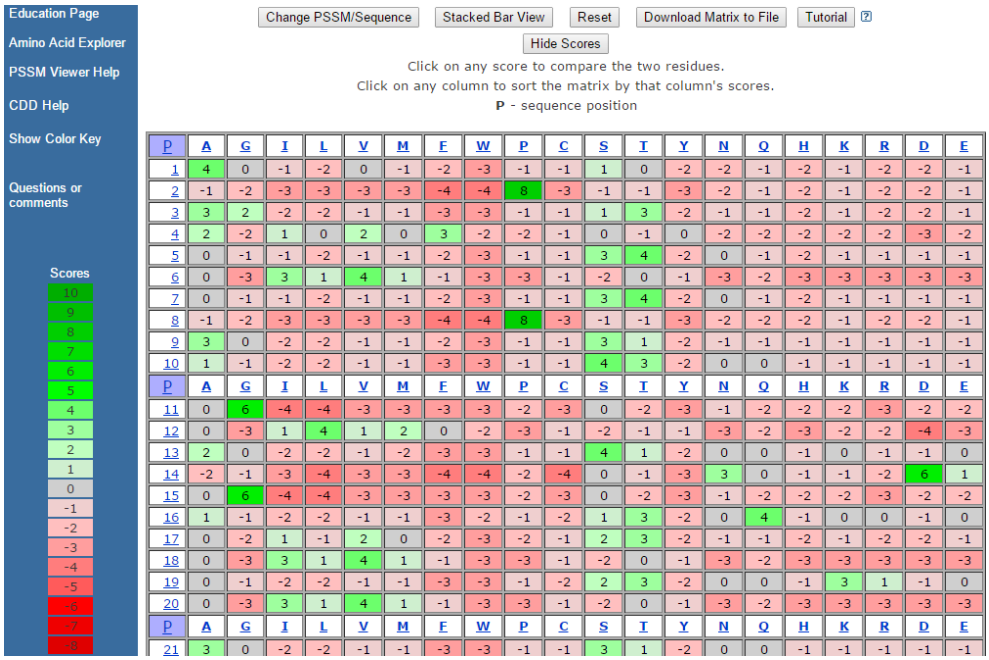
Como relatado na revisão bibliográfica, inicialmente as proteínas eram representadas de forma ortogonal. Um exemplo foi o experimento de Qian e Sejnowski (QIAN; SEJNOWSKI, 1988), que utilizou uma representação ortogonal para os aminoácidos, de modo que cada aminoácido corresponde a um vetor com 20 dígitos. Cada um desses dígitos

indica um dos aminoácidos e o aminoácido é indicado com um número 1 na posição do vetor que correspondente àquele aminoácido, da seguinte maneira;

- 10000000000000000000 – Alanina;
- 01000000000000000000 – Arginina;
  ⋮
- 00000000000000000001 – Valina.

Com o objetivo de agregar mais informação ao aprendizado de maneira a melhorar o acerto da predição, Jones propôs o uso de matrizes de pontuação de posições específicas (PSSM, do inglês *position-specific scoring matrix*) (JONES, 1999), uma vez que estas estruturas, obtidas a partir do alinhamento de uma família de proteínas são capazes de representar informação evolutiva de uma dada proteína. Os dados foram obtidos do *Protein Data Bank* (PDB) (BERMAN et al., 2000), que constrói uma PSSM a partir de uma proteína dada como entrada. Uma busca é realizada no PDB de modo a encontrar, através de alinhamento múltiplo, proteínas que pertençam à mesma família da proteína dada como entrada.

Figura 16 – Matriz de alinhamento (PSSM) para a proteína com PDB ID 1ACX.



P	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	4	0	-1	-2	0	-1	-2	-3	-1	-1	1	0	-2	-2	-1	-2	-1	-2	-2	-1
2	-1	-2	-3	-3	-3	-3	-4	-4	8	-3	-1	-1	-3	-2	-1	-2	-1	-2	-2	-1
3	3	2	-2	-2	-1	-1	-3	-3	-1	-1	1	3	-2	-1	-1	-2	-1	-2	-2	-1
4	2	-2	1	0	2	0	3	-2	-2	-1	0	-1	0	-2	-2	-2	-2	-2	-3	-2
5	0	-1	-1	-2	-1	-1	-2	-3	-1	-1	3	4	-2	0	-1	-2	-1	-1	-1	-1
6	0	-3	3	1	4	1	-1	-3	-3	-1	-2	0	-1	-3	-2	-3	-3	-3	-3	-3
7	0	-1	-1	-2	-1	-1	-2	-3	-1	-1	3	4	-2	0	-1	-2	-1	-1	-1	-1
8	-1	-2	-3	-3	-3	-3	-4	-4	8	-3	-1	-1	-3	-2	-2	-2	-1	-2	-2	-1
9	3	0	-2	-2	-1	-1	-2	-3	-1	-1	3	1	-2	-1	-1	-1	-1	-1	-1	-1
10	1	-1	-2	-2	-1	-1	-3	-3	-1	-1	4	3	-2	0	0	-1	-1	-1	-1	-1
P	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	0	6	-4	-4	-3	-3	-3	-3	-2	-3	0	-2	-3	-1	-2	-2	-2	-3	-2	-2
12	0	-3	1	4	1	2	0	-2	-3	-1	-2	-1	-1	-3	-2	-3	-2	-2	-4	-3
13	2	0	-2	-2	-1	-2	-3	-3	-1	-1	4	1	-2	0	0	-1	0	-1	-1	0
14	-2	-1	-3	-4	-3	-3	-4	-4	-2	-4	0	-1	-3	3	0	-1	-1	-2	6	1
15	0	6	-4	-4	-3	-3	-3	-3	-2	-3	0	-2	-3	-1	-2	-2	-2	-3	-2	-2
16	1	-1	-2	-2	-1	-1	-3	-2	-1	-2	1	3	-2	0	4	-1	0	0	-1	0
17	0	-2	1	-1	2	0	-2	-3	-2	-1	2	3	-2	-1	-1	-2	-1	-2	-2	-1
18	0	-3	3	1	4	1	-1	-3	-3	-1	-2	0	-1	-3	-2	-3	-3	-3	-3	-3
19	0	-1	-2	-2	-1	-1	-3	-3	-1	-2	2	3	-2	0	0	-1	3	1	-1	0
20	0	-3	3	1	4	1	-1	-3	-3	-1	-2	0	-1	-3	-2	-3	-3	-3	-3	-3
P	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
21	3	0	-2	-2	-1	-1	-3	-3	-1	-1	3	1	-2	0	0	-1	-1	-1	-1	-1

Visualização no PSSM Viewer (NCBI, 2014).

A partir da correspondência com a informação fornecida pela matriz de alinhamento, cada resíduo passa a ser representado por uma sequência de 20 números, como mostrado na Figura 16, os quais serão, de fato, utilizados como entrada para a rede neural. A proteína, entretanto, não pode ser fornecida por completo à rede neural, uma vez que isso pode causar *overfitting*. Além disso, as interações de curta distância, que ocorrem entre aminoácidos

próximos e são responsáveis pela formação da estrutura secundária, em geral não são codificadas nessa situação. Para tanto, como já proposto por Qian e Sejnowski (QIAN; SEJNOWSKI, 1988) e é uma abordagem popular entre os métodos de predição existentes, foram usadas janelas de aminoácidos como entrada para as redes neurais, nas quais a predição é feita para o aminoácidos central. O objetivo é avaliar o impacto dos resíduos vizinhos na predição da estrutura do resíduo central, de modo a codificar as interações que ocorrem nesse nível da estrutura. O tamanho escolhido para a janela é de 15 resíduos, valor também obtido empiricamente, já que, assim como ocorre com o número de nós da rede, não existe um método definido que forneça um número ótimo para esse caso. A camada de entrada, portanto, é composta por $15\,resíduos \times 20 = 300\,nós$.

Na rede neural que corresponde à fase de estrutura-estrutura, a entrada fornecida é o resultado combinado da predição das redes da fase sequência-estrutura i.e. 3-uplas que representam a predição para cada aminoácido das sequências de entrada. Assim como na primeira fase, a segunda fase tem como entrada janelas de 15 aminoácidos. Portanto, a entrada da rede da segunda fase contém $15\,resíduos \times 3 = 45\,nós$. Uma vez que a codificação de uma estrutura secundária no conjunto de teste e no resultado da saída da primeira fase é feita com base na posição que possui maior valor no vetor, duas possibilidades existem: a primeira é o uso dos valores dados como saída da primeira fase sem qualquer alteração; a segunda é a obtenção da classe e o fornecimento de um vetor que contenha 1 na posição correspondente a classe, assim como é feito no treinamento da primeira fase. A partir de experimentos, foi observado que a segunda possibilidade piorou o resultado da predição e, assim sendo, somente a primeira possibilidade foi alvo de testes e uma análise mais aprofundada.

4.3.2 Camada Intermediária

O número de nós escolhido foi 30, 45 e 60, respectivamente correspondentes a cada uma das três redes da primeira fase do preditor. Para a obtenção desse número, foram realizadas tentativas com número de nós intermediários iguais com 30 e 100 nós nas camadas intermediárias de todas as redes; bem como com conjuntos de valores distintos maiores, como 100, 125 e 150; 150, 175 e 200; e 200, 250 e 300, respectivamente. Observou-se que, na camada intermediária, números de nós entre 30 e 100, como o mesmo número de nós na

entrada e saída, produzem resultados em uma faixa de valores semelhantes[2]. Valores maiores que esses, principalmente acima de 250 nós, reduziam a acurácia da predição.

4.3.3 Camada de Saída

Na camada de saída, é utilizada uma codificação ortogonal que representa cada um dos três tipos de estrutura secundária. Desse modo, três nós são necessários e, no conjunto de teste, a classificação dada como saída obedece às seguintes correspondências:

- 100 – hélice-α
- 010 – folha-β
- 001 – *coil*

Após o treinamento, na fase de teste da rede neural, os três nós das três redes que compõem a fase sequência-estrutura do preditor são combinados através de alguns métodos distintos. A discussão sobre os resultados é feita no capítulo de resultados.

Considerando os nós a_{ij} como o nó da posição *j* no vetor de saída pertencente à rede *i*, os métodos de combinação são definidos pelas equações 4.1 e 4.2.

- Somatório

$$\sum_{i=1}^{3} (a_{i1}, a_{i2}, a_{i3}) \tag{4.9}$$

- Máximo

$$(max(a_{i1}), max(a_{i2}), max(a_{i3})) \tag{4.10}$$

A classe resultante na predição é aquela correspondente ao índice de maior valor na tripla, considerando que a ordem é (hélice-α, fita-β, *coil*).

4.4 DADOS DE TREINAMENTO

Os dados utilizados como entrada foram as proteínas do conjunto CB513, o qual é composto por 513 proteínas não-homólogas (CUFF; BARTON, 2000) . Um fator importante é o uso no treinamento e teste de diversos métodos de predição de estrutura secundária de proteínas, o que possibilita uma comparação mais precisa entre as taxas de acerto dos

---

2 As acurácias alcançadas em cada um dos casos serão foco de discussão no capítulo 5, Resultados.

métodos, uma vez que o número de proteínas, a proporção de classes de estrutura secundária, ou o tipo de treinamento realizado, como o uso de validação cruzada ou de grupos distintos de treinamento e teste, podem influenciar no resultado da predição de estrutura secundária feita por um dado método (AVDAGIC et al., 2009).

A priori um número de maior de proteínas no conjunto de treinamento pode agregar mais informação ao aprendizado do método (CHANDONIA; KARPLUS, 1996), de modo que o conjunto CB513 foi escolhido em detrimento a outros conjuntos também populares, como o CB396 ou RS126.

4.4.1 Obtenção e formatação dos dados

Toda a informação de proteínas utilizadas foram extraídas do *Protein Data Bank* (BERMAN et al., 2000). Os PDB IDs das proteínas que formam o conjunto CB513 foram obtidos em (KOUNTOURIS; HIRST, 2009).As matrizes de alinhamento correspondentes a cada uma das protéinas foram obtidas através da ferramenta PSI-BLAST, disponibilizada pelo *Nation Center for Biotechnology Information* (NCBI), a qual realiza o alinhamento múltiplo da proteína de entrada com aquelas que estão presentes na base de dados, de maneira a identificar similaridades e, como consequência, uma família de proteínas. A ferramenta foi utilizada através da biblioteca Biopython, da linguagem de programação Python, de modo a automatizar a obtenção da informação e sua formatação de acordo com a entrada adequada às redes neurais. Uma das entradas da ferramenta PSI-BLAST são as sequências das proteínas a serem submetidas ao alinhamento múltiplo, em formato FASTA, tendo sido obtidos também no *Protein Data Bank*.

4.5 EXECUÇÃO DOS TESTES

As redes neurais foram implementadas utilizando-se a biblioteca *Encog*, para a linguagem *Java*. Esta biblioteca foi selecionada pela rapidez de execução em relação a outras bibliotecas analisadas, como a *Neuroph*, por possuir uma boa documentação e demandar uma implementação simples, com classes que representam instâncias de entrada, conjuntos instâncias e redes neurais básicas, além dos algoritmos de correção de erros necessários, nomeadamente o RPROP. Além disso, uma característica que não foi encontrada em outras bibliotecas é a possiblidade de controlar o número de iterações, de modo a controlar o critério de parada. A vantagem direta desse elemento é o fato de que outras bibliotecas analisadas, como a *Neuroph*, em Java, e *pybrain*, em *Python*, necessitam de uma quantidade de tempo

muito grande para atingir a convergência, tendo sido realizados testes em que ambos executaram durante dias sem devolver resultados. Os arquivos de entrada foram resultantes da formatação feita após a obtenção dos dados, ambos na linguagem *Python*.

Os testes foram realizados em uma máquina com sistema operacional *Windows*, com com processador *Intel Core* i5, 6 GB de memória RAM e 1 TB de disco rígido. Uma vez que cada fase do preditor utiliza os resultados de três ANNs distintas, foram utilizados *threads* da linguagem *Java* de modo a paralelizar e diminuir o tempo de treinamento.

5. ANÁLISE DOS RESULTADOS

5.1 RESULTADOS ALCANÇADOS

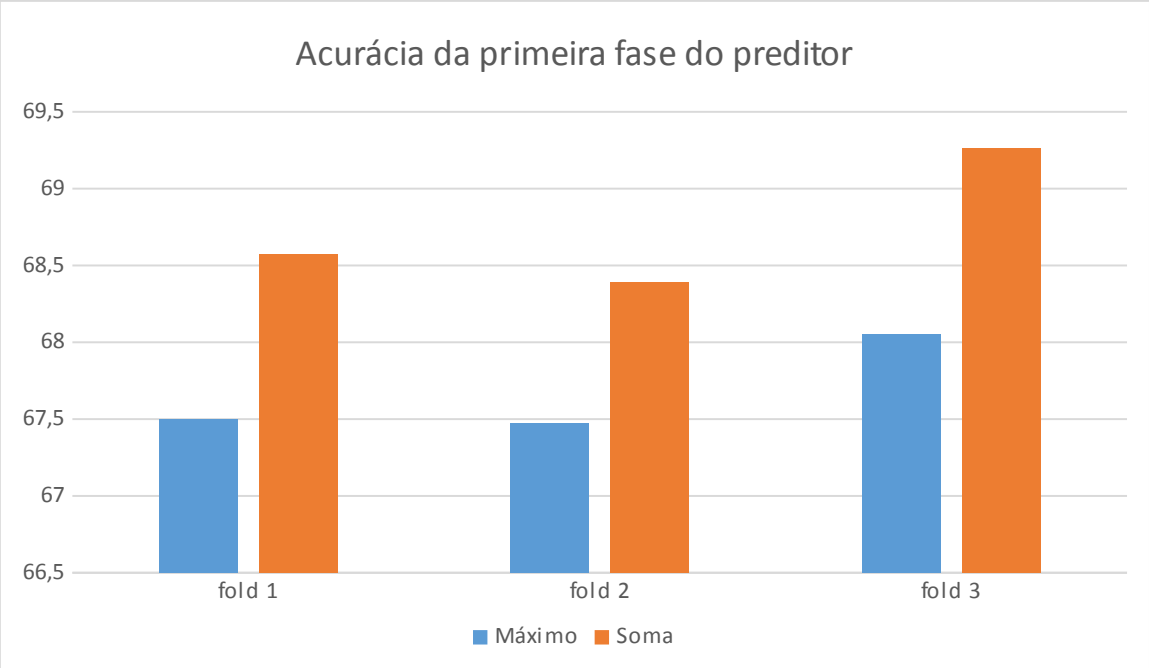
Nesta seção, serão relatados os resultados alcançados com os testes realizados com as diversas configurações e os dados enumerados no capítulo 4.

Algumas tentativas foram realizadas para obter o número ótimo de iterações do algoritmo de retropropagação. Tentativas com 100, 300, 500, 1000, 5000 e 20000 foram realizados, observando-se o erro da rede como parâmetro. De fato, cerca de 500 iterações do algoritmo alcançam um bom resultado e cerca de 700 a 800 iterações, o resultado não melhora significativamente, isto é, não melhora mais que 0,5%; e o erro da rede passa a permanecer praticamente constante, reduzindo-se em um fator da ordem de 10^{-3} a cada iteração, fator este que também diminui, chegando à ordem de 10^{-6} por volta de 10000 iterações.

Isso mostra que o treinamento exaustivo da rede neural não melhora necessariamente o resultado da predição. Dessa forma, em redes neurais com mesma estrutura (número de nós de entrada, intermediários, de saída, número de camadas intermediárias) o conjunto de dados mostra-se o único elemento capaz de melhorar do resultado da predição.

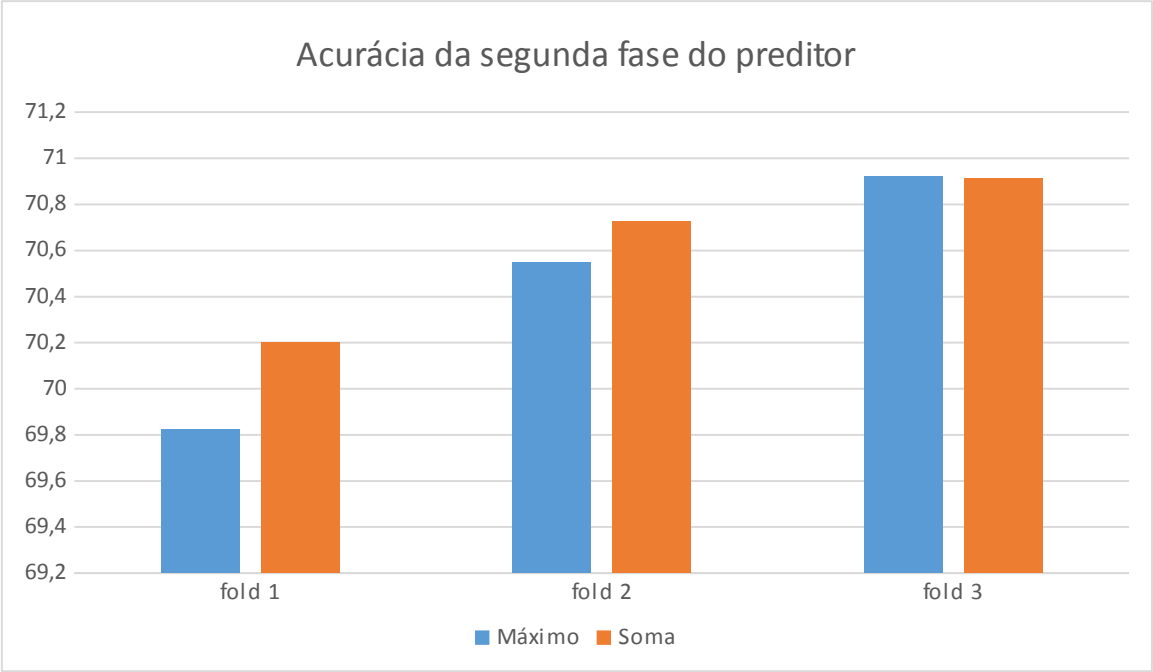
A acurácia atingida considerando-se somente a primeira fase é apresentada no Gráfico 3.

Gráfico 3 – Acurácia alcançada na primeira fase do preditor.



A acurácia atingida considerando a segunda fase do preditor é apresentada no Gráfico 4.

Gráfico 4 - Acurácia alcançada na segunda fase do preditor.



As acurácias atingidas pela primeira fase do método utilizando o máximo e a soma como métodos de combinação foram 67,67% e 68,74%, respectivamente. Tratando-se da segunda fase, as acurácias para os mesmos métodos de combinação foram 70,43% e 70,61%.

Tabela 1 – Acurácias obtidas nos testes.

Métodos de Combinação	Primeira fase	Segunda fase
Máximo	67,67%	70,39%
Soma	68,74%	70,55%
Rede Única	-	70,48%

5.2 DISCUSSÃO DOS RESULTADOS

Em comparação com os métodos mais atuais de predição de estrutura secundária com redes neurais, o resultado alcançado com preditor proposto pode ser considerado abaixo da média. Entretanto, em uma análise mais atenta, pode-se observas que os preditores mais modernos utilizam técnicas que melhoram os resultados, como o uso de funções matemáticas de pontuação que analisam a probabilidade de um dado resíduo da cadeia pertencer a uma determinada classe de estrutura secundária.

Com os resultados obtidos, pode-se observar que uma predição consensual, isto é, resultante do treinamento de diversos modelos e a combinação dos resultados, melhora a

performance da predição. Além disso, como já mostrado por Qian e Sejnowski (QIAN; SEJNOWSKI, 1988), o uso dos resultados da predição de uma rede neural como entrada em uma nova predição também melhoram o resultado final, uma vez que a segunda fase de predição, a fase estrutura-estrutura, filtra os resultados dados pela primeira fase, denominada fase sequência-estrutura. No preditor implementado neste trabalho, a melhora do resultado foi de cerca de 2% a 3%.

5.2.1 Comparacão com preditor de Huang e Chen

Embora o resultado do preditor proposto tenha sido inferior ao preditor idealizado por Huang e Chen (HUANG; CHEN, 2013), diversos fatores podem ser enumerados como motivadores da discrepância dos resultados, que é de cerca de 9%. O primeiro desses fatores são os dados de entrada. Haja vista a influência que os dados de entrada têm sobre o aprendizado é esperado que isso ocorra, uma vez que, além de fazerem uso de PSSMs, como é feito nesta proposta, utilizam uma série de outros dados que tem como objetivo agregar informação sobre como os aminoácidos de uma determinada proteína interagem de maneira a atingir sua conformação nativa.

Entretanto, o teste tem como objetivo frisar a importância desse tipo de avaliação de modo a definir a potencialidade que um método de aprendizado possui de contribuir, por exemplo, com a predição de um determinado tipo de estrutura (SHARMA; OM, 2014). Um problema que pode ser utilizado como motivação para esse tipo de testes é a predição de folhas-β, as quais são mais difíceis de predizer, dada a sua proporção em relação às outras estruturas secundárias existentes. O principal objetivo desse tipo de análise é a possibilidade do uso dos modelos de aprendizado de forma coordenada, de modo a melhorar a taxa de acerto na predição da estrutura secundária, de maneira análoga ao que ocorre com preditores que combinam resultados de outros preditores e atingem bons resultados, como o CONCORD (WEI; THOMPSON; FLOUDAS, 2011).

A associação de modelos de aprendizado tem um potencial maior que a combinação de preditores treinados, como feito pelo CONCORD, já que o treinamento de cada um dos modelos pode ser feito em separado, de modo que os pontos identificados como fortes em cada um deles, como a identificação de regiões ou estruturas específicas, podem ser potencializados por treinamentos feitos de formas distintas.

6. CONSIDERAÇÕES FINAIS

Neste trabalho foi realizada uma atualização do estado da arte apresentado em (MELO, 2005), com ênfase nos trabalhos relacionados ao uso de métodos de aprendizagem supervisionada para o problema de Predição da Estrutura Secundária de Proteínas. A acurácia foi um dos principais elementos observados no levantamento, uma vez que é uma das características determinantes na avaliação de um preditor de estrutura secundária.

A partir desse levantamento, pode-se observar que, embora o avanço alcançado nos métodos mais recentes tenha sido significativo, com acurácias que variam de 75% a 83%, ainda existe uma margem para que o limite teórico de 88% seja atingido. A importância da melhora do resultado estatístico de um preditor está relacionada ao impacto que falhas na predição podem ter em estudos posteriores.

Adicionalmente, embora tenha ocorrido avanço na predição de folhas-β, estas estruturas ainda são um desafio aos preditores, dadas a sua propoção reduzida em relação às demais estruturas secundárias e interações de longa distâncias a que estão submetidas. O preditor SCORPION (YASEEN; LI, 2014) obteve melhora na predição de folhas-β, contudo, a taxa de acerto para estas estruturas ainda é inferior à de hélices-α e *coils*.

Uma abordagem alternativa que pode ser utilizada para o problema da predição de estrutura secundária é um tratamento diferenciado a cada uma das estruturas. Uma implementação desta abordagem pode incluir o uso de diferentes modelos de aprendizado supervisionado, de modo a utilizar potencialidades de cada modelo na predição de uma estrutura específica. O objetivo é melhorar o resultado estatístico do preditor associando os pontos fortes de cada um dos modelos de aprendizado utilizados. Por esse motivo, também foi proposta uma análise comparativa entre preditores que utilizam ANNs e SVMs como modelo de aprendizado.

Foi desenvolvido um preditor de estrutura secundária, baseado no preditor implementado por Melo (MELO, 2005). A proposta feita neste trabalho foi a adição ao preditor descrito por Melo uma segunda fase, a qual recebe a predição de estrutura secundária feita na primeira fase. O objetivo desta proposta é filtrar as classes resultantes da predição feita na primeira fase, de maneira a melhorar a acurácia do preditor. Esta abordagem foi motivada pela afirmação feita por Qian e Sejnowski (QIAN; SEJNOWSKI, 1988) de que a filtragem da predição pode melhorar a taxa de acerto.

A acurácia alcançada pelo preditor desenvolvido foi de 70,9%, utilizando-se o conjunto CB513 (CUFF; BARTON, 2000) e *3-fold cross-validation*. Foram realizados diversos testes para avaliar a arquitetura que alcançou a melhor acurácia, de modo a obter um número ótimo de nós intermediários na ANNs e um número de iterações do algoritmo de atualização dos pesos, neste caso o RPROP, que maximizasse o resultado. Os valores encontrados foram 30, 45 e 60, para as ANNs da primeira e segunda fases, e cerca de 800 iterações para o RPROP. Observou-se que um número de iterações mais que 800 não melhora significativamente o resultado e um número muito alto pode reduzir a acurácia em cerca de 2%. Além disso, também constatou-se que a segunda fase do preditor melhorou a acurácia da predição realizada pela primeira fase.

Em comparação com o preditor desenvolvido por Huang e Chen (HUANG; CHEN, 2013), a acurácia foi cerca de 9% menor. Em contrapartida, Huang e Chen utilizam outras fontes de informação sobre as proteínas além das PSSMs, de modo que, para que a comparação possa ser feita de maneira mais precisa, é necessário que o treinamento de ambos seja feito com o mesmo conjunto de dados. O poder computacional disponível para a realização dos testes não era suficiente para a execução de testes com essa quantidade de informação, portanto, análises posteriores devem ser realizadas.

Como discutido, a associação de resultados provenientes de diversas predições, utilizando-se os mesmos modelos de aprendizado ou não, têm potencial para melhorar o resultado da predição de estrutura secundária. Aliada a isto, a melhora dos conjuntos de dados, de maneira a manter uma proporção de estruturas secundárias que garanta um resultado ótimo, a aplicação de funções matemáticas de pontuação, bem como a inclusão de dados que adicionem informação sobre as proteínas e sua estrutura são essenciais para um avanço na eficácia dos métodos de predição.

6.1 RECOMENDAÇÕES E TRABALHOS FUTUROS

- ✓ Associação de resultados provenientes de mais de uma predição;
- ✓ Análise das potencialidades de cada modelo de aprendizado (redes neurais, SVMs) para a predição de regiões e/ou estruturas secundárias específicas;
- ✓ Associação de resultados de diversos modelos de aprendizado de maneira a utilizar suas potencialidades de reconhecimento de regiões e/ou estruturas secundárias específicas;

- ✓ Uso de pontuações resultantes da aplicação de funções matemáticas que analisem possam facilitar inferências sobre as interações que ocorre entre os aminoácidos e foram as estruturas de mais alto nível (secundária, terciária e quaternária);
- ✓ Uso de outros algoritmos mais eficientes de atualização dos pesos da rede neural;
- ✓ Desenvolvimento de técnicas mais eficientes que possam avaliar o número de nós intermediários necessários para um resultado ótimo em um determinado conjunto de dados;
- ✓ Associação de diferentes modelos de aprendizado supervisionado para a predição de uma mesma estrutura ou de estruturas diferentes;
- ✓ Uso de abordagens mais avançadas, como aprendizado profundo e implementações distribuídas de redes neurais.

REFERÊNCIAS

ANFINSEN, C. Principles that govern the folding of protein chains. **Science**, v. 181, n. 4096, p. 223–230, 1973.

AREL, I.; ROSE, D. C.; KARNOWSKI, T. P. Deep Machine Learning - A New Frontier in Artificial Intelligence Research. **IEEE Computational Intelligence Magazine**, v. 5, n. November, p. 13–18, 2010.

AVDAGIC, Z. et al. Artificial Intelligence in Prediction of Secondary Protein Structure Using CB513 Database. **Summit on Translational Bioinformatics**, v. 2009, p. 1–5, 2009.

BERMAN, H. M. et al. The protein data bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 2000.

BETTELLA, F.; RASINSKI, D.; KNAPP, E. W. Protein secondary structure prediction with SPARROW. **Journal of chemical information and modeling**, v. 52, n. 2, p. 545–56, 27 mar. 2012.

BRANDEN, C.; TOOZE, J. **Introduction to protein structure**. 2nd. ed. New York, New York, USA: Garland Publishing, Inc., 1991.

BURGES, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. **Data Mining and Knowledge Discovery**, v. 2, p. 121–167, 1998.

CHANDONIA, J.; KARPLUS, M. The importance of larger data sets for protein secondary structure prediction with neural networks. **Protein Science**, p. 768–774, 1996.

COLE, C.; BARBER, J. D.; BARTON, G. J. The Jpred 3 secondary structure prediction server. **Nucleic acids research**, v. 36, n. Web Server issue, p. W197–201, 1 jul. 2008.

COMPIANI, M.; CAPRIOTTI, E. Computational and Theoretical Methods for Protein Folding. **Biochemistry**, p. 8601–8624, 2013.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v. 20, p. 273–297, 1995.

CUFF, J.; BARTON, G. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. **Proteins: Structure, Function, and …**, v. 511, n. March, p. 502–511, 2000.

ESWAR, N. et al. Protein Structure Modeling With MODELLER. **Methods in molecular biology**, p. 1–25, 2008.

FRISHMAN, D.; ARGOS, P. Seventy-five percent accuracy in protein secondary structure prediction. **Proteins**, v. 27, n. 3, p. 329–35, mar. 1997.

HART, W. E.; ISTRAIL, S. Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. **Journal of computational biology : a journal of computational molecular cell biology**, v. 4, n. 1, p. 1–22, jan. 1997.

HUANG, Y.; CHEN, S. Extracting Physicochemical Features to Predict Protein. **The Scientific World Journal**, v. 2013, 2013.

IKRAM, A. A. et al. Neural network based cloud computing platform for bioinformatics. **2013 IEEE Long Island Systems, Applications and Technology Conference (LISAT)**, p. 1–6, 2013.

IMMING, P.; SINNING, C.; MEYER, A. Drugs, their targets and the nature and number of drug targets. **Nature reviews. Drug discovery**, v. 5, p. 821–835, 2007.

ISMAEEL, A. G.; ABLAHAD, A. A. Novel Method for Mutational Disease Prediction using Bioinformatics Techniques and Backpropagation Algorithm. **IRACST – Engineering Science and Technology: An International Journal**, v. 3, n. 1, p. 150–156, 2013.

JONES, D. T. Protein secondary structure prediction based on position-specific scoring matrices. **Journal of molecular biology**, v. 292, n. 2, p. 195–202, 17 set. 1999.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, v. 22, n. 12, p. 2577–637, dez. 1983.

KING, R. D.; STERNBERG, M. J. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. **Protein science : a publication of the Protein Society**, v. 5, n. 11, p. 2298–310, nov. 1996.

KOHAVI, R. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**International Joint Conference on Artificial Intelligence. 1995

KOUNTOURIS, P.; HIRST, J. D. **CB513 Dataset**. Disponível em: <http://comp.chem.nottingham.ac.uk/disspred/datasets/CB513>. Acesso em: 10 out. 2014.

LEVINTHAL, C. How to fold graciously. **Mössbaun Spectroscopy in Biological Systems Proceedings**, v. 24, n. 41, p. 22–24, 1969.

LIN, H.-N. et al. Improving protein secondary structure prediction based on short subsequences with local structure similarity. **BMC genomics**, v. 11 Suppl 4, n. Suppl 4, p. S4, jan. 2010.

LORENA, A.; CARVALHO, A. DE. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

MARTIN, J.; GIBRAT, J.-F.; RODOLPHE, F. Analysis of an optimal hidden Markov model for secondary structure prediction. **BMC structural biology**, v. 6, p. 25, jan. 2006.

MCCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, 1943.

MELO, J. C. B. DE. **Análise de Estruturas de Proteínas**. [s.l.] Universidade Federal de Pernambuco, 2005.

MONTGOMERY PETTITT, B. The unsolved “solved-problem” of protein folding. **Journal of biomolecular structure & dynamics**, v. 31, n. 9, p. 1024–7, jan. 2013.

NCBI. **PSSM Viewer**. Disponível em: <http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi>. Acesso em: 20 dez. 2014.

NELSON, D. L.; COX, M. M. **Principles of Biochemistry**. 4th. ed. [s.l.] Lehninger, 2008. p. 1130

OUALI, M.; KING, R. D. Cascaded multiple classifiers for secondary structure prediction. **Protein science : a publication of the Protein Society**, v. 9, n. 6, p. 1162–76, jun. 2000.

PAVLOPOULOU, A.; MICHALOPOULOS, I. State-of-the-art bioinformatics protein structure prediction tools (Review). **International journal of molecular medicine**, v. 28, n. 3, p. 295–310, set. 2011.

POLLASTRI, G. et al. Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. **PROTEINS: Structure, Function, and Genetics**, v. 235, n. July 2001, p. 228–235, 2002.

QIAN, N.; SEJNOWSKI, T. J. Predicting the secondary structure of globular proteins using neural network models. **Journal of molecular biology**, v. 202, n. 4, p. 865–84, 20 ago. 1988.

QU, W. et al. Improving protein secondary structure prediction using a multi-modal BP method. **Computers in biology and medicine**, v. 41, n. 10, p. 946–59, out. 2011.

RCSB. **RCSB PDB - Holdings Report**. Disponível em: <http://www.rcsb.org/pdb/statistics/holdings.do>. Acesso em: 6 ago. 2014.

RIEDMILLER, M.; BRAUN, H. RPROP-A fast adaptive learning algorithm. **Proceedings of ISCIS VII**, n. September 1988, 1992.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958.

ROST, B. Review : Protein Secondary Structure Prediction Continues to Rise. **Journal of Structural Biology**, v. 218, p. 204–218, 2001.

ROST, B.; EYRICH, V. A. EVA : Large-Scale Analysis of Secondary Structure Current Implementation of EVA. **PROTEINS: Structure, Function, and Genetics**, v. 199, n. October 2001, p. 192–199, 2002.

ROST, B.; SANDER, C. Combining evolutionary information and neural networks to predict protein secondary structure. **Proteins**, v. 19, p. 55–72, 1994.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning internal representations by error propagation. In: **Parallel Distributed Processing: Explorations in the Microstructure of Cognition**. Cambridge, MA: MIT Press, 1986. v. 1p. 318–362.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3rd. ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

SARASWATHI, S. et al. Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction. **Journal of molecular modeling**, v. 18, n. 9, p. 4275–89, set. 2012.

SELA, M.; WHITE, F. H.; ANFINSEN, C. B. Reductive cleavage of disulfide bridges in ribonuclease. **Science**, v. 125, n. 3250, p. 691–692, 1957.

SETUBAL, J. C.; MEIDANIS, J. **Introduction to computational molecular biology**. Boston, MA: PWS Publishing Company, 1997.

SHARMA, N.; OM, H. Using MLP and SVM for predicting survival rate of oral cancer patients. **Network Modeling Analysis in Health Informatics and Bioinformatics**, v. 3, n. 1, p. 58, 6 maio 2014.

UNIPROT. **UniProt Release 2014_07**. Disponível em: <ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt>. Acesso em: 6 ago. 2014.

WEI, Y.; THOMPSON, J.; FLOUDAS, C. A. CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. **Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 468, n. 2139, p. 831–850, 18 nov. 2011.

WERBOS, P. J. **Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences**. [s.l.] Harvard University, 1974.

XU, Y.; XU, D. Protein Threading using PROSPECT: Design and Evaluation. **PROTEINS: Structure, Function, and Genetics**, v. 40, n. March, p. 343–354, 2000.

XU, Y.; XU, D.; LIANG, J. **Computational methods for protein structure prediction and modeling - Volume 2: Structure Prediction**. 1st. ed. Oak Ridge, TN: Springer, 2007a. p. 335

XU, Y.; XU, D.; LIANG, J. **Computational methods for protein structure prediction and modeling - Volume 1: Basic Characterization**. 1st. ed. Oak Ridge, TN: Springer, 2007b.

YASEEN, A.; LI, Y. Context-based features enhance protein secondary structure prediction accuracy. **Journal of Chemical Information and Modeling**, v. 54, n. 3, p. 992–1002, 24 mar. 2014.