

Universidade de São Paulo

Programa Interunidades de Pós-Graduação em Bioinformática

JOSÉ GERALDO DE CARVALHO PEREIRA

**REDES NEURAIS RESIDUAIS PROFUNDAS E AUTÔMATOS
CELULARES COMO MODELOS PARA PREDIÇÃO QUE
FORNECEM INFORMAÇÃO SOBRE A FORMAÇÃO DE
ESTRUTURAS SECUNDÁRIAS PROTEICAS**

Tese apresentada ao Pro-
grama Interunidades de
Pós-Graduação em Bioin-
formática para obtenção
do Título de Doutor em
Bioinformática.

Dr. Paulo Sergio Lopes Oliveira
Orientador

Campinas, Fevereiro de 2018

ABSTRACT

The process of self-organization of the protein structure is known as folding. Although we know the structure of many proteins, for a majority of them, we do not have enough understanding to describe in details how the structure is organized from its amino acid sequence. In this work, we developed two methods for secondary structure prediction using models that have the potential to provide detailed information about the prediction process. One of these models was constructed using cellular automata, a type of dynamic model where it is possible to obtain spatial and temporal information. The other model was developed using deep residual neural networks. With this model it is possible to extract spatial and probabilistic information from its multiple internal layers of convolution. The accuracy of the prediction obtained by this model was $\sim 78\%$ for residues that showed consensus in the structure assigned by the DSSP, STRIDE, KAKSI and PROSS methods. Such value is higher than that obtained by other methods which perform the prediction of secondary structures from the amino acid sequence only.

RESUMO

O processo de auto-organização da estrutura proteica a partir da cadeia de aminoácidos é conhecido como enovelamento. Apesar de conhecermos a estrutura tridimensional de muitas proteínas, para a maioria delas, não possuímos uma compreensão suficiente para descrever em detalhes como a estrutura se organiza a partir da sequência de aminoácidos. É bem conhecido que a formação de núcleos de estruturas locais, conhecida como estrutura secundária, apresenta papel fundamental no enovelamento final da proteína. Desta forma, o desenvolvimento de métodos que permitam não somente predizer a estrutura secundária adotada por um dado resíduo, mas também, a maneira como esse processo deve ocorrer ao longo do tempo é muito relevante em várias áreas da biologia estrutural. Neste trabalho, desenvolvemos dois métodos de predição de estruturas secundárias utilizando modelos com o potencial de fornecer informações mais detalhadas sobre o processo de predição. Um desses modelos foi construído utilizando autômatos celulares, um tipo de modelo dinâmico onde é possível obtermos informações espaciais e temporais. O outro modelo foi desenvolvido utilizando redes neurais residuais profundas. Com este modelo é possível extrair informações espaciais e probabilísticas de suas múltiplas camadas internas de convolução, o que

parece refletir, em algum sentido, os estados de formação da estrutura secundária durante o enovelamento. A acurácia da predição obtida por esse modelo foi de ~78% para os resíduos que apresentaram consenso na estrutura atribuída pelos métodos DSSP, STRIDE, KAKSI e PROSS. Tal acurácia, apesar de inferior à obtida pelo PSIPRED, o qual utiliza matrizes PSSM como entrada, é superior à obtida por outros métodos que realizam a predição de estruturas secundárias diretamente a partir da sequência de aminoácidos.

*We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.*

— Donald E. Knuth

AGRADECIMENTOS

Agradeço aos meus pais por todo apoio, suporte e incentivo durante todos esses anos. E por estarem sempre presentes em toda a minha vida. Amo vocês!

Agradeço ao Paulão, meu orientador, chefe e amigo, pela oportunidade de fazer parte do seu grupo de pesquisa e pela confiança demonstrada desde que tomei a decisão de voltar a Campinas.

Agradeço minha namorada linda, Ângela Saito, por todo o apoio nesses últimos meses. Por ter escutado e aturado, pacientemente, as reclamações sobre as coisas que não davam certo. E mais pacientemente ainda quando eu começava a falar sem parar sobre as coisas que finalmente estavam dando certo. É muito bom conversar com você. E estar com você! E em breve voltaremos a andar de bike...

Agradeço ao amigos do lab, Rodrigo, Heldinho e João e ao Felipe, que foi do grupo. E por fim, agradeço aos amigos, amigas e colegas do LNBio.

SUMÁRIO

I INTRODUÇÃO	1
1 INTRODUÇÃO	5
1.1 Proteínas e sua organização estrutural	5
1.2 Enovelamento de proteínas	5
1.3 Objetivo e justificativa	10
II DESENVOLVIMENTO	11
2 MÉTODOS DE ATRIBUIÇÃO DE ESTRUTURA SECUNDÁRIA	13
2.1 Introdução	13
2.1.1 DSSP	13
2.1.2 STRIDE	15
2.1.3 KAKSI	17
2.1.4 PROSS	18
2.2 Materiais e Métodos	19
2.2.1 Conjunto de dados	19
2.3 Resultados	21
2.3.1 Características das estruturas secundárias atribuídas	22
2.3.2 Características dos dissensos	23
3 MÉTODOS DE REFERÊNCIA PARA A PREDIÇÃO DE ESTRUTURA SECUNDÁRIA	29
3.1 Introdução	29
3.1.1 Redes neurais	29
3.1.2 Redes neurais para a predição a partir da sequência	30
3.1.3 Redes neurais para a predição a partir da matriz de substituição específica por posição (Position Specific Scoring Matrix (PSSM))	32
3.2 Materiais e Métodos	33
3.2.1 Avaliação do desempenho	33
3.2.2 Rede neural similar ao modelo de Holley e Karplus	35
3.2.3 Rede neural que utiliza PSSM	36
3.3 Resultados	37
3.3.1 Rede neural similar ao modelo de Holley e Karplus	37
3.3.2 Análise da predição com o PSIPRED	39
4 AUTÔMATOS CELULARES	47
4.1 Introdução	47
4.1.1 Autômatos celulares	47
4.1.2 Autômatos celulares elementares	48

4.1.3	Autômatos celulares aplicados à predição de estruturas secundárias	48
4.2	Materiais e métodos	50
4.2.1	Autômatos celulares	50
4.2.2	Otimização do conjunto de regras	53
4.3	Resultados	58
4.3.1	Seleção do modelo de autômato celular	58
4.3.2	Análise da predição de estruturas secundárias	58
4.3.3	Exemplos de evolução do autômato celular para predizer a estrutura secundária	61
5	REDES NEURAIS RESIDUAIS PROFUNDAS	65
5.1	Introdução	65
5.1.1	Redes neurais residuais	65
5.2	Materiais e métodos	67
5.2.1	Arquitetura da rede residual	67
5.2.2	Blocos	69
5.2.3	Número de blocos e canais	70
5.2.4	Treinamento	71
5.3	Resultados	71
5.3.1	Análise do treinamento	71
5.3.2	Acurácia por tipo de estrutura secundária	71
5.3.3	Acurácia em relação ao métodos de atribuição	72
5.3.4	Distribuição da acurácia	72
5.3.5	Distribuição do tamanho das estruturas secundárias preditas	72
5.3.6	Similaridade entre aminoácidos	72
5.3.7	Análise do processo de predição	73
III	CONCLUSÃO	89
6	DISCUSSÃO	91
6.1	Métodos de atribuição de estrutura secundária	91
6.2	Métodos de referência para a predição de estrutura secundária	91
6.3	Autômatos celulares para a predição de estruturas secundárias	92
6.4	Redes neurais residuais	94
6.4.1	Representação do aminoácidos	94
6.4.2	Configuração dos blocos da rede residual	94
6.4.3	Acurácia dos modelos de predição	96
7	CONCLUSÕES E PERSPECTIVAS	99
	REFERÊNCIAS BIBLIOGRÁFICAS	101

LISTA DE FIGURAS

- Figura 1.1 Classificação hierárquica da organização estrutural das proteínas. [6](#)
- Figura 2.1 Mesoestados $60^\circ \times 60^\circ$ para descrever os resíduos de acordo com os ângulos torcionais Φ e Ψ . [19](#)
- Figura 2.2 Mesoestados $30^\circ \times 30^\circ$ para descrever os resíduos de acordo com os ângulos torcionais Φ e Ψ . [20](#)
- Figura 2.3 Estrutura secundária atribuída aos resíduos das proteínas do conjunto de dados. Entre os resíduos que apresentaram consenso entre os métodos, aproximadamente 29,5% foram classificados como hélices (H); 14,5% como fitas (E); 26,5% como coils (C). 24,2% dos resíduos não apresentaram consenso entre as estruturas secundárias atribuídas pelos quatro métodos (Disenso); 5,2% são resíduos não visualizados na estrutura atômica resolvida, mas presentes na sequência da proteína analisada experimentalmente e 0,04% dos resíduos não tiveram a estrutura secundária atribuída por todos os quatro métodos (Erro*). Acima das barras estão indicados os números de resíduos. [22](#)
- Figura 2.4 Comparação pareada entre os métodos DSSP, STRIDE, KAKSI e PROSS da fração [0,1] de resíduos com a mesma estrutura secundária atribuída. É possível observar que a atribuição mais similar ocorre entre os métodos DSSP e STRIDE. [23](#)
- Figura 2.5 Número de estruturas secundárias atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS. [24](#)
- Figura 2.6 Distribuição do comprimento de hélices por método de atribuição de estrutura secundária. [25](#)
- Figura 2.7 Distribuição do comprimento de fitas por método de atribuição de estrutura secundária. [26](#)
- Figura 2.8 Distribuição do comprimento dos coils por método de atribuição de estrutura secundária. [26](#)

- Figura 2.9 Relação entre o número de dissensos e o comprimento da cadeia polipeptídica. Distribuição da fração de resíduos que não apresentaram consenso (dissensos) por proteína. A estrutura da RNA polimerase do bacteriófago N4 (PDB ID 3Q23) apresentou um grande número de dissensos. Esses dissensos foram causados pela atribuição feita pelo STRIDE, o qual classificou um grande número de resíduos de hélices como voltas (*turns*). [27](#)
- Figura 2.10 Proporção dos quatro tipos de dissensos que os resíduos podem apresentar. [27](#)
- Figura 2.11 Ocorrências de dissensos por região relativa à estrutura secundária precedente e sucedente. Ex. C?H indica uma região de dissenso imediatamente posterior a um coil e anterior a uma hélice. [28](#)
- Figura 2.12 Frequência dos aminoácidos em regiões de dissenso. A frequência foi normalizada pelo número total de ocorrências do aminoácido nas proteínas do conjunto de dados. [28](#)
- Figura 3.1 Esquema demonstrando a arquitetura diferente entre *Perceptrons* e *Multi Layer Perceptron*, a qual possui camadas ocultas de neurônios. [30](#)
- Figura 3.2 Rede neural utilizada nos trabalhos de Holley e Karplus [57] e Chandonia e Karplus [58]. A rede neural utiliza como entrada uma janela de 17 aminoácidos da proteína, cada um codificado em um vetor binário de tamanho 21 (20 aa + 1 posição que indica ausência de aminoácidos). Na camada oculta foram testadas várias configurações, diferindo entre si pelo número de neurônios. A camada de saída possuía 2 neurônios, uma representando a saída para hélice e outro para fita. Todos os neurônios possuíam funções de ativação sigmoide. Os rótulos de treinamento foram $(1, 0) \rightarrow \text{hélice}$, $(0, 1) \rightarrow \text{fita}$, $(0, 0) \rightarrow \text{coil}$. [32](#)

- Figura 3.3 Método de predição de estrutura secundária PSIPRED. Inicialmente a sequência de resíduos da proteína alvo é alinhada com um banco de dados para a construção de uma PSSM. A primeira rede neural artificial recebe como entrada janelas de 15 resíduos da PSSM e emite 3 valores entre 0 e 1 correspondentes a cada elemento de estrutura secundária. A segunda rede neural utiliza como entrada janelas de 15 resíduos contendo os valores preditos pela rede anterior e emite 3 valores correspondentes a estrutura secundária predita para o resíduo no centro da janela (Figura adaptada de [59]) 34
- Figura 3.4 A principal diferença encontra-se na camada de saída onde foram utilizados 3 neurônios, ao invés de 2, cada um representando uma estrutura secundária. Em seguida, os valores dos neurônios passam por uma função Softmax para que a somatória das saídas seja igual a 1 e assim, a saída representa a probabilidade da estrutura secundária para cada aminoácido. Os rótulos utilizados foram $(1, 0, 0) \rightarrow$ hélice, $(0, 1, 0) \rightarrow$ fita, $(0, 0, 1) \rightarrow$ coil. 36
- Figura 3.5 Treinamento das redes neurais com diferentes números de neurônios na camada oculta. O gráfico mostra o desempenho da rede nos conjuntos de treinamento e validação. 38
- Figura 3.6 Distribuição da acurácia das redes com 32 e 64 neurônios na camada oculta para as proteínas do conjunto de teste. 39
- Figura 3.7 Distribuição dos comprimentos dos elementos de estrutura secundária preditos por redes neurais similares ao modelo de Holley e Karplus. (A) Modelo com 32 neurônios na camada oculta. (B) Modelo com 64 neurônios na camada oculta. 40
- Figura 3.8 Distribuição dos valores de Q_3 obtidos pelo PSIPRED para cada proteína do conjunto. A comparação foi feita com as estruturas secundárias definidas pelos quatro métodos atribuição e com as regiões de consenso entre eles. 41
- Figura 3.9 Distribuição da acurácia obtida pelo PSIPRED na predição de hélices. 41
- Figura 3.10 Distribuição da acurácia obtida pelo PSIPRED na predição de fitas. 45

- Figura 3.11 Distribuição da acurácia obtida pelo PSIPRED na predição de coils. 45
- Figura 3.12 Distribuição dos comprimentos dos elementos de estrutura secundária preditas utilizando o PSIPRED. 46
- Figura 4.1 Espaço discreto e unidimensional dos autômatos celulares representado por um conjunto linear de células. Cada célula corresponde a uma região do espaço. 48
- Figura 4.2 Estados discretos dos Autômato Celular Elementar (**ACE**) representados com 0 e 1. Cada célula encontra-se em um estado. 48
- Figura 4.3 Exemplo de um conjunto de regras para um **ACE**. Os padrões de três células no tempo t ocasionam a mudança do estado da célula central no tempo $t+1$. A figura representa a “regra 110” dos **ACE**. Ao todo, os **ACE** possuem 2^8 (256) regras possíveis, pois há dois estados possíveis para cada um dos 8 elementos. As regras diferem entre si pelos estados $t+1$ dos elementos. 49
- Figura 4.4 Esquema da evolução de um **ACE**. Cada célula no estado t evolui para o estado $t+1$ e depois para o estado $t+2$ através da aplicação das regras da figura 4.3. 49
- Figura 4.5 Exemplos de padrões produzidos por diferentes regras dos **ACE**. 50
- Figura 4.6 Exemplo de parte dos dados da probabilidade inicial enviada pelo mestre aos escravos. Os escravos construirão as regras candidatas de acordo com essa informação. 54
- Figura 4.7 Exemplo contendo elementos de uma regra candidata produzido pelo escravo. 55
- Figura 4.8 Concatenação da sequências de proteínas do conjunto de treinamento. (A) Esquema geral da concatenação. (B) Colocação de estados de ínicio/fim da sequência para impedir a troca de informação entre proteínas. 56
- Figura 4.9 Acurácia dos modelos de Autômato Celular (**AC**) testados em um conjunto reduzido de proteínas. O **AC** 4, o qual utiliza um maior número de contextos, apresentou maior capacidade de reproduzir a estrutura secundária. 59

- Figura 4.10 Valor médio observado durante a evolução do Estimation of distribution algorithm ([EDA](#)) por 1000 gerações utilizando as métricas Class Balance Accuracy ([CBA](#)), Matthews Correlation Coeficient ([MCC](#)) e Cross Entropy ([CE](#)) como *fitness*. [60](#)
- Figura 4.11 Distribuição da acurácia (Q_3) apresentada pelo [AC](#) 4 para proteínas do conjunto de treinamento e teste. [60](#)
- Figura 4.12 Distribuição da acurácia na predição de hélices (Q_H) apresentada pelo [AC](#) 4 para proteínas do conjunto de treinamento e teste. [61](#)
- Figura 4.13 Distribuição da acurácia na predição de fitas (Q_E) apresentada pelo [AC](#) 4 para proteínas do conjunto de treinamento e teste. [61](#)
- Figura 4.14 Distribuição da acurácia na predição de coils (Q_C) apresentada pelo [AC](#) 4 para proteínas do conjunto de treinamento e teste. [62](#)
- Figura 4.15 Distribuição do comprimento de estruturas secundárias preditas para o conjunto de treinamento. Independentemente da métrica utilizada na otimização das regras, o [AC](#) 4 produziu estruturas com comprimento inferior ao observado experimentalmente para as três classes de estruturas secundárias: hélice (H), fita (E) e coil (C). [62](#)
- Figura 4.16 Evolução do AC4 com a regra otimizada por [EDA](#) utilizando [CBA](#) como função de *fitness*. O domínio PUB da proteína *Peptide-N(4)-(N-acetyl-beta-glucosaminyl) asparagine amidase* (PDB ID: 2CCQ) apresentou uma das maiores acurácia com $Q_3 = 90,23\%$, $Q_H = 84,31\%$, $Q_E = 100\%$ e $Q_C = 86,36\%$. Estados que representam hélices estão em vermelho, fitas em amarelo e coils em verde. [63](#)
- Figura 4.17 Evolução do AC4 com a regra otimizada por [EDA](#) utilizando [CBA](#) como função de *fitness*. A proteína sem função descrita (PDB ID: 2RLD) apresentou uma das piores acurácia com $Q_3 = 30\%$, $Q_H = 60,00\%$ e $Q_C = 0\%$. Estados que representam hélices estão em vermelho, fitas em amarelo e coils em verde. [63](#)

- Figura 5.1 Codificação da sequência de aminoácidos da proteína no formato *one hot encoding*. Os códigos '>' e '<' representam, respectivamente, a ausência de um aminoácido anterior e posterior a proteína. [68](#)
- Figura 5.2 Diagrama da arquitetura da rede neural residual utilizada neste trabalho. A entrada da rede consiste na sequência de resíduos de uma proteína no formato *one-hot encoding*. A saída contém três probabilidades para cada resíduo, cada uma representa a probabilidade de uma estrutura secundária: hélice, fita ou coil. Para facilitar, apenas dois blocos estão representados e $C=32$ canais (ou filtros). No trabalho foram testadas redes com 4, 11 ou 21 blocos e 8, 16 ou 32 canais. [75](#)
- Figura 5.3 Esquema de um neurônio em uma camada de convolução 3×1 . Cada neurônio utiliza os valores dos canais de 3 colunas, multiplica esses valores pelos pesos, soma, e atribui o resultado a um canal na posição correspondente. Os neurônios de um mesmo canal são idênticos, logo, o resultado irá variar de acordo com a entrada. Neurônios de canais diferentes tem pesos diferentes. [76](#)
- Figura 5.4 Aprendizado da rede residual com 4 blocos demonstrado pela redução da entropia cruzada ao longo de 1000 épocas. Os valores são para os dados do conjunto de treinamento e validação nas redes com 8, 16 e 32 canais (ou filtros). [76](#)
- Figura 5.5 Aprendizado da rede residual com 8 blocos demonstrado pela redução da entropia cruzada ao longo de 1000 épocas. Os valores são para os dados do conjunto de treinamento e validação nas redes com 8, 16 e 32 canais (ou filtros). A rede com 32 canais apresenta sinais de sobreajuste. [77](#)

Figura 5.6	Aprendizado da rede residual com 21 blocos demonstrado pela redução da entropia cruzada ao longo de 1000 épocas. Os valores são para os dados do conjunto de treinamento e validação nas redes com 8, 16 e 32 canais (ou filtros). A rede com 32 canais apresenta fortes sinais de sobreajuste com a redução do erro no conjunto de treinamento e o aumento do erro no conjunto de validação.	77
Figura 5.7	Distribuição de acurácia da predição para as proteínas do conjunto de teste.	79
Figura 5.8	Distribuição do comprimento das estruturas secundárias preditas pelas redes residuais. As hélices, fitas e coils preditos apresentam comprimento similar ao observado experimentalmente.	80
Figura 5.9	Agrupamento da representação dos aminoácidos aprendida pela primeira camada oculta da rede residual com 4 blocos.	81
Figura 5.10	Agrupamento da representação dos aminoácidos aprendida pela primeira camada oculta da rede residual com 11 blocos.	82
Figura 5.11	Agrupamento da representação dos aminoácidos aprendida pela primeira camada oculta da rede residual com 21 blocos.	83
Figura 5.12	Estados internos da rede residual durante a predição do domínio N-terminal da proteína RssB (PDB ID: 3EOD). As três primeiras linhas correspondem, respectivamente, a sequência de resíduos, a estrutura secundária predita e a estrutura secundária atribuída. Nas demais linhas temos a estrutura secundária com maior probabilidade em cada um dos 21 blocos. Hélices são representadas em vermelho, fitas em amarelo e coils em verde.	84
Figura 5.13	Domínio N-terminal da proteína RssB (PDB ID: 3EOD). As cores representam a estrutura secundária predita com maior probabilidade. Resíduos preditos como hélices são representadas em vermelho, fitas em amarelo e coils em verde.	85

Figura 5.14

Estados internos da rede residual durante a predição da proteína RmlC de *Streptococcus suis* (PDB ID: 1NXM). As três primeiras linhas correspondem, respectivamente, a sequência de resíduos, a estrutura secundária predita e a estrutura secundária atribuída. Nas demais linhas temos a estrutura secundária com maior probabilidade em cada um dos 21 blocos. Hélices são representadas em vermelho, fitas em amarelo e coils em verde. [86](#)

Figura 5.15

Proteína RmlC de *Streptococcus suis* (PDB ID: 1NXM). As cores representam a estrutura secundária predita com maior probabilidade. Resíduos preditos como hélices são representadas em vermelho, fitas em amarelo e coils em verde. Em azul é representada a outra cadeia do homodímero. [87](#)

Figura 6.1

Configurações de blocos para redes residuais. (a) Configuração original proposta por He e colaboradores [\[74\]](#), (b) Variação também proposta por He e colaboradores [\[80\]](#), (c) Bloco proposto por Zagoruyko e Komodakis que substitui as duas camadas de *batch normalization* por uma de *dropout* [\[75\]](#). [95](#)

LISTA DE TABELAS

Tabela 3.2	Redes neurais treinadas usando como dados de entrada a estrutura atribuída por diferentes métodos ou somente resíduos com consenso entre eles. A tabela mostra a influência dos dados de treinamento na acurácia durante a predição. (A tabela continua na 3.4). 42
Tabela 3.4	Continuação da tabela 3.2. 43
Tabela 3.5	Acurácia obtida na predição de hélices, fitas e coils para os conjuntos de treinamento, validação e teste em redes neurais com diversos números de neurônios na camada oculta. 44
Tabela 4.1	Número de elementos no conjunto de regras e número de regras possíveis para os quatro modelos testados. 53
Tabela 4.2	Matriz de confusão C^k para o cálculo do CBA. 56
Tabela 5.1	Acurácia apresentada pelas redes residuais em resíduos com consenso na atribuição da estrutura secundária. 78
Tabela 5.3	Acurácia (Q_3) das redes residuais em relação a estrutura secundária atribuída por diferentes métodos computacionais. 79
Tabela 6.1	Comparação de métodos de predição de estruturas secundárias. (a) Consideramos como parâmetros do autômato celular (AC4) o número de elementos no conjunto de regras. Nos demais modelos os parâmetros correspondem ao número de pesos que são otimizados durante o treinamento. (b) A acurácia da rede neural convolucional profunda (Deep Convolutional Neural Network (DCNN)) corresponde ao dado do artigo [71]. Para os demais casos, a acurácia foi calculada utilizando nossos dados e com os resíduos que apresentam consenso na atribuição da estrutura secundária. 97

LISTA DE SIGLAS E ACRÔNIMOS

AC Autômato Celular

ACE Autômato Celular Elementar

CBA Class Balance Accuracy

CE Cross Entropy

DCNN Deep Convolutional Neural Network

EDA Estimation of distribution algorithm

LTU Linear Threshold Unit

MCC Matthews Correlation Coeficient

MLP Multi Layer Perceptron

PDB Protein Data Bank

PSSM Position Specific Scoring Matrix

PReLU Parametric Rectified Linear Unit

ReLU Rectified Linear Unit

Parte I
INTRODUÇÃO

ORGANIZAÇÃO DA TESE

Este trabalho está organizado em 3 partes. No capítulo 1 fazemos uma introdução sobre a organização estrutural de proteínas e seu processo de enovelamento.

A segunda parte, nomeada de Desenvolvimento, contém 4 capítulos. A divisão em capítulos busca facilitar a compreensão de etapas importantes.

No capítulo 2, Métodos de Atribuição de Estruturas Secundárias, descrevemos a preparação dos dados utilizados ao longo do trabalho para compor os conjuntos de treinamento, validação e teste. Tais dados consistem de um conjunto de estruturas proteicas resolvidas experimentalmente e com alta qualidade, e da estrutura secundária atribuída aos seus resíduos através de métodos computacionais. Quatro métodos computacionais de atribuição foram utilizados e há uma breve descrição de como eles funcionam. Por fim, realizamos uma comparação entre os resultados das estruturas atribuídas por eles. Como neste trabalho são utilizados métodos de aprendizado de máquina (*machine learning*) e otimização, esses dados são essenciais para o treinamento dos métodos de predição e para a avaliação da acurácia.

No capítulo 3, Métodos de Referência para a Predição de Estrutura Secundária, selecionamos dois métodos de predição, publicados e com alto número de citações, para servirem de comparação aos métodos desenvolvidos neste trabalho.

Um dos métodos realiza a predição da estrutura secundária utilizando diretamente a sequência de aminoácidos da proteína, sendo, nesse aspecto, similar aos métodos propostos neste trabalho. Esse método, baseado em redes neurais artificiais, foi proposto em 1989. Na época, haviam poucas estruturas resolvidas experimentalmente para serem usadas no treinamento da rede neural. Assim, ao invés de utilizarmos a rede neural original, optamos por implementar uma rede neural similar e a treinamos utilizando o nosso conjunto de dados contendo um maior número de estruturas proteicas. Essa abordagem possibilitou uma comparação mais justa da capacidade preditiva do método.

O outro método escolhido, PSIPRED, representa o estado da arte na predição de estruturas secundárias. Assim como o método anterior, ele faz uso de redes neurais artificiais. Entretanto, o grande diferencial deste método está no uso de matrizes de substituição de aminoácidos construídas através do alinhamento de proteínas similares (**PSSM - Position Specific Scoring Matrix**). Portanto, ao invés de utilizar somente a sequência de aminoácidos de uma proteína, ele obtém,

através do alinhamento, informações como regiões mais conservadas e substituições mais frequentes em cada posição. Como essas informações tem uma relação direta com a estrutura proteica, os métodos que utilizam [PSSM](#) atingem uma acurácia maior. O PSIPRED é um método atualizado frequentemente e por isso, compararmos diretamente os resultados da predição no nosso conjunto de proteínas.

O capítulo 4, Autômatos Celulares, descreve o desenvolvimento de um novo método de predição de estruturas secundárias utilizando autômatos celulares. A busca de um conjunto de regras para os autômatos celulares, conhecido como "Problema Inverso", foi realizada utilizando um algoritmo de estimativa de distribuição (AED, em inglês *Estimation of distribution algorithm EDA*). Essa busca tem como objetivo encontrar um conjunto de regras capaz de reproduzir o padrão de elementos de estrutura secundária a partir da sequência de aminoácidos.

O capítulo 5, Redes Neurais Residuais Profundas, descreve a utilização de um método de aprendizado profundo (*Deep Learning*) como uma alternativa aos autômatos celulares.

Na última parte, que contém o capítulo 6 - Discussão - e 7 - Conclusão - discutimos os resultados obtidos ao longo do trabalho, nossas conclusões e quais as perspectivas futuras.

INTRODUÇÃO

1.1 PROTEÍNAS E SUA ORGANIZAÇÃO ESTRUTURAL

Proteínas são cadeias polipeptídicas compostas por aminoácidos unidos em uma sequência definida. Tal sequência determina todas as propriedades particulares da proteína, incluindo sua função biológica e sua estrutura tridimensional.

A estrutura proteica é comumente classificada de maneira hierárquica em estrutura primária, secundária, terciária e quaternária.

A estrutura primária corresponde a sequência linear de aminoácidos na cadeia polipeptídica. A estrutura secundária corresponde à padrões conformacionais locais os quais são formados, sobretudo, por ligações de hidrogênio. A estrutura terciária é o arranjo espacial da estrutura secundária e, consequentemente, da estrutura primária. Já a estrutura quaternária é a organização espacial contendo mais de uma estrutura terciária e, portanto, mais de uma cadeia polipeptídica (Figura 1.1) [1].

Dentre os principais padrões observados em estruturas proteicas estão as hélices α , descritas pela primeira vez por Pauling, Corey e Branson em 1951 [2]. Outro padrão estrutural importante são as folhas β descritas por Pauling e Corey também em 1951 [3].

Apesar de haverem diversos outros padrões estruturais identificados, usualmente as diferentes estruturas secundárias costumam ser agrupadas em três classes: hélices, estruturas β e "coil". Sendo as regiões de "coil" definidas na prática como "nenhuma das outras". Assim, a característica que define uma região de coil é a ausência de conformações estruturais sequencialmente repetitivas [4].

1.2 ENOVELAMENTO DE PROTEÍNAS

O problema do enovelamento de proteínas é a questão de como são formadas ou organizadas suas estruturas atômicas tridimensionais. Essa questão surgiu no final da década de 50, logo após a resolução atômica da primeira estrutura proteica por Kendrew e colaboradores [5], trabalho no qual se observou experimentalmente, segundo o próprio autor, uma complexidade maior que as antecipadas pelas teorias da época sobre estruturas proteicas. Posteriormente, Anfinsen [6] realizou experimentos que demonstraram que a ribonuclease poderia ser reversamente desnaturada/renaturada *in vitro*, e que em condições desnaturantes, tanto a estrutura quanto a função eram perdidas, no entanto, ambas eram recuperadas ao retornarem às condições fisiológicas.

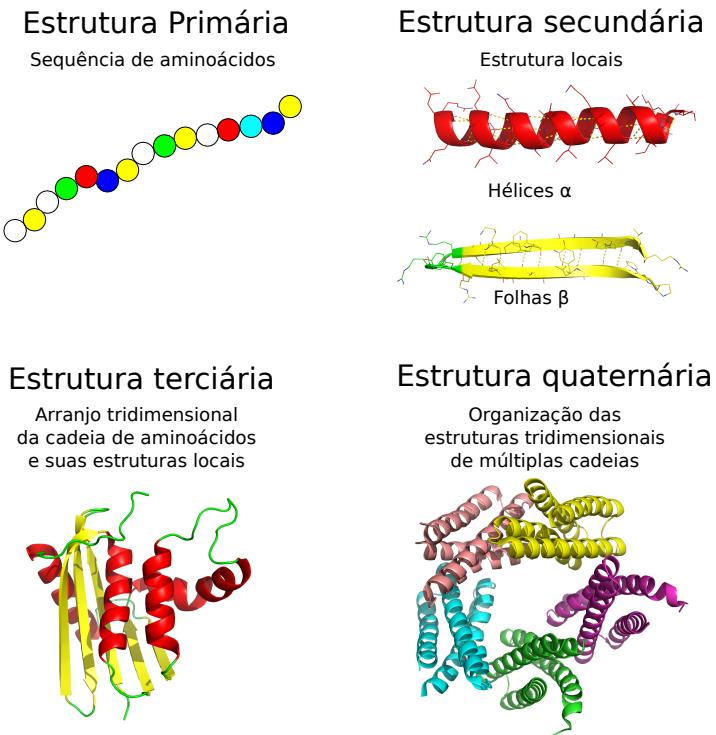


Figura 1.1: Classificação hierárquica da organização estrutural das proteínas.

gicas. Concluiu-se que, apesar da grande complexidade observada, as proteínas se auto-organizavam estruturalmente, assim sendo, apenas a informação contida em sua sequência de aminoácidos seria suficiente para definir sua estrutura e que esta determinaria a sua função. A explicação de Anfinsen para esta auto-organização estrutural foi dada através da hipótese termodinâmica, a qual postula que em condições fisiológicas a população proteica atinge um mínimo de energia livre de Gibbs no seu estado nativo [7].

Dessa forma, devido ao princípio da relação estrutura \leftrightarrow função e a resultados experimentais que demonstraram que a estrutura é determinada pela sequência de aminoácidos, diversos trabalhos buscaram prever a estrutura de uma proteína a partir da sua sequência de resíduos. Alguns dos primeiros trabalhos a discutir uma forma de predição foram publicados por Levinthal [8, 9]. Nestes, o autor menciona que o número de configurações estruturais possíveis para uma cadeia polipeptídica é imenso, sendo impossível explorar todas as conformações possíveis para se determinar qual sua estrutura nativa, ou de menor energia livre. Apesar disso, as proteínas são capazes de se enovelarem espontaneamente e adotar a conformação nativa rapidamente, numa escala de segundos ou menos [10]. Esta observação ficou popularmente conhecida como Paradoxo de Levinthal. Entretanto, Levinthal não considerou isso como um resultado absurdo, mas baseou-se nessa análise para concluir que um mecanismo aleató-

rio para o enovelamento não seria válido [11]. Segundo Levinthal [9], uma possível explicação para a eficiência observada no processo seria a formação rápida de interações locais que acelerariam e guiariam o enovelamento:

"We feel that protein folding is speeded and guided by the rapid formation of local interactions, which then determine the further folding of the peptide."

Apesar da sugestão de Levinthal para explicar um possível mecanismo de enovelamento ter sido publicado há 45 anos, o desafio de se prever as estruturas tridimensionais das proteínas a partir de suas sequências de aminoácidos, mesmo obtendo grande progresso no últimos anos, ainda permanece sem uma solução definitiva [12], sendo os métodos experimentais, mais especificamente, os métodos de cristalografia de proteínas por difração de raios-X e o de ressonância magnética nuclear, ainda as principais formas de se obter um modelo estrutural com resolução atômica. Métodos experimentais de resolução da estrutura proteica apresentam diversas dificuldades técnicas. Para a cristalografia por difração de raios-X, é necessária a obtenção de proteína em alto grau de pureza e a obtenção de monocristais, que muitas vezes é o fator limitante do processo. Por outro lado, a ressonância magnética nuclear exige concentrações bastante altas de proteína purificada de meios com diferentes isótopos e ainda, existe limitação quanto ao tamanho da proteína analisada. Essas limitações experimentais são evidenciadas pela disparidade entre o número de estruturas resolvidas experimentalmente (~127 mil depositadas no Protein Data Bank ([PDB](#))) e o número de proteínas com sequência de aminoácido conhecidas (~107 milhões depositadas no UniProtKB/-TrEMBL – dados de 01/2018). Dessa forma, a busca por métodos computacionais capazes de prever estruturas proteicas continua uma área de grande interesse científico, tanto como uma forma de se conhecer melhor o mecanismo de enovelamento como também na utilização da informação estrutural para responder diversas questões biológicas e desenvolver novos medicamentos [13].

Os métodos computacionais desenvolvidos e utilizados para construir modelos estruturais das proteínas podem ser classificados em dois tipos: (1) modelagem comparativa e (2) modelagem *de novo*, sendo considerados *de novo* todos os métodos que não utilizam informações provenientes de proteínas com estruturas similares [14]. Dentre os métodos de predição estrutural *de novo*, os que tem apresentado melhor desempenho são os que utilizam uma técnica de montagem de fragmentos (*fragment assembly*) como o I-Tasser [15] e o Rosetta [16]. Esses métodos utilizam fragmentos extraídos de proteínas com estrutura resolvida experimentalmente, os quais são posteriormente reunidos de acordo com a sequência de aminoácidos da proteína a qual se deseja construir um modelo. A utilização de fragmentos tem como

objetivo acelerar a busca pelo modelo correto, entretanto, este ainda é um método caro computacionalmente e informações como a predição da estrutura secundária e de contatos não-locais entre resíduos são comumente utilizadas para restringir o número de fragmentos a serem testados, consequentemente reduzindo o espaço de busca e acelerando um pouco mais a modelagem da estrutura [14].

A outra categoria de métodos de modelagem, a modelagem comparativa, necessita que estruturas similares à da proteína que se deseja modelar tenham sido previamente resolvidas experimentalmente. Os métodos de modelagem comparativa baseiam-se no princípio de que, em proteínas homólogas, a estrutura é mais conservada do que a sequência de aminoácidos. Assim, proteínas que tenham uma identidade entre as sequências maior que 30% e que por isso apresentam evidência de que são homólogas, podem ser modeladas caso uma delas tenha estrutura resolvida [17]. Isso não significa que proteínas com identidade sequencial menor que 30% não possam apresentar estruturas tridimensionais similares, entretanto, a identificação dessas proteínas homólogas e o alinhamento entre as sequências, ambos passos essenciais durante a modelagem comparativa, tornam-se mais suscetíveis a erros [17].

Na tentativa de contornar as dificuldades na identificação e no alinhamento de proteínas homólogas com baixa identidade sequencial (< 30%) foram desenvolvidos métodos de reconhecimento de enovelamento, os quais englobam métodos de comparação sequêncial-estrutura e métodos de alinhavamento (threading) [18]. O diferencial desses métodos em relação ao simples alinhamento entre sequências primárias está na utilização de informações estruturais, dentre elas, a estrutura secundária, a exposição ao solvente, entre outros, para descrever o ambiente em que o resíduo se encontra na proteína.

Esse ambientes alteram os padrões de substituições de aminoácidos como demonstrado por Overington e colaboradores (1990) [19]. Consequentemente, a utilização desses ambientes na construção de matrizes de substituição ou nos métodos de alinhavamento permite uma estimativa da estrutura tridimensional conhecida que melhor acomoda a sequência de aminoácidos da proteína que se deseja modelar. Outro argumento que justifica a aplicação de tal método é a existência aparente de um número finito de enovelamentos adotados pelas proteínas, o qual alguns autores estimam ser entre 1.000 e 10.000 [20, 21]. Portanto, com o aumento do número de estruturas resolvidas, acredita-se que futuramente esses enovelamentos estarão completamente representados nos bancos de dados, possibilitando a modelagem de um número cada vez maior de proteínas [22].

Os métodos de modelagem comparativa não fornecem informação sobre as etapas de enovelamento da proteína, ou mesmo o mecanismo, pois baseiam-se apenas na estrutura nativa para a construção do modelo [14]. Essas informações sobre as etapas de enovelamento

podem ser importantes para melhorar a predição de estruturas terciárias como pode ser observado no trabalho de Giri e colaboradores [23]. Neste trabalho, os autores analisaram experimentalmente o enovelamento de proteínas com topologias diferentes, mas com alta identidade entre as sequências (30%, 77% e 88%), e notaram que, as diferenças entre as topologias surgem logo no início do processo de enovelamento. Proteínas como essa, caso fossem modeladas comparativamente, provavelmente resultariam em modelos estruturais incorretos, o que a princípio poderia ser evitado com alguns métodos *de novo* como, por exemplo, os baseados em dinâmica molecular, ou por algum outro método capaz de obter informações sobre estágios intermediários do enovelamento a partir da sequência.

A hipótese termodinâmica, apesar de explicar o enovelamento, não fornece informações sobre qual o mecanismo de enovelamento adotado pelas cadeias polipeptídicas na transição entre os estados desenovelado e nativo [7]. Consequentemente, diversos mecanismos de enovelamento foram propostos [24–26]. De maneira geral, esses mecanismos podem se distinguir em dois tipos: hierárquico e não-hierárquico. No mecanismo hierárquico de enovelamento, acredita-se que o processo se inicia com estruturas que são formadas localmente na sequência e que comumente apresentem baixa estabilidade. A interação entre essas estruturas locais produziria estruturas intermediárias, com complexidade crescente e maior estabilidade até atingir a conformação nativa. Diferentemente, no mecanismo de enovelamento não-hierárquico, as interações não locais não apenas estabilizariam as estruturas locais, mas seriam as responsáveis por determiná-las [27]. Não há na literatura um consenso sobre qual tipo de mecanismo – hierárquico ou não-hierárquico – descreve com maior fidelidade o enovelamento proteico, uma vez que há evidências experimentais e de simulação computacional que, segundo a interpretação dos autores, ora sustentam um mecanismo, ora outro [28–32]. Devido a essas evidências, alguns autores propõem que, na realidade, ambos os mecanismos possam ocorrer, sendo, portanto, não apenas a estrutura proteica, mas também o processo de enovelamento, determinados pela sequência de aminoácidos [29].

Dentre as evidências que apontam para um mecanismo de enovelamento hierárquico estão estudos de proteínas como a alfa-lactoalbumina, a apo-mioglobina, a RNase H, a barnase e o citocromo c, em que análises do processo de enovelamento indicam a ocorrência de uma rápida formação de estrutura secundária semelhante a observada na proteína nativa (*native-like*) e que a mesma é estabilizada em estruturas intermediárias (*molten globules*), ou seja, antes da proteína atingir sua conformação nativa [27]. Outras evidências que sugerem a existência de um mecanismo hierárquico são os padrões na sequência de aminoácidos que ocorrem imediatamente após as extremidades N e C terminal de hélices α [27, 33–35] e fitas β [36], os quais acredita-se

que atuem como sinais de término ou “parada” desses elementos de estrutura secundária. Alguns trabalhos ainda demonstram haver preferência de algumas trincas de aminoácidos por determinadas conformações e estruturas secundárias [37, 38], sendo que algumas trincas, interessantemente, não foram observadas uma única vez em alguns tipos de estruturas secundárias das proteínas analisadas [38].

Alguns trabalhos de simulação computacional [39, 40] também identificaram que algumas sequências polipeptídicas, correspondentes a pequenas regiões de proteínas, apresentam maior propensão a adotar uma estrutura secundária, ou mesmo uma conformação similar a observada na estrutura nativa da proteína original. Posteriormente, Srinivasan e Rose [41] realizaram simulações para demonstrar que essas propensões por estruturas secundárias surgiam de impedimentos estéricos entre os átomos de resíduos consecutivos na sequência. Eles notaram ainda que a propensão correspondia a estrutura secundária de estados intermediários da proteína, sendo por vezes conservada na estrutura nativa e, em outras, alterada devido a interações não locais.

1.3 OBJETIVO E JUSTIFICATIVA

Nosso objetivo neste trabalho foi desenvolver métodos de predição de estruturas secundárias que forneçam informações sobre o processo de predição ou que nos permitam extrair tal informação. Acreditamos que uma informação detalhada do processo de predição poderá estar relacionada ao processo físico de formação da estrutura secundária e, consequentemente, ao enovelamento. Assim, tal método poderá auxiliar na predição da estrutura tridimensional tanto por métodos de modelagem comparativa quanto por métodos *de novo*.

Neste trabalho exploramos dois modelos capazes de fornecer informações do processo de predição. O primeiro modelo utiliza autômatos celulares, o qual, devido as suas características, fornece diretamente a informação do processo de predição. O segundo modelo, baseado em redes neurais residuais profundas, nos permite extrair a informação do processo de predição a partir das camadas de neurônios internas.

Ambos os métodos, uma vez otimizados, são de rápida execução e exigem poucos recursos computacionais, permitindo, assim, a aplicação em larga escala. Isto possibilitará diversos tipos de aplicações, como o estudo do impacto de mutações durante a formação da estrutura, a comparação entre proteínas homólogas, entre proteínas com sequências similares, mas estruturas distintas, além de auxiliar no *design* de novas sequências que possuam um enovelamento desejado.

Parte II
DESENVOLVIMENTO

2

MÉTODOS DE ATRIBUIÇÃO DE ESTRUTURA SECUNDÁRIA

2.1 INTRODUÇÃO

Os métodos computacionais de atribuição de estruturas secundárias surgiram com o objetivo de automatizar e reduzir a subjetividade da atribuição manual feita durante o processo de resolução experimental de estruturas. Portanto, tais métodos são diferentes dos métodos de predição. Enquanto na atribuição o objetivo é identificar os elementos de estrutura secundária a partir da estrutura trimidimensional da proteína e, portanto, é necessário que sua estrutura atômica já tenha sido resolvida, na predição o objetivo é prever os elementos de estrutura secundária a partir da sequência de aminoácidos, ou seja, quando a estrutura atômica ainda não é conhecida.

Em geral, os métodos de atribuição buscam por um conjunto de padrões estruturais, como distâncias e ângulos entre átomos, capazes de representar a estrutura secundária atribuída por especialistas. Entretanto, como a atribuição feita por diferentes especialistas pode diferir devido ao uso de padrões subjetivos e, sobretudo, pela dificuldade ao se selecionar parâmetros objetivos que reproduzam a atribuição feita por especialistas, os métodos computacionais apresentam variações entre si.

Diferentes autores sugerem a impossibilidade de se eleger qual o melhor método, uma vez que todos estão corretos segundo os princípios adotados por eles [42–45]. Assim, métodos que utilizam a informação da estrutura secundária de proteínas, como os métodos de predição de estruturas secundárias, acabaram optando, possivelmente de forma natural, por escolherem um método de atribuição como padrão. Consequentemente, podemos observar o método DSSP como um padrão de referência por ser o mais utilizado.

Nesse trabalho, nós comparamos quatro métodos de atribuição que utilizam diferentes parâmetros para classificar elementos de estrutura secundária. O objetivo disso foi entender como a atribuição varia entre os métodos e como isso afeta a acurácia dos métodos de predição de estrutura secundária.

2.1.1 DSSP

Em 1983, Kabsch e Sander publicaram o algoritmo de atribuição de estruturas secundárias de proteínas que viria a ser o mais utilizado até os dias atuais, o DSSP (*Dictionary of Protein Secondary Structure*) [42].

No trabalho, os autores afirmam que a atribuição de estruturas secundárias a partir das coordenadas atômicas de estruturas proteicas é um problema de reconhecimento de padrões. Nesse contexto, eles optaram por identificar esses padrões através de ligações de hidrogênio entre átomos da cadeia principal.

A utilização das ligações de hidrogênio da cadeia principal ao invés de ângulos Φ e Ψ ou de posições relativas de C_α foi justificada pela simplicidade. Assim, a presença ou ausência, de ligações de hidrogênio poderiam ser avaliadas por um simples critério energético, enquanto que outras características precisariam do ajuste numérico de um número maior de parâmetros.

Na época havia um pouco mais de 100 estruturas depositadas no Protein Data Bank.

O DSSP define as ligações de hidrogênio utilizando um modelo eletrostático. Nesse modelo, uma ligação de hidrogênio HB ocorrerá se, e somente se, a energia de ligação E for menor que -0.5kcal/mol . Para o cálculo são utilizadas as cargas parciais $+q_1, -q_1$ nos átomos C e O , e $-q_2, +q_2$ nos átomos N e H , onde $q_1 = 0.42e$ e $q_2 = 0.20e$.

$$E < -0.5\text{kcal/mol} \implies HB = \text{Verdade} \quad (2.1)$$

onde

$$E = q_1 q_2 (1/r(ON) + 1/r(CH) - 1/r(OH) - 1/r(CN)) * f \quad (2.2)$$

Na equação 2.2, $r(AB)$ é a distância interatômica entre A e B em Ångström e f é o fator dimensional $f = 332\text{\AA kJ/(e}^2\text{mol)}$.

Os autores afirmam que, por este modelo, uma boa ligação de hidrogênio teria aproximadamente -3kcal/mol . Assim, a escolha do limiar como -0.5kcal/mol torna o modelo mais tolerante a erros nas coordenadas atômicas e a ligações de hidrogênios bifurcadas [42].

Uma vez definido o modelo para identificar ligações de hidrogênio, elas são procuradas e anotadas na cadeia polipeptídica em duas classes ou padrões elementares: (1) padrão *n-Turn* e (2) padrão *Bridge*.

O padrão *n-Turn*, onde $n \in \{3, 4, 5\}$, apresentam uma ligação de hidrogênio entre o CO do resíduo i e o NH do resíduo $i + n$.

$$n\text{-Turn} \iff HB(i, i + n), n \in \{3, 4, 5\} \quad (2.3)$$

O padrão *Bridge* pode ocorrer de duas formas, a paralela e a anti-paralela.

$$\text{Paralela} \iff [HB(i-1, j) \wedge HB(j, i+1)] \vee [HB(j-1, i) \wedge HB(i, j+1)] \quad (2.4)$$

$$\text{Antiparalela} \iff [HB(i, j) \wedge HB(j, i)] \vee [HB(i-1, j+1) \wedge HB(j-1, i+1)] \quad (2.5)$$

sendo que as sequências $i - 1, i, i + 1$ e $j - 1, j, j + 1$ não apresentam sobreposição (*overlap*) de resíduos entre si.

As ocorrências dos padrões elementares *n-Turn* e *Bridge* ao longo da cadeia polipeptídica são utilizadas para a atribuição dos elementos de estrutura secundária. Como exemplos, repetições consecutivas do padrão *4-Turn* indicam a ocorrência de uma hélice α , enquanto resíduos consecutivos com padrão *Bridge* formam uma fita de uma folha β .

No trabalho, os autores mencionam que o algoritmo proposto produz hélices mais curtas, com um resíduo a menos em cada extremidade da hélice, em relação a anotação seguindo as regras da IUPAC. Outra característica do algoritmo é que hélices que apresentem algumas ligações de hidrogênio ausentes, são mantidas como uma hélice única ao invés de múltiplas hélices com *kinks*. O mesmo ocorre com resíduos que formam *bulges* em fitas, sendo os mesmos anotados como parte integrante da fita.

2.1.2 STRIDE

O método de atribuição de estrutura secundárias STRIDE foi desenvolvido com o objetivo de reproduzir, com o máximo de acurácia possível, a anotação manual [44]. No método são utilizados como critérios tanto as energias de ligações de hidrogênio quanto a propensão dos ângulos torcionais Φ e Ψ da cadeia polipeptídica.

A energia das ligações de hidrogênio são calculadas de acordo com a seguinte função:

$$E_{hb} = E_r * E_t * E_p \quad (2.6)$$

onde E_r (2.7) é o termo de distância, E_p (2.8) e E_t (2.9) descrevem as propriedades direcionais da ligação de hidrogênio.

$$E_r = \frac{-3E_m r_m^8}{r^8} + \frac{-4E_m r_m^6}{r^6} \quad (2.7)$$

onde r é a distância entre o nitrogênio N e o oxigênio O da cadeia principal de resíduos diferentes, $E_m = -2,8 \text{ kcal/mol}$ e $r_m = 3,0 \text{ \AA}$.

Os termos angulares são representados pelas seguintes funções:

$$E_p = \cos^2 p \quad (2.8)$$

e

$$E_t = \begin{cases} (0.9 + 0.1 \sin 2t_i) \cos t_0 & 0 < t_i \leq 90^\circ \\ K_1(K_2 - \cos^2 t_i)^3 \cos t_0 & 90^\circ < t_i \leq 110^\circ \\ 0 & t_i > 110^\circ \end{cases} \quad (2.9)$$

onde $K_1 = 0,9/\cos^6 110^\circ$, $K_2 = \cos^2 110^\circ$ e os ângulos t_i e t_0 são, respectivamente, os desvios angulares da ligação de hidrogênio.

Na equação 2.9, a colocação de ambos " \leq " foi feita por nós, uma vez que no artigo original [42] os autores utilizam somente "<".

A propensão dos ângulos torcionais para hélices α e fitas β são calculadas respectivamente como:

$$P_i^\alpha = \begin{cases} \frac{N_i^\alpha}{N_i^{total}} & se -180^\circ < \Psi < 10^\circ e -120^\circ < \Phi < 45^\circ \\ 0 & caso contrário \end{cases} \quad (2.10)$$

e

$$P_i^\beta = \begin{cases} \frac{N_i^\beta}{N_i^{total}} & se -180^\circ < \Psi < 0^\circ e -180^\circ < \Phi < -120^\circ ou 45^\circ < \Phi < 180^\circ \\ 0 & caso contrário \end{cases} \quad (2.11)$$

onde N_i^α e N_i^β são, respectivamente, os números de resíduos definidos como hélice e fita em um quadrante Φ e Ψ de dimensões $20^\circ \times 20^\circ$, e N_i^{total} é o número total de resíduos nesse quadrante.

A hélice mínima é definida por duas ligações de hidrogênio consecutivas, ou seja, entre resíduos $k, k+4$ e $k+1, k+5$, utilizando a função:

$$E_{hb}^{k,k+4} \left(1 + W_1^\alpha + W_2^\alpha \frac{P_k^\alpha + P_{k+4}^\alpha}{2} \right) < T_1^\alpha \quad (2.12)$$

Caso a condição seja verdadeira, os resíduos centrais $k+1, k+2, k+3$ e $k+4$ são classificados como hélice. Já os resíduos $k, k+5$ serão classificados como hélice se respeitarem as condições adicionais:

$$P_k^\alpha < T_2^\alpha \quad (2.13)$$

$$P_{k+5}^\alpha < T_3^\alpha \quad (2.14)$$

Nas funções 2.12, 2.13 e 2.14, W_1^α , W_2^α , T_1^α , T_2^α e T_3^α são pesos e limites empíricos que foram otimizados.

A folha β mínima é definida por duas ligações de hidrogênio consecutivas através das seguintes funções:

$$\begin{cases} E_{hb1} \left(1 + W_1^\beta + W_2^\beta \cdot CONF_{Antiparalela} \right) < T_{Antiparalela}^\beta \\ E_{hb2} \left(1 + W_1^\beta + W_2^\beta \cdot CONF_{Antiparalela} \right) < T_{Antiparalela}^\beta \end{cases} \quad (2.15)$$

$$\begin{cases} E_{hb1} \left(1 + W_1^\beta + W_2^\beta \cdot CONF_{Paralela} \right) < T_{Paralela}^\beta \\ E_{hb2} \left(1 + W_1^\beta + W_2^\beta \cdot CONF_{Paralela} \right) < T_{Paralela}^\beta \end{cases} \quad (2.16)$$

onde

$$CONF = \frac{P_{Int1}^\beta + P_{Int2}^\beta}{2} \quad (2.17)$$

$W_1^\beta, W_2^\beta, T_{Paralela}^\beta, T_{Antiparalela}^\beta$ são pesos e limites empíricos que foram otimizados.

A otimização dos pesos foi realizada de forma a aumentar a acurácia (Q_3) entre a atribuição feita pelo método e a atribuição manual. Utilizando 223 proteínas, os autores definiram os seguintes valores: $W_1^\alpha = W_2^\alpha = 1$, $T_1^\alpha = 230.0$, T_2^α e $T_3^\alpha = 0.06$, $W_1^\beta = W_2^\beta = 0.2$, $T_{Paralela}^\beta = -240.0$ e $T_{Antiparalela}^\beta = -310.0$.

2.1.3 KAKSI

KAKSI é um método de atribuição de estruturas secundárias proposto por Martin e colaboradores [46]. Esse método foi desenvolvido utilizando padrões de distâncias entre carbonos alfa (C_α) e de ângulos Φ e Ψ .

A heurística de atribuição de estrutura secundária busca primeiramente por hélices, sendo que um resíduo é classificado como hélice se respeitar os critérios de distâncias entre C_α ou os critérios de ângulos Φ e Ψ . Em seguida, é feito a classificação dos resíduos em fitas. Somente os resíduos não-hélice podem ser classificados em fitas, e para tal, eles precisam repetir os critérios de distância entre C_α e os critérios de ângulos Φ e Ψ .

1. Critérios para classificação de hélices:

a) Distância entre C_α

Todas as distâncias entre C_α em uma janela de seis resíduos $[i, i+5]$ precisam estar dentro do intervalo $[M_\alpha - \epsilon_H \times SD_\alpha, M_\alpha + \epsilon_H \times SD_\alpha]$, onde M_α e SD_α são, respectivamente, a distância média e o desvio padrão observado em hélices α .

b) Ângulos Φ e Ψ

Todos os pares de ângulos Φ e Ψ em uma janela de quatro resíduos precisam satisfazer as condições: $\Phi < 0^\circ$ e $-90^\circ < \Psi < 60^\circ$. Além disso, ao menos um par de ângulos precisa estar em uma região densamente povoada, com densidade $> \sigma_H$.

2. Critérios para classificação de folhas:

a) Distância entre C_α

Todos as distâncias entre C_α em duas janelas de três resíduos precisam estar no intervalo $[M_\beta - \epsilon_b \times SD_\beta, M_\beta + \epsilon_b \times SD_\beta]$, onde M_β e SD_β são, respectivamente, a distância média e o desvio padrão observado em folhas β .

b) Ângulos Φ e Ψ

Cada par de ângulos Φ e Ψ presente na zona povoada de resíduos em folhas β incrementa um contador em 1. Quando um resíduo central da janela apresenta $-120^\circ < \Psi < 50^\circ$ (fora da região de fitas), o contador é reiniciado em zero. Esse critério é satisfeito se o contator $\geq \sigma_b$.

Além dos critérios acima, há outros para detecção de *kink* em hélices e um critério de correção de segmentos, que altera um resíduo para o estado de coil quando ocorre continuidade de segmentos hélice-fita, ou fita-hélice, tornando as hélices 1 resíduo menor.

Os vários parâmetros necessários ao método foram ajustados empiricamente utilizando um conjunto de 2880 domínios estruturais, com identidade sequencial inferior à 40%, resolvidos por cristalografia e com resolução superior a 2.25 Å.

2.1.4 PROSS

PROSS é um método de atribuição de estruturas secundárias que utiliza somente os ângulos torcionais Φ e Ψ da cadeia principal. Originalmente, cada resíduo era classificado em um mesoestado (*mesostate*) de acordo com o valor de seus ângulos torcionais [41]. Cada mesoestado equivale a um quadrante de dimensões $60^\circ \times 60^\circ$, totalizando 36 mesoestados possíveis (Figura 2.1). Em seguida, cada resíduo é classificado em hélice α , fita β , volta β , P_{II} ou coil de acordo com um conjunto de regras. São elas:

1. Hélices α : Uma região é identificada como hélice se houverem 5 ou mais resíduos contínuos no conjunto de mesoestados {O, P}.
2. Fita β : Uma região é definida como fita se houverem 3 ou mais resíduos contínuos no conjunto de mesoestados {L, G, F, A, R, M}.
3. Volta β : São definidos como voltas β todos os pares de dipeptídeos com combinações no conjunto {OO, OP, OJ, PO, PP, PJ, JO, JP, JJ, Mo, Mp, Mj, Ro, Rp, Rj, oo, op, oj, po, pp, pj, jo, jp, jj, mO, mP, mJ, rO, rP, rJ}.
4. P_{II}: Resíduos que não foram classificados como fita, mas que estão no conjunto {M, R} são categorizados como poliprolina II.
5. Coil: Todos os outros resíduos não classificados como hélices α , fitas β , voltas β ou P_{II} são classificados como coil.

Posteriormente, outra versão do método PROSS passou a utilizar 144 mesoestados com área $30^\circ \times 30^\circ$ (Figura 2.2). As regras continuaram sendo estabelecidas de forma semelhante a feita anteriormente, mas utilizando estados mais específicos. Por exemplo:

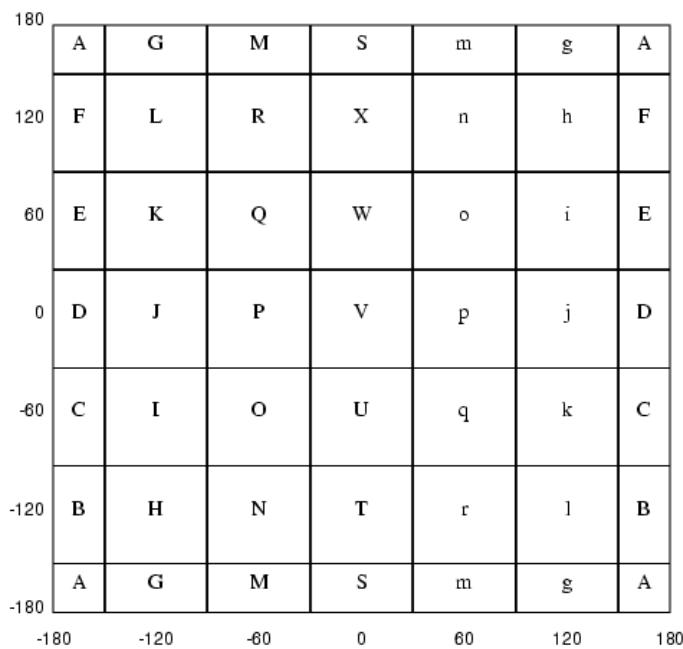


Figura 2.1: Mesoestados $60^\circ \times 60^\circ$ para descrever os resíduos de acordo com os ângulos torcionais Φ e Ψ .

1. Hélices α : Uma região é identificada como hélice se houverem 5 ou mais resíduos contínuos no conjunto de mesoestados {De, Df, Ed, Ee, Ef, Fe}.
2. Fita β : Uma região é definida como fita se houverem 3 ou mais resíduos contínuos no conjunto de mesoestados {Bj, Bk, Bl, Cj, Ck, Cl, Dj, Dk, Dl}.

Segundo os autores, a utilização somente dos ângulos torcionais da cadeia principal apresenta a vantagem de representar, corretamente, as fitas β como estruturas secundárias [41]. Métodos como o DSSP, ao utilizarem ligações de hidrogênio, acabam definindo as folhas β como estruturas secundárias, as quais, no contexto físico, seriam mais apropriadamente classificadas como estruturas terciárias.

2.2 MATERIAIS E MÉTODOS

2.2.1 Conjunto de dados

O conjunto de proteínas utilizado ao longo deste trabalho foi obtido do banco de dados "TOP8000-best-hom50" (atualizado em 2015). Esse banco de dados é organizado pelo Richardson Lab da Universidade de Duke e está disponível em github.com/rlabduke/reference_data. As estruturas proteicas do banco de dados estão separadas por cadeias e atendem aos seguintes critérios:

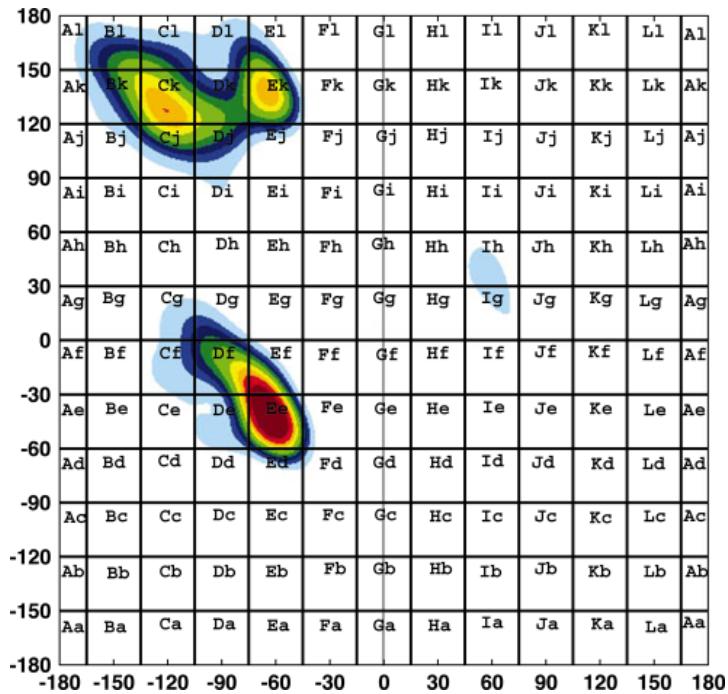


Figura 2.2: Mesoestados $30^\circ \times 30^\circ$ para descrever os resíduos de acordo com os ângulos torcionais Φ e Ψ .

- Resolução $< 2,0\text{\AA}$
- MolProbity score $< 2,0$
- $\leq 5\%$ dos resíduos apresentando comprimentos de ligação anormais ($> 4\sigma$)
- $\leq 5\%$ dos resíduos apresentando ângulos de ligação anormais ($> 4\sigma$)
- $\leq 5\%$ dos resíduos com desvios anormais do C_β ($> 0,25\text{\AA}$)
- identidade sequencial entre elas $< 50\%$ (HOM₅₀)

A restrição de identidade sequencial inferior à 50% tem o objetivo de reduzir a redundância no conjunto.

A estrutura secundária dessas proteínas foram atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS. Proteínas que possuíam resíduos indeterminados na sequência de aminoácidos ou que apresentaram algum erro de execução durante a atribuição da estrutura secundária por algum dos quatro métodos foram removidas do conjunto.

O DSSP e o STRIDE classificam os resíduos em oito categorias: H, G, I, E, S, B, T e C. Nós realizamos o reagrupamento dessas categorias em apenas três: hélice (H), fita (E) e coil (C). O reagrupamento foi o seguinte:

- Hélices (H) = H (hélice α), G (hélice 3_{10}) e I (hélice π)

- Fitas (E) = E (fita)
- Coil (C) = S (curva), B (ponte β), T (volta) e C (coil)

O PROSS classifica os resíduos em 5 categorias: H, E, T, P e C. Elas foram reagrupadas como:

- Hélices (H) = H (hélices)
- Fitas (E) = E (Fitas)
- Coil (C) = T (volta), P (P_{II}) e C (coil)

O KAKSI classifica os resíduos apenas nas três categorias (hélice, fita e coil) e, por isso, não foi necessário um reagrupamento.

O conjunto de dados final é composto por 6749 cadeias polipeptídicas das 7233 presentes no banco de dados "TOP8000-best-hom50".

Nos capítulos seguintes, descrevemos o desenvolvido de métodos de predição utilizando redes neurais artificiais e autômatos celulares. Durante o desenvolvimento dos métodos o conjunto foi dividido em três subconjuntos:

- Subconjunto de treinamento: 5000 proteínas
- Subconjunto de validação: 500 proteínas
- Subconjunto de teste: 1249 proteínas

A separação das proteínas entre os subconjuntos foi feita de maneira aleatória. A proporção aproximada foi de $3/4$ ($\sim 75\%$) dos dados no subconjunto de treinamento e $1/4$ ($\sim 25\%$) para os subconjuntos restantes.

2.3 RESULTADOS

A comparação das estruturas secundárias atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS demonstrou que há diferenças significativas entre eles. Aproximadamente 24% dos 1,7 milhão de resíduos presentes no conjunto de dados não apresentaram consenso na estrutura secundária atribuída pelos quatro métodos (Figura 2.3).

As similaridades obtidas pela comparação pareada dos métodos de atribuição indicam que, com exceção do DSSP x STRIDE, aproximadamente 85% dos resíduos apresentam consenso entre dois métodos (Figura 2.4). Considerando a comparação conjunta dos quatro métodos, a qual apresenta consenso para 74,4% dos resíduos, podemos inferir que as posições de dissenso variam entre os métodos.

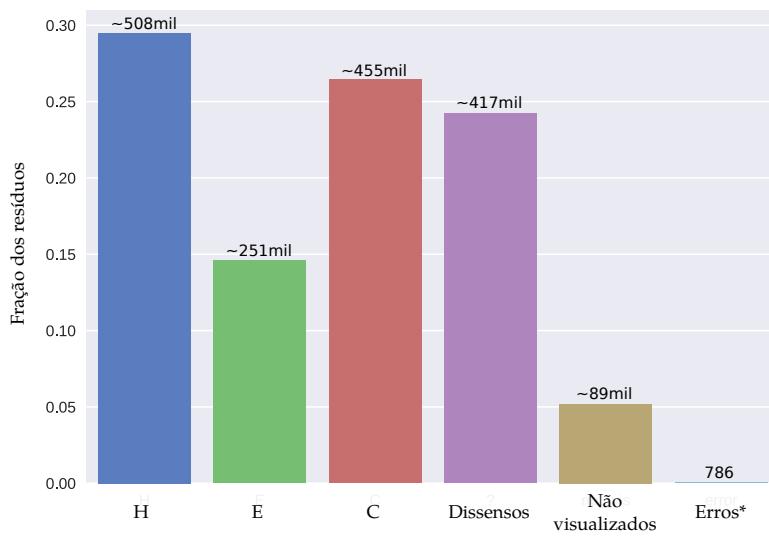


Figura 2.3: Estrutura secundária atribuída aos resíduos das proteínas do conjunto de dados. Entre os resíduos que apresentaram consenso entre os métodos, aproximadamente 29,5% foram classificados como hélices (H); 14,5% como fitas (E); 26,5% como coils (C). 24,2% dos resíduos não apresentaram consenso entre as estruturas secundárias atribuídas pelos quatro métodos (Dissenso); 5,2% são resíduos não visualizados na estrutura atômica resolvida, mas presentes na sequência da proteína analisada experimentalmente e 0,04% dos resíduos não tiveram a estrutura secundária atribuída por todos os quatro métodos (Erro*). Acima das barras estão indicados os números de resíduos.

2.3.1 Características das estruturas secundárias atribuídas

Em relação ao número de estruturas secundárias atribuídas, o DSSP atribuiu o maior número de hélices, ~67 mil. Esse valor é apenas 5,7% maior que o número atribuído pelo STRIDE, mas 31,6% maior que o KAKSI e 42% maior que o atribuído pelo PROSS (Figura 2.5). Quanto as fitas, o método que atribuiu o maior número foi o STRIDE com ~68mil. O número de fitas atribuídas por cada método apresentou uma variação menor que as hélices. O número de fitas atribuídas pelo DSSP foi apenas 0,03% menor, pelo PROSS, 0,3% menor e pelo KAKSI, 0,5% menor (Figura 2.5).

A distribuição do comprimento das estruturas secundárias também difere entre os métodos de atribuição. A mediana do comprimento das hélices atribuídas pelo DSSP é de 8 resíduos, 9 no STRIDE, 10 no PROSS e de 12 resíduos no KAKSI (Figura 2.6). Hélices mais curtas são mais frequentes tanto no DSSP quanto no STRIDE. Caso duas ou mais hélices curtas sejam unidas e anotadas como uma hélice longa por outros métodos, teríamos um menor número de hélices. Isso ex-

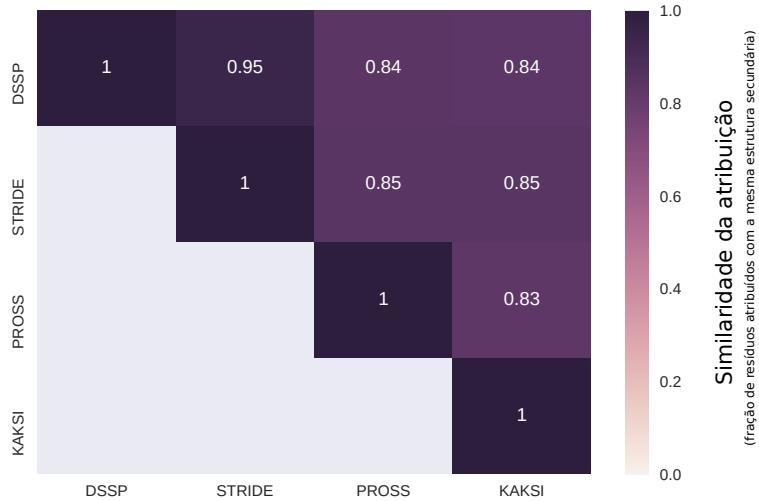


Figura 2.4: Comparação pareada entre os métodos DSSP, STRIDE, KAKSI e PROSS da fração [0,1] de resíduos com a mesma estrutura secundária atribuída. É possível observar que a atribuição mais similar ocorre entre os métodos DSSP e STRIDE.

plicaria, ao menos parcialmente, o fato dos métodos KAKSI e PROSS anotarem um menor número de hélices (Figura 2.5).

As distribuições dos comprimentos das fitas apresentam mediana de 5 resíduos para os quatro métodos de atribuição (Figura 2.7).

As distribuições dos comprimentos dos coils apresentam mediana de 4 resíduos para os quatro métodos de atribuição. O comprimento dos coils para o DSSP e o STRIDE são muito similares, o que também foi observado em fitas.

2.3.2 Características dos dissensos

O número de dissensos observados nas proteínas apresenta uma relação linear com o comprimento da cadeia polipeptídica. Assim, em média, aproximadamente 25% dos resíduos não apresentam consenso entre os métodos de atribuição analisados (Figura 2.9).

Na comparação entre as estruturas secundárias atribuídas para cada resíduo, quatro tipos de dissensos podem ocorrer. O tipo $H \leftrightarrow C$, onde o resíduo é anotado como hélice por alguns métodos e fita por outros, e da mesma forma os tipos $E \leftrightarrow C$, $H \leftrightarrow E$ e $H \leftrightarrow C \leftrightarrow E$. A análise dos dissensos demonstrou que os dois tipos dominantes são $H \leftrightarrow C$, com 48,7%, e $E \leftrightarrow C$, com 50,5%, totalizando 99,2% dos casos de dissenso. Dentre os demais tipos, $H \leftrightarrow E$ ocorre 0,1% e $H \leftrightarrow C \leftrightarrow E$ ocorre em 0,7% dos casos (Figura 2.10).



Figura 2.5: Número de estruturas secundárias atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS.

Outra característica analisada é a região relativa onde as regiões de dissenso ocorrem. Para isso, definimos 9 tipos de regiões que diferem pelo tipo de estrutura secundária que precede e sucede o dissenso. Por exemplo, C?H é um dissenso (simbolizado por "?") antecedido por um coil (C) e sucedido por uma hélice (H), o que indicaria que há variações em relação ao resíduo que inicia a hélice.

Notamos que a maioria dos dissensos ("?") ocorre em regiões de transição entre coil e as estruturas secundárias hélice ou fita. Assim, os tipos mais frequentes foram C?E, E?C, H?C e C?H (Figura 2.11). Isso indica que os métodos de atribuição tem pouca precisão em anotar onde começam e terminam as estruturas secundárias.

Os tipos E?H e H?E foram pouco frequentes, o que era esperado uma vez que a transição entre esses dois tipos de estruturas costumam ocorrer pelo intermédio de regiões de coil.

O tipo C?C também apresentou um grande número de ocorrências. Isso indica que alguns métodos de atribuição podem falhar na atribuição de regiões inteiras de estruturas secundárias, por exemplo, ignorando a presença de uma hélice inteira ou de uma fita.

A frequência relativa dos aminoácidos em regiões de dissenso mostrou que o resíduo mais comum nessas regiões é a glicina. Acreditamos que um dos motivos seja a maior variação dos ângulos Φ e Ψ que esse resíduo pode apresentar, fato que poderia dificultar a atribuição da estrutura secundária por métodos que utilizam tais ângulos, como o PROSS, ou a posição relativa de carbonos alfa, como o KAKSI.

Em relação aos aminoácidos hidrofóbicos, parece haver uma menor tendência deles ocorrerem em regiões de dissenso. É possível

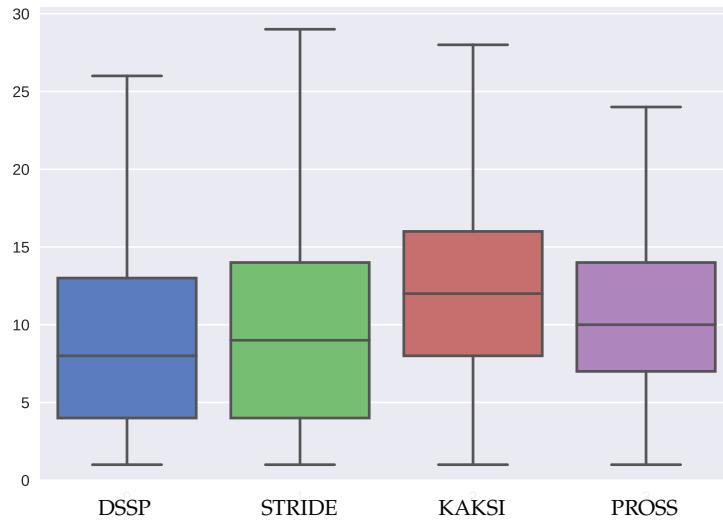


Figura 2.6: Distribuição do comprimento de hélices por método de atribuição de estrutura secundária.

que a maior frequência desses aminoácidos em regiões internas da estrutura proteica limite a variabilidade dos estados conformacionais e apresente ligações de hidrogênio mais próximas ao padrão ideal. Isso poderia explicar um maior consenso na atribuição da estrutura secundária por diferentes métodos.

Observamos ainda uma maior frequência do aspartato (D) em relação ao glutamato (E) e da asparagina (N) em relação à glutamina (Q). Essas diferenças podem estar relacionadas à maior propensão do aspartato e da asparagina em relação ao glutamato e à glutamina nos inícios e finais de hélice (*helix capping*) [47, 48] e inícios e finais de fitas (*beta-sheet capping*) [49].

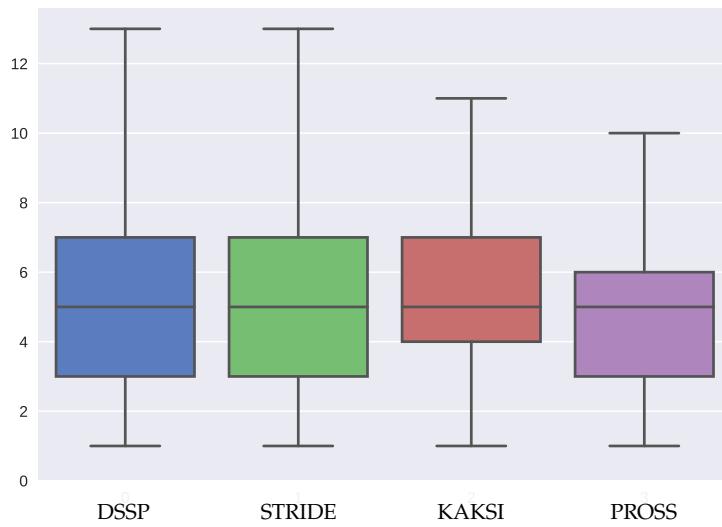


Figura 2.7: Distribuição do comprimento de fitas por método de atribuição de estrutura secundária.

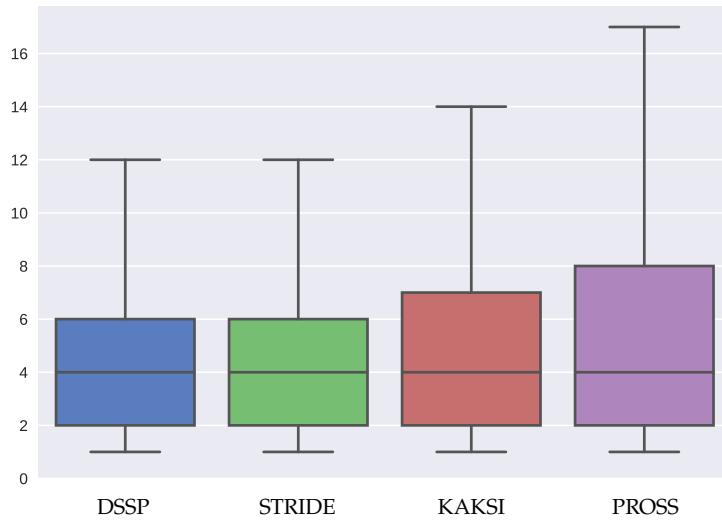


Figura 2.8: Distribuição do comprimento dos coils por método de atribuição de estrutura secundária.

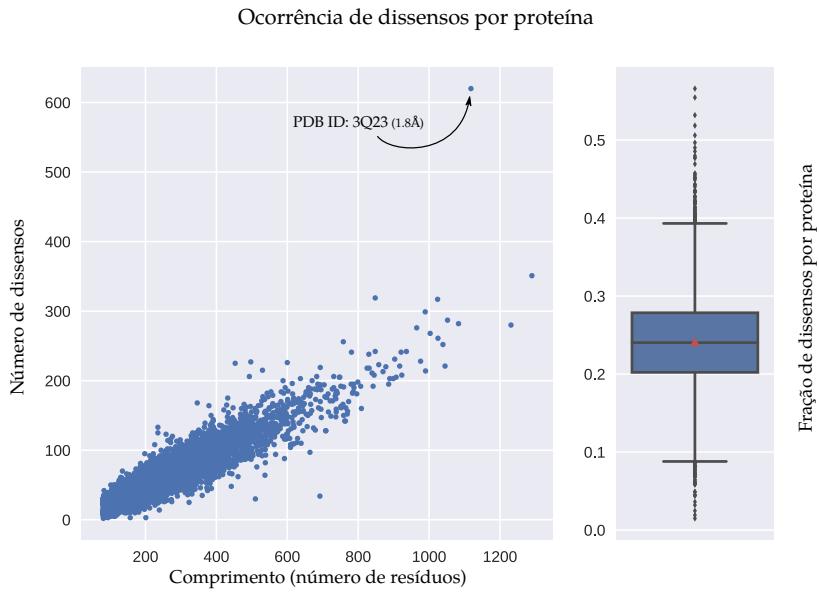


Figura 2.9: Relação entre o número de dissensos e o comprimento da cadeia polipeptídica. Distribuição da fração de resíduos que não apresentaram consenso (dissensos) por proteína. A estrutura da RNA polimerase do bacteriófago N4 (PDB ID 3Q23) apresentou um grande número de dissensos. Esses dissensos foram causados pela atribuição feita pelo STRIDE, o qual classificou um grande número de resíduos de hélices como voltas (*turns*).

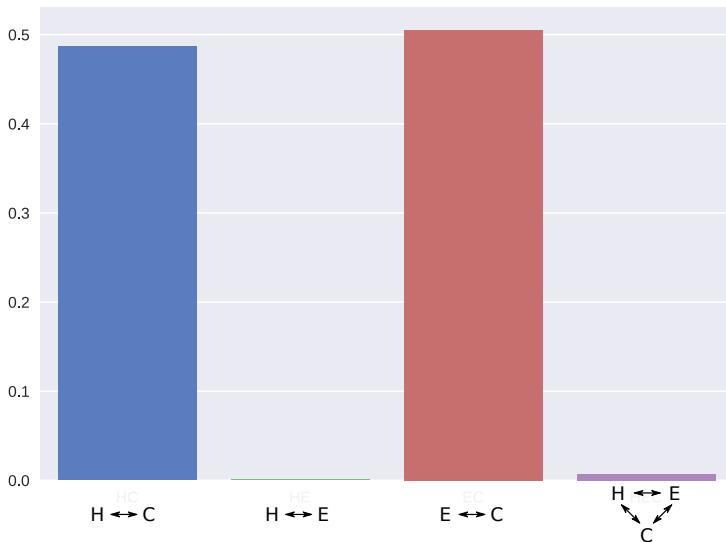


Figura 2.10: Proporção dos quatro tipos de dissensos que os resíduos podem apresentar.



Figura 2.11: Ocorrências de dissensos por região relativa à estrutura secundária precedente e sucedente. Ex. C?H indica uma região de dissenso imediatamente posterior a um coil e anterior a uma hélice.

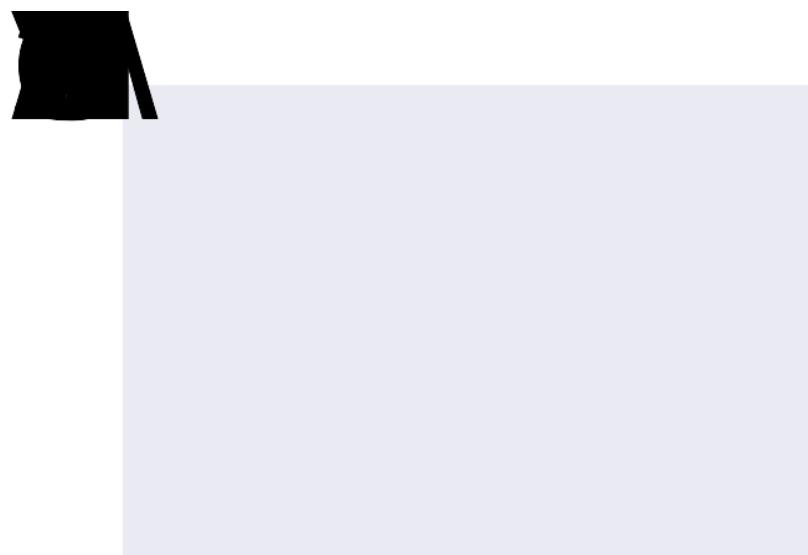


Figura 2.12: Frequência dos aminoácidos em regiões de dissenso. A frequência foi normalizada pelo número total de ocorrências do aminoácido nas proteínas do conjunto de dados.

3

MÉTODOS DE REFERÊNCIA PARA A PREDIÇÃO DE ESTRUTURA SECUNDÁRIA

3.1 INTRODUÇÃO

Neste capítulo analisamos o resultado de dois métodos que utilizam redes neurais para a predição da estrutura secundária. A diferença entre os métodos está sobretudo nos dados de entrada. Enquanto em um deles a predição é realizada a partir da sequência de aminoácidos das proteínas, no outro é realizada a partir de uma matriz de substituição específica por posição ([PSSM](#) - *Position Specific Scoring Matrix*) produzida pelo PSI-BLAST [50]. Métodos baseados em matrizes [PSSM](#) são o estado da arte para a predição de estruturas secundárias.

O método que utiliza apenas a sequência de aminoácidos como entrada foi proposto em 1989, época em que havia em torno de apenas 350 estruturas disponíveis no *Protein Data Bank* ([PDB](#)). Assim, um dos objetivos desse capítulo foi avaliar qual acurácia poderia ser atingida por um método similar utilizando um conjunto de dados maior. Isso nos fornece uma medida de acurácia mais justa para ser usada como uma linha de base na comparação com os métodos propostos neste trabalho.

O método que utiliza [PSSM](#) tem passado por atualizações frequentes, assim, ele foi apenas aplicado as proteínas do conjunto de dados para termos uma referência do estado da arte.

3.1.1 Redes neurais

As redes neurais artificiais são modelos computacionais inspirados na estrutura neural biológica e capazes de aprender a reconhecer padrões através de um processo de treinamento. Este processo de treinamento consiste em fornecer dados com padrões conhecidos para que a rede possa, através de um procedimento matemático, identificar características que sejam relevantes para a classificação. Assim, uma vez treinada, a rede neural é capaz de utilizar seu "conhecimento" previamente adquirido e aplicá-lo aos novos dados.

O primeiro neurônio artificial foi proposto na década de 40 por McCulloch e Pitts e chamado de Linear Threshold Unit ([LTU](#)) [51]. Posteriormente, Rosenblatt utilizou o [LTU](#) para criar, em 1957, o *Perceptron*, a primeira rede neural artificial capaz de aprender a reconhecer padrões [52] (Figura 3.1). Entretanto, em 1969, no livro "Perceptrons" de Minsky e Papert, os autores demonstraram a incapacidade do *Perceptron* em classificar padrões não-linearmente separáveis como o XOR

Perceptron Multi Layer Perceptron

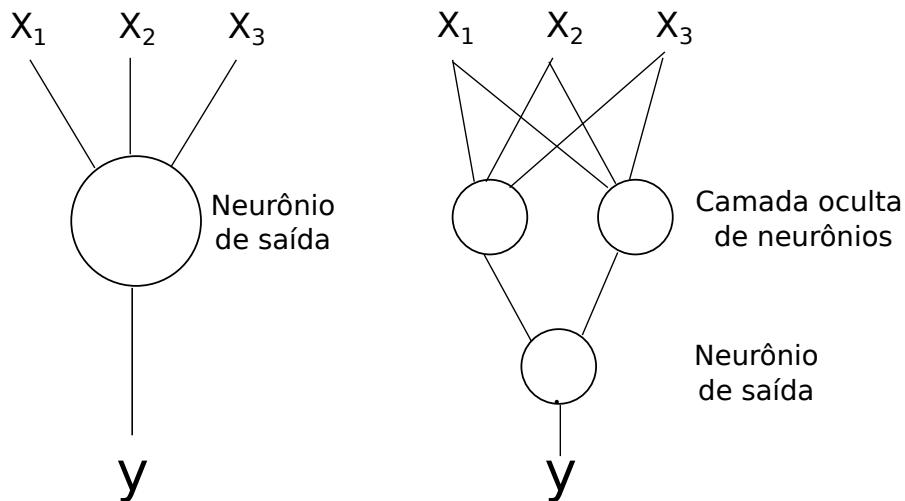


Figura 3.1: Esquema demonstrando a arquitetura diferente entre *Perceptrons* e *Multi Layer Perceptron*, a qual possui camadas ocultas de neurônios.

Apesar do nome Multi Layer Perceptron ainda ser amplamente utilizado, este nome está incorreto segundo uns dos próprios autores, Geoffrey Hinton. O erro deve-se ao fato mecanismo de treinamento do Perceptron ser diferente do backpropagation.

[53]. Apesar de Grossberg ter demonstrado modelos de redes neurais capazes de lidar com problemas não-linearmente separáveis apenas 3 anos após a publicação do livro, acredita-se que o livro de Minsky e Papert tenha atrasado o desenvolvimento das redes neurais até a década de 80, quando elas voltaram a ser estudadas.

Em 1986, Rumelhart, Hinton e Willians publicaram o trabalho onde descrevem o algoritmo de *back-propagation* [54]. Esse algoritmo tornou capaz o treinamento de redes neurais como múltiplas camadas de neurônios chamadas de Multi Layer Perceptron (MLP). As MLP utilizam camadas de neurônios ocultos, ou seja, neurônios que estão conectados a outros neurônios, e devido a esta arquitetura, são capazes de solucionar problemas de regressão não-linear e/ou classificar dados em classes não-linearmente separáveis (Figura 3.1).

3.1.2 Redes neurais para a predição a partir da sequência

Em 1988, Qian e Sejnowski [55] foram possivelmente os primeiros a publicarem um trabalho utilizando redes neurais artificiais para a predição da estrutura secundária de proteínas. Segundo os autores, a inspiração surgiu de um trabalho que aplicava redes neurais artificiais na conversão de texto em fonemas (*text-to-speech*) denominado NETtalk [56].

A melhor arquitetura testada por eles consistia de duas redes neurais em sequência. A primeira recebia como entrada 13 resíduos da sequência polipeptídica e como saída emitia 3 valores no intervalo entre 0 e 1 correspondentes aos elementos de estrutura secundária,

coil, hélice e fita. Cada resíduo foi codificado em um vetor binário com 21 elementos representando os 20 aminoácidos e um elemento indicando a ausência de aminoácidos.

A segunda rede neural recebia como entrada os valores de saída da rede anterior para uma janela de 13 resíduos e emitia novamente 3 valores no intervalo entre 0 e 1 representando a pontuação para os 3 elementos de estrutura secundária. O elemento com o maior valor correspondia a estrutura secundária predita.

Os autores exploraram ainda a utilização de uma camada oculta de neurônios, mas esta, apesar de diminuir o erro de classificação durante o treinamento, foi incapaz de reduzir o erro no conjunto de teste.

Segundo os autores, a acurácia de 64,3% (Q3) atingida pela rede neural artificial, apesar de superior aos métodos anteriores, foi decepcionante. A explicação mais plausível, segundo eles, era que a influência de resíduos mais distantes na sequência, e que portanto não estão contidos na janela de entrada, precisaria ser considerada na previsão. Caso isso fosse confirmado, eles acreditavam que seriam necessários novos métodos para considerar esses efeitos. Eles concluíram o trabalho com a hipótese de que um banco de dados maior de proteínas homólogas poderia possibilitar que uma rede neural artificial aprendesse a equivalência dos aminoácidos em diferentes contextos na proteína.

Outro modelo de rede neural aplicado à previsão de estruturas secundárias foi proposto por Holley e Karplus [57]. Este modelo também utilizou uma janela de resíduos como entrada, mas neste caso, com tamanho 17. A camada de saída possuía 2 neurônios, um indicando hélice e outro indicando fita. A previsão de um coil era definida caso nenhum dos neurônios de saída atingisse um valor maior que o limiar.

Nesta rede, os autores utilizaram uma camada de neurônios oculta, sendo a camada com dois neurônios a que demonstrou maior acurácia no conjunto de teste.

A acurácia do método foi de aproximadamente 63%. Os autores testaram ainda uma codificação de aminoácidos com características como hidrofobicidade, carga e flexibilidade da cadeia principal, entretanto, a acurácia observada nesse modelo foi de 61% e portanto inferior a codificação binária.

O método de previsão proposto por Holley e Karplus [57], e que demonstrou acurácia de 63%, foi treinado utilizando 48 estruturas proteicas resolvidas, de um conjunto de 62 estruturas selecionadas. Posteriormente, Chandonia e Karplus [58] [58] demonstraram que o aumento no número de estruturas do conjunto de treinamento poderia aumentar a acurácia da previsão utilizando redes neurais. Entretanto, como observado por eles, o aumento do conjunto de dados poderia requerer modificações na arquitetura da rede neural para pro-

Rede Neural utilizada por Holley e Karplus (1989)

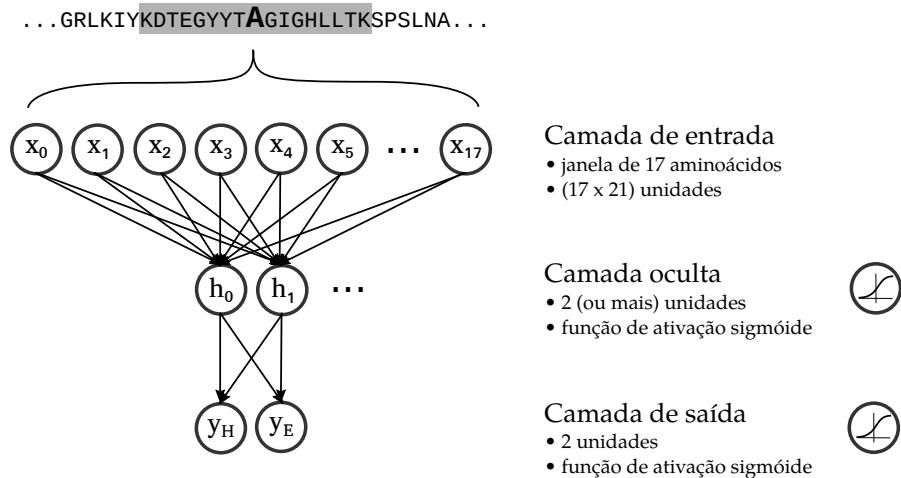


Figura 3.2: Rede neural utilizada nos trabalhos de Holley [57] e Chandonia e Karplus [58]. A rede neural utiliza como entrada uma janela de 17 aminoácidos da proteína, cada um codificado em um vetor binário de tamanho 21 (20 aa + 1 posição que indica ausência de aminoácidos). Na camada oculta foram testadas várias configurações, diferindo entre si pelo número de neurônios. A camada de saída possuía 2 neurônios, uma representando a saída para hélice e outro para fita. Todos os neurônios possuíam funções de ativação sigmoide. Os rótulos de treinamento foram $(1, 0) \rightarrow$ hélice, $(0, 1) \rightarrow$ fita, $(0, 0) \rightarrow$ coil.

duzir uma melhor acurácia. Assim, com um conjunto de 318 estruturas proteicas e aumentando o número de neurônios da camada oculta de 2 para 8, eles conseguiram uma acurácia de 67%.

O modelo de Holley e Karplus [57], por ser construído com uma arquitetura mais convencional que utiliza uma rede neural com uma camada oculta, ao invés de duas redes em sequência como a de Qian e Sejnowski [55], foi o método escolhido para ser testado.

3.1.3 Redes neurais para a predição a partir da matriz de substituição específica por posição (PSSM)

Em 1999, Jones [59] publicou o método de predição PSIPRED. O principal diferencial desse método em relação ao método de Holley e Karplus [57] foi a utilização de matrizes de substituição específica por posição (*Position Specific Scoring Matrix - PSSM*) ou *profile*.

A PSSM é produzida através do método de alinhamento PSI-Blast (*Position-Specific Iterative Blast*) [50]. No PSI-Blast, o primeiro alinhamento entre a sequência da proteína alvo e as proteínas de um banco de dados é feito utilizando uma matriz de substituição de aminoácidos como por exemplo a BLOSUM62 [60]. Na etapa seguinte, as proteínas com maior similaridade nesse primeiro alinhamento são

utilizadas para a construção de uma nova matriz de substituição. No entanto, essa segunda matriz tem probabilidades específicas para os aminoácidos em cada posição da sequência da proteína alvo, ou seja, essa matriz é uma **PSSM**. A **PSSM** é então utilizada para um segundo alinhamento e os resultados são utilizados para atualizar as probabilidades da **PSSM**. Esse processo é iterativo e ocorre por um número definido de passos ou até os resultados convergirem, isto é, quando não forem encontradas novas proteínas nos resultados do alinhamento com o banco de dados [50].

A **PSSM** construída no alinhamento representa um perfil (*profile*) dos resíduos para uma família de proteínas homólogas. Esse perfil contém informações sobre regiões mais conservadas e regiões que apresentam maior variabilidade dos resíduos, assim como locais onde ocorreram inserções e deleções.

Como a conservação de resíduos na proteína tem relação com a estrutura, por exemplo, resíduos expostos ao solvente sofrem menor pressão evolutiva que resíduos não expostos. Assim, por conter indiretamente a informação do ambiente de cada resíduo, a **PSSM** contém mais informação que a simples sequência de resíduos.

O PSIPRED é composto de duas redes neurais artificiais. A primeira rede neural recebe uma janela correspondente a 15 resíduos na **PSSM**. Essa rede neural possui 75 neurônios na camada oculta e 3 neurônios de saída. Cada neurônio de saída representa um tipo de estrutura secundária, coil, hélice ou fita, e emitem valores entre 0 e 1.

A segunda rede neural também utiliza uma janela referente a 15 resíduos com 3 valores para cada resíduo. Estes valores correspondem aos emitidos pela primeira rede neural. Esta segunda rede neural possui 60 neurônios na camada oculta e 3 neurônios na camada de saída. Assim como na primeira rede neural, cada neurônio de saída corresponde a um elemento de estrutura secundária. O neurônio da camada de saída que apresentar o maior valor indica a estrutura predita para o resíduo do centro da janela. Portanto, para a predição da estrutura secundária de uma proteína é necessário aplicar a primeira rede neural a todas as janelas possíveis e então, aplicar a segunda rede neural também a todas as janelas possíveis contendo os resultados da primeira rede neural.

O PSIPRED apresentou acurácia de 78,3% (Q_3) no conjunto de teste com 187 proteínas. Essa medida foi calculada comparando a estrutura secundária predita com a atribuída pelo DSSP.

3.2 MATERIAIS E MÉTODOS

3.2.1 Avaliação do desempenho

A avaliação da acurácia da rede foi medida utilizando a métrica Q_3 . O Q_3 é a média dos valores de Q_i onde i pode ser hélice (H), fita (E)

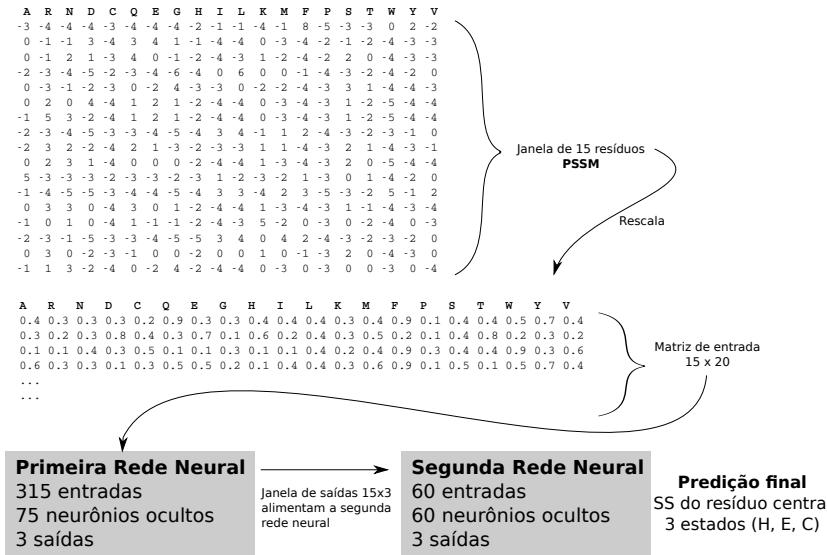


Figura 3.3: Método de predição de estrutura secundária PSIPRED. Inicialmente a sequência de resíduos da proteína alvo é alinhada com um banco de dados para a contrução de uma **PSSM**. A primeira rede neural artificial recebe como entrada janelas de 15 resíduos da **PSSM** e emite 3 valores entre 0 e 1 correspondentes a cada elemento de estrutura secundária. A segunda rede neural utiliza como entrada janelas de 15 resíduos contendo os valores preditos pela rede anterior e emite 3 valores correspondentes a estrutura secundária predita para o resíduo no centro da janela (Figura adaptada de [59])

ou coil (C) (Equação 3.1). O Q_i é definido como a porcentagem de resíduos corretamente preditos no estado i em relação ao número total de resíduos observados experimentalmente nesse mesmo estado.

$$Q_3 = \frac{Q_H + Q_E + Q_C}{3} \quad (3.1)$$

e

$$Q_i = \frac{VP_i}{Total_i} \quad (3.2)$$

onde i pode ser hélice (H), fita (E) ou coil (C), VP_i é o número de resíduos corretamente preditos para o estado i e $Total_i$ é o número de resíduos observados experimentalmente no estado i .

3.2.2 Rede neural similar ao modelo de Holley e Karplus

A rede neural artificial implementada foi do tipo *feed-forward* com uma camada de neurônios oculta. A camada de entrada recebe 17 resíduos, cada um codificado em um vetor binário (*one-hot encoding*) com 22 posições. Cada uma das 22 posições representam um aminoácido ou uma posição "vazia" N ou C terminal. A representação dessas posições vazias é necessária para a predição dos primeiros e últimos oito resíduos.

A camada oculta possui neurônios com funções de ativação sigmoide (Eq. 3.3). Variamos o número de neurônios na camada oculta de 2 à 128 com o objetivo de analisar a acurácia máxima que conseguíamos obter.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3.3)$$

Na camada de saída optamos por utilizar 3 neurônios ao invés de 2 como no trabalho original de Holley e Karplus [57]. Acrescentamos ainda uma função Softmax, que normaliza os valores entre 0 e 1 de forma que a somatória seja 1 (Eq. 3.4). Assim, os valores de saída da rede neural representam as probabilidades dos elementos de estrutura secundária (hélice, fita e coil) do resíduo central da janela de 17 aminoácidos.

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=i}^N e^{x_j}} \quad (3.4)$$

Essa alteração permitiu utilizarmos a entropia cruzada (Eq. 3.5) como função de custo, ao invés da função de erro quadrático médio (*Mean squared error* - MSE) utilizado no trabalho original.

$$H(p, q) = - \sum_i p_i \log q_i \quad (3.5)$$

Na equação da entropia cruzada (Eq. 3.5) i são as três classes (hélice, fita e coil) para cada resíduo, p é a probabilidade observada, a qual será 1 para a classe atribuída ao resíduo, e q é a probabilidade predita pela rede neural para a classe.

O conjunto de dados contendo informação da estrutura secundária de 6749 proteínas foi dividido em 3 subconjuntos: (1) treinamento com 5000 proteínas, (2) validação com 500 proteínas e (3) teste com 1249 proteínas.

Além do número de neurônios na camada oculta, foram testadas a predição utilizando como referência para o aprendizado a estrutura secundária atribuída pelos métodos DSSP, STRIDE, KAKSI e PROSS e também o consenso entre elas. No caso que utiliza o consenso, resíduos com estruturas secundárias diferentes não são utilizados no aprendizado.

Cada resíduo foi rotulado como $(1, 0, 0) \rightarrow \text{hélice}$, $(0, 1, 0) \rightarrow \text{fita}$ ou $(0, 0, 1) \rightarrow \text{coil}$.

Rede Neural HK modificada

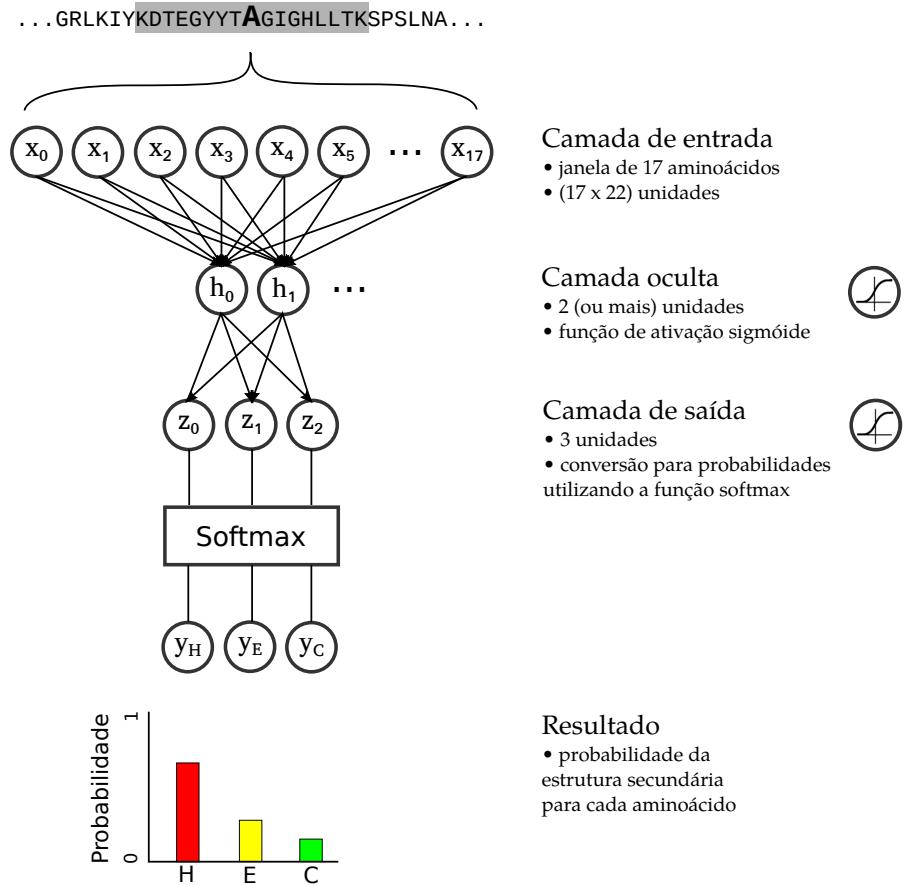


Figura 3.4: A principal diferença encontra-se na camada de saída onde foram utilizados 3 neurônios, ao invés de 2, cada um representando uma estrutura secundária. Em seguida, os valores dos neurônios passam por uma função Softmax para que a somatória das saídas seja igual a 1 e assim, a saída representa a probabilidade da estrutura secundária para cada aminoácido. Os rótulos utilizados foram $(1, 0, 0) \rightarrow$ hélice, $(0, 1, 0) \rightarrow$ fita, $(0, 0, 1) \rightarrow$ coil.

O algoritmo de otimização utilizado no treinamento foi o Adam [61] com taxa de aprendizado de 0.001.

3.2.3 Rede neural que utiliza PSSM

O método de predição PSIPRED [59] foi escolhido para testar o estado da arte em predição de estruturas secundárias. A PSSM de cada proteína foi gerada localmente com o PSI-Blast utilizando o banco de dados de sequências não redundantes (*nr*) [50].

3.3 RESULTADOS

3.3.1 Rede neural similar ao modelo de Holley e Karplus

3.3.1.1 Influência da atribuição na predição

Os resultados demonstraram que os resíduos que apresentam consenso entre os métodos de atribuição de estrutura secundária são preditos com maior acurácia independentemente do método de atribuição utilizado durante o treinamento (Consenso, DSSP, STRIDE, KAKSI ou PROSS) (Tabelas 3.2 e 3.4)

Na rede com maior acurácia, com 64 neurônios na camada oculta e treinada apenas com os resíduos que apresentaram consenso, a acurácia no conjunto de teste para os dados de consenso foi de 75,27%, enquanto a acurácia em relação aos métodos de atribuição individuais foi de aproximadamente 68%. Os valores da tabela abaixo correspondem a média de cinco treinamentos para cada rede. O desvio padrão para todos os casos foi inferior a 0,006.

Considerando que 74,4% dos resíduos apresentam consenso entre os métodos de atribuição de estrutura secundária (ver seção 2.3) e uma acurácia na predição próxima 75% para os mesmos, é possível estimar a acurácia média para as regiões de dissenso como próxima a 48%.

$$0,744 * 0,75 + 0,256 * Q_3dissenso \simeq 0,68$$

$$Q_3dissenso \simeq 0,476$$

É interessante notar que a predição utilizando somente os dados em que há consenso apresentou maior acurácia para o método de predição PROSS do que para os demais métodos de atribuição. Entretanto, quando a rede neural é treinada com os dados de atribuição do PROSS, a rede neural apresenta a menor acurácia para os resíduos com consenso.

3.3.1.2 Análise do treinamento

Analisamos a influência do número de neurônios na camada oculta durante o treinamento por 1000 épocas. Cada época corresponde à utilização de todos os dados do conjunto de treinamento, ou seja, o conjunto é apresentado à rede neural 1000 vezes para que os pesos, ou parâmetros, sejam otimizados buscando a redução da entropia cruzada.

Durante o treinamento, o conjunto de validação também é apresentado à rede para o cálculo a entropia cruzada. No entanto, ao contrário do conjunto de treinamento, o conjunto de validação não é utilizado para a otimização dos pesos. A comparação da entropia cruzada (Equação 3.5 do conjunto de treinamento com a do conjunto de

validação permite analisarmos se há a ocorrência de sobreajuste (*overfitting*) da rede neural para os dados de treinamento. Um sobreajuste indica que a rede neural está memorizando os dados de treinamento ao invés de aprender um padrão geral para ser aplicado à novos casos. Em outras palavras, quando ocorre o sobreajuste a rede neural tem uma redução da capacidade de generalização.

Na figura 3.5 é possível observar que a rede com 128 neurônios na camada oculta apresenta uma maior diferença entre a entropia cruzada no conjunto de treinamento e o conjunto de validação. Notamos ainda que a entropia cruzada próxima a época 1000 apresenta uma tendência de continuar diminuindo e, consequentemente, uma tendência em aumentar a diferença da entropia cruzada entre os conjuntos. Esse fato, juntamente com a semelhança entre os valores de entropia cruzada para o conjunto de validação nas redes com 32, 64 e 128 neurônios, são fortes indicativos da ocorrência de sobreajuste na rede com 128 neurônios na camada oculta.

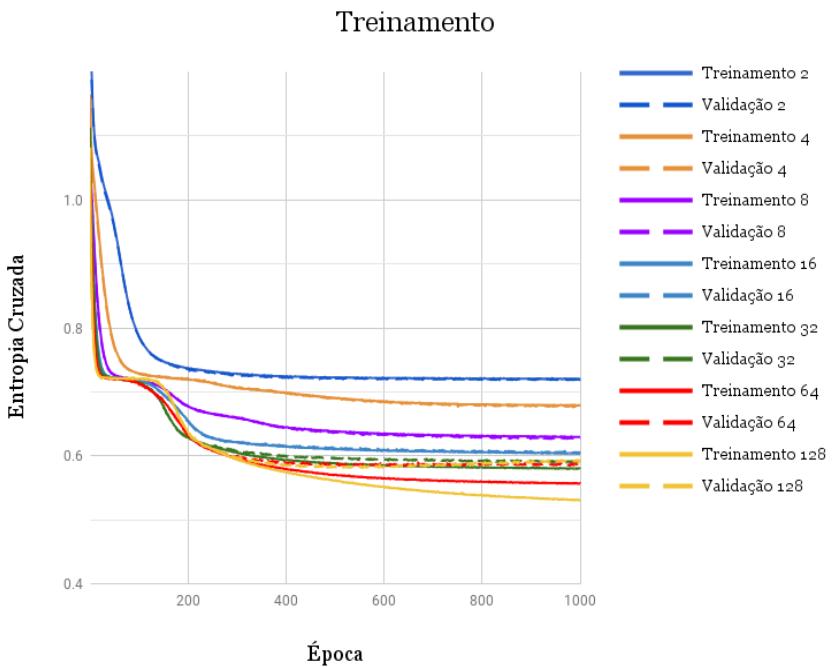


Figura 3.5: Treinamento das redes neurais com diferentes números de neurônios na camada oculta. O gráfico mostra o desempenho da rede nos conjuntos de treinamento e validação.

3.3.1.3 Acurácia da predição por tipos de estrutura secundária

A rede neural similar ao modelo de Holley e Karplus apresentou uma maior acurácia para hélices (~ 80%), seguido de coils (~ 75%) e fitas β (~ 60%). Acreditamos que uma das causas desta menor acurácia seja a dependência de resíduos distantes na formação de folhas β , fator

que não é considerado num método que utiliza janelas de entrada. Outra possível causa da menor acurácia é o menor número, nos conjuntos de dados, de resíduos que compõem as fitas β em relação à hélices e coils (Figura 2.3).

3.3.1.4 Distribuição da acurácia por proteína

A distribuição da acurácia para as proteínas do conjunto de teste é similar nas duas melhores redes neurais treinadas, as com 32 e 64 neurônios na camada oculta 3.6. A acurácia das redes para a predição de fitas (Q_E) apresentou uma distribuição com maior espalhamento.

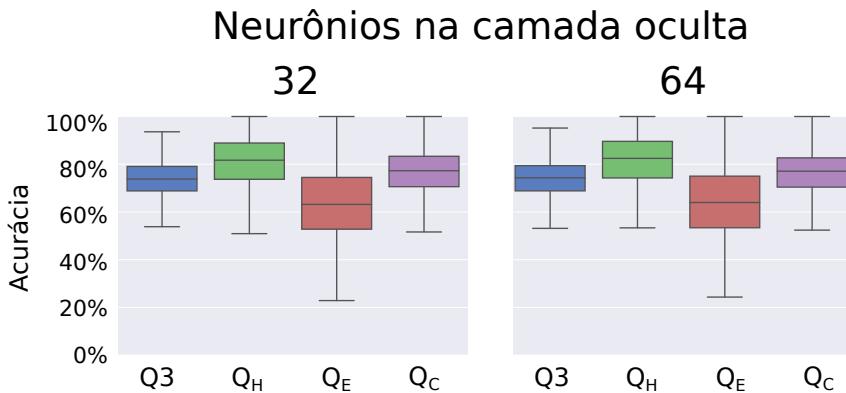


Figura 3.6: Distribuição da acurácia das redes com 32 e 64 neurônios na camada oculta para as proteínas do conjunto de teste.

3.3.1.5 Distribuição do tamanho das estruturas secundárias

A análise do comprimento das estruturas secundárias preditas para as proteínas do conjunto de teste apresentou uma distribuição diferente da observada nos métodos de atribuição (Comparação da figura 3.7 com 2.6, 2.7 e 2.8).

A alta ocorrência de elementos de estruturas secundárias preditos com poucos resíduos de comprimento é, provavelmente, uma consequência do fato da rede neural fazer a predição para cada resíduo individualmente. Isto implica que a estrutura secundária predita para o resíduo r_i não terá qualquer efeito nos resíduos vizinhos, anteriores ou posteriores a r_i . Assim, apesar da alta acurácia observada na predição de hélices ($Q_H > 80\%$) é possível observar que aproximadamente 50% das hélices preditas tem menos de 4 resíduos, o que difere muito dos dados experimentais.

3.3.2 Análise da predição com o PSIPRED

O PSIPRED apresentou uma alta acurácia na predição de estruturas secundárias atingindo valores superiores aos reportados no artigo original ($Q_3 \simeq 78\%$). A comparação da estrutura predita com a estrutura

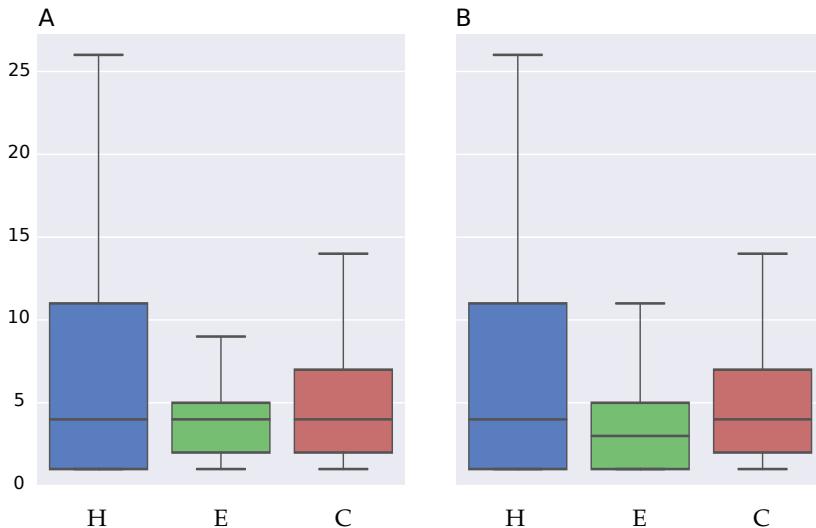


Figura 3.7: Distribuição dos comprimentos dos elementos de estrutura secundária preditos por redes neurais similares ao modelo de Holley e Karplus. (A) Modelo com 32 neurônios na camada oculta. (B) Modelo com 64 neurônios na camada oculta.

atribuída pelos quatro métodos apresentou uma acurácia mediana de aproximadamente 85% (Figura 3.8). De acordo com os valores medianos, o DSSP foi o método que atribuiu a estrutura secundária mais semelhante a estrutura predita. Este resultado era esperado uma vez que o PSIPRED utiliza os dados do DSSP como referência durante o treinamento da rede neural artificial.

Assim como no método testado anteriormente, similar ao de Holley e Karplus, a acurácia observada para as regiões de consenso apresentaram uma acurácia superior. Para o PSIPRED o valor mediano foi de aproximadamente 93%. Considerando-se as porcentagens de 74,4% para os resíduos que apresentam consenso na atribuição da estrutura secundária e 25,6% dos resíduos que não apresentam consenso, podemos inferir que a acurácia para as regiões de dissenso ficou na faixa de 60 a 65%.

A comparação da predição para cada elemento de estrutura secundária demonstra que o PSIPRED apresenta uma maior acurácia para regiões de hélice e a menor acurácia é observada para as regiões de fitas (Figuras 3.9, 3.10 e 3.11).

A distribuição dos comprimentos das estruturas secundárias preditas pelo PSIPRED é similar a distribuição dos comprimentos observados experimentalmente (Comparação da figura 3.12 com 2.6, 2.7 e 2.8).

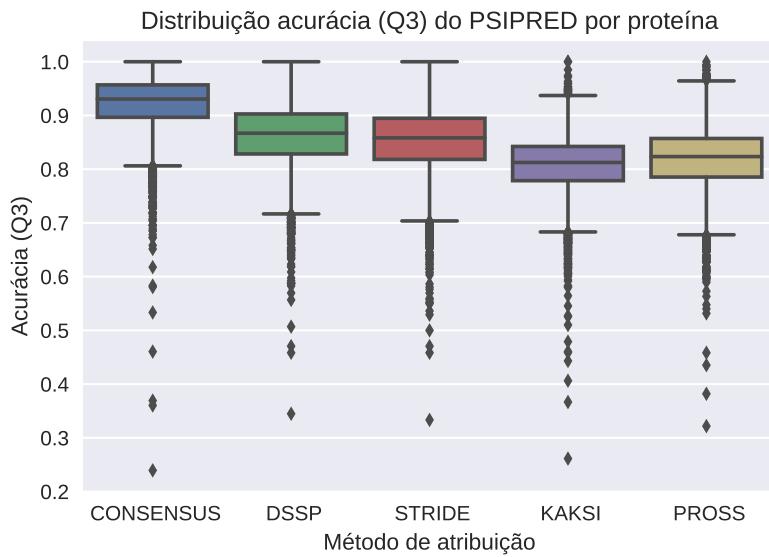


Figura 3.8: Distribuição dos valores de Q3 obtidos pelo PSIPRED para cada proteína do conjunto. A comparação foi feita com as estruturas secundárias definidas pelos quatro métodos atribuição e com as regiões de consenso entre eles.

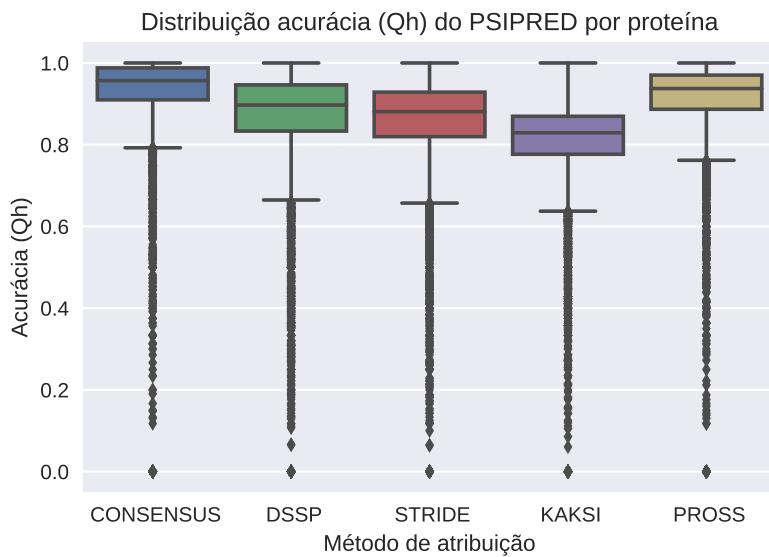


Figura 3.9: Distribuição da acurácia obtida pelo PSIPRED na predição de hélices.

CAMADA OCULTA (# de neurônios)	FONTE	TREINAMENTO		TESTE	
		CONSENSO	DSSP	STRIDE	KAKSI
		(%)	(%)	(%)	(%)
2	CONSENSO	68,72	62,66	62,42	61,90
	DSSP	68,26	62,92	62,44	61,62
	STRIDE	68,38	62,70	62,60	61,68
	KAKSI	67,80	61,76	61,58	62,20
4	PROSS	67,40	62,32	61,84	61,20
	CONSENSO	71,24	64,90	64,54	64,00
	DSSP	70,60	65,00	64,40	63,44
	STRIDE	70,92	65,04	64,82	63,78
8	KAKSI	69,62	63,10	63,08	63,90
	PROSS	69,96	64,48	63,96	63,28
	CONSENSO	73,27	66,40	66,23	65,93
	DSSP	73,03	66,90	66,43	65,53
16	STRIDE	72,97	66,57	66,50	65,63
	KAKSI	72,43	65,40	65,40	66,40
	PROSS	71,93	65,87	65,50	64,90
	CONSENSO	74,47	67,37	67,23	66,93
	DSSP	73,93	67,77	67,30	66,30
	STRIDE	73,97	67,40	67,47	66,50
	KAKSI	73,77	66,23	66,37	67,70
	PROSS	73,07	66,73	66,37	65,87

Tabela 3.2: Redes neurais treinadas usando como dados de entrada a estrutura atribuída por diferentes métodos ou somente resíduos com consenso entre eles. A tabela mostra a influência dos dados de treinamento na acurácia durante a predição. (A tabela continua na [3.4](#)).

CAMADA OCULTA (# de neurônios)	TREINAMENTO		TESTE			
	FONTE	CONSENSO	DSSP	STRIDE	KAKSI	
		(%)	(%)	(%)	(%)	
32	CONSENSO	75,17	67,87	67,80	67,57	68,83
	DSSP	74,60	68,43	67,97	66,83	68,97
	STRIDE	74,77	68,23	68,30	67,17	68,77
	KAKSI	74,27	66,57	66,73	68,20	67,50
64	PROSS	73,83	67,33	67,00	66,43	70,20
	CONSENSO	75,27	67,93	68,00	67,67	68,77
	DSSP	74,77	68,70	68,20	66,93	69,07
	STRIDE	74,90	68,43	68,53	67,23	68,83
128	KAKSI	74,47	66,67	66,87	68,47	67,57
	PROSS	73,90	67,43	67,10	66,50	70,37
	CONSENSO	75,20	67,97	67,90	67,63	68,73
	DSSP	74,57	68,63	68,10	66,80	68,90
	STRIDE	74,77	68,27	68,47	67,17	68,60
	KAKSI	74,30	66,50	66,77	68,43	67,47
	PROSS	73,87	67,33	67,03	66,50	70,23

Tabela 3.4: Continuação da tabela 3.2.

CAMADA OCULTA (# de neurônios)	DADOS	ENTROPIA CRUZADA	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)
2	TREINAMENTO	0,721	68,8	74,8	50,1	72,6
	VALIDAÇÃO	0,720	68,8	73,9	51,0	72,5
	TESTE	0,723	68,7	74,0	50,5	72,8
4	TREINAMENTO	0,680	71,0	76,5	55,2	73,5
	VALIDAÇÃO	0,680	71,0	74,5	57,8	73,8
	TESTE	0,677	71,2	76,3	55,9	74,2
8	TREINAMENTO	0,630	73,6	79,3	58,7	75,5
	VALIDAÇÃO	0,627	73,6	79,9	57,8	74,6
	TESTE	0,634	73,3	79,0	58,2	75,1
16	TREINAMENTO	0,604	74,8	80,4	61,3	76,1
	VALIDAÇÃO	0,607	74,7	79,0	61,5	76,9
	TESTE	0,610	74,5	81,0	60,8	76,4
32	TREINAMENTO	0,580	75,9	81,3	63,2	76,9
	VALIDAÇÃO	0,591	75,4	81,2	60,7	76,7
	TESTE	0,597	75,2	81,8	61,1	76,3
64	TREINAMENTO	0,557	77,0	82,3	64,7	77,8
	VALIDAÇÃO	0,580	75,8	81,3	62,1	76,7
	TESTE	0,595	75,2	81,8	61,2	75,7
128	TREINAMENTO	0,532	78,1	83,2	66,4	79,0
	VALIDAÇÃO	0,595	75,4	81,1	63,9	74,9
	TESTE	0,600	75,2	80,8	62,8	75,7

Tabela 3.5: Acurácia obtida na predição de hélices, fitas e coils para os conjuntos de treinamento, validação e teste em redes neurais com diversos números de neurônios na camada oculta.

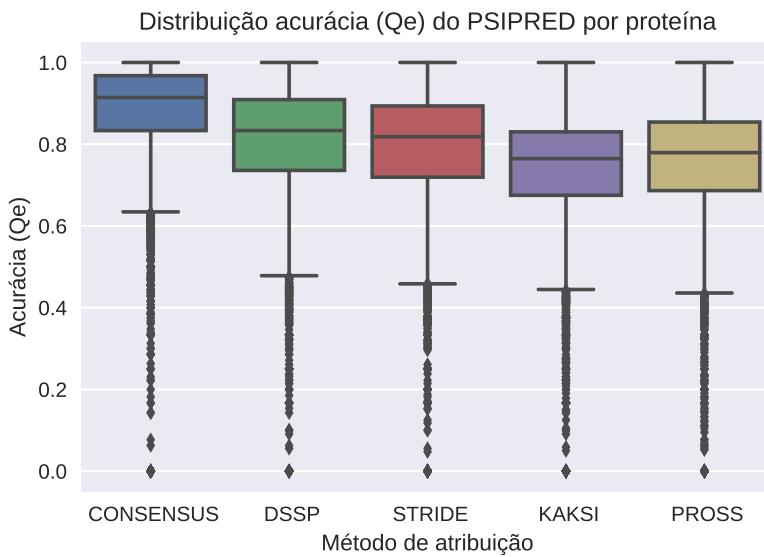


Figura 3.10: Distribuição da acurácia obtida pelo PSIPRED na predição de fitas.

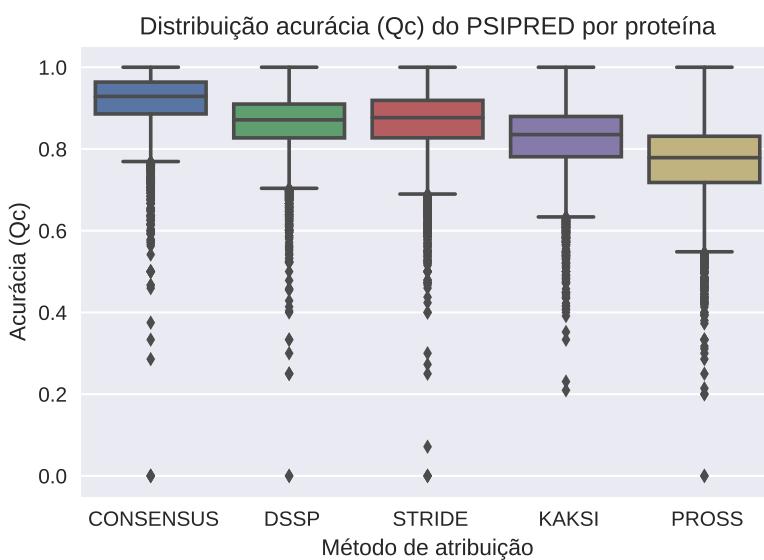


Figura 3.11: Distribuição da acurácia obtida pelo PSIPRED na predição de coils.

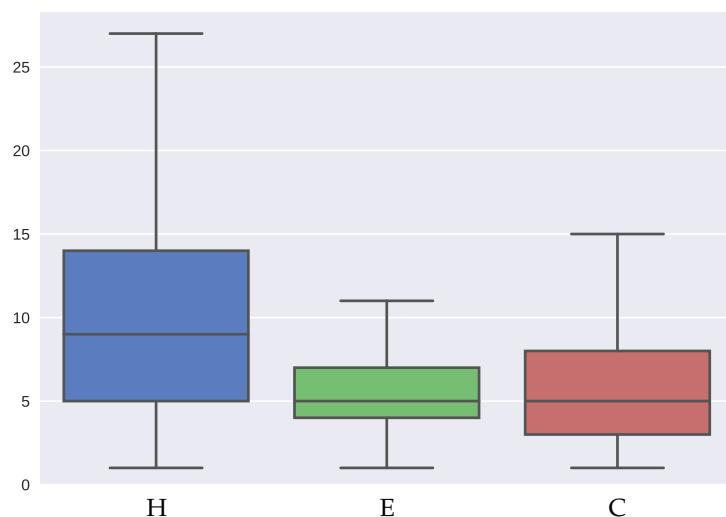


Figura 3.12: Distribuição dos comprimentos dos elementos de estrutura secundária preditas utilizando o PSIPRED.

4

AUTÔMATOS CELULARES

4.1 INTRODUÇÃO

Neste capítulo descrevemos a implementação e os resultados dos autômatos celulares aplicados a predição de estruturas secundárias. Foram testados quatro modelos de autômatos celulares que diferem essencialmente pelos estados utilizados para representar os elementos de estruturas secundárias.

No primeiro modelo testado tais estados representam apenas um dos três elementos de estrutura secundária, hélice, fita ou coil. Nos demais modelos, os estados que representam a estrutura secundária conservam, parcialmente, a informação do resíduo que originou a estrutura secundária.

A construção de autômatos celulares que reproduzam um padrão desejado implica na busca por um conjunto de regras capazes de conduzir a evolução do autômato celular **AC** de um estado inicial, conhecido, até um estado final, também conhecido. Neste trabalho, o estado inicial é a sequência de resíduos da proteína e o estado final, sua estrutura secundária. Essa busca por um conjunto de regras, chamada de problema inverso, é um problema de otimização e para isso foi utilizado um algoritmo de estimativa de distribuição (**EDA** - *Estimation of distribution algorithm*), um tipo de algoritmo evolutivo.

4.1.1 Autômatos celulares

Os autômatos celulares foram inventados na década de 40 por John von Neumann baseando-se em sugestões de seu colega, o matemático Stanislaw Ulam [62]. Autômatos celulares são modelos matemáticos para representar sistemas complexos que consistem num conjunto de células espacialmente discretas. Cada uma das células estão em um estado dentre um conjunto finito de estados possíveis. Os autômatos celulares evoluem paralelamente, ou seja, o estado de cada célula evolui de maneira síncrona em passos discretos de tempo e de acordo com regras simples, locais e determinísticas. A evolução dos **AC** produz complexidade a partir do efeito cooperativo de elementos simples - as regras e as células - tratando-se portanto de uma complexidade emergente [63].

Autômatos celulares tem sido utilizados em diversos campos de pesquisa como por exemplo na modelagem de sistemas: (1) biológicos, desde eventos intracelulares, como redes de interação proteicas, até estudo de populações; (2) químicos, na modelagem cinética de

sistemas moleculares e no crescimento de cristais; (3) físicos, para o estudos sistemas dinâmicos, desde a interação entre partículas até o agrupamento de galáxias [64].

4.1.2 Autômatos celulares elementares

O tipo mais simples de **AC** são os autômatos celulares elementares (**ACE**). Devido a sua simplicidade e a semelhança com os modelos propostos neste trabalho, acreditamos que explicá-los nesta introdução seja útil para facilitar uma melhor compreensão.

Os **ACE** possuem espaço unidimensional, o qual pode ser representado como um conjunto linear de células (Figura 4.1).

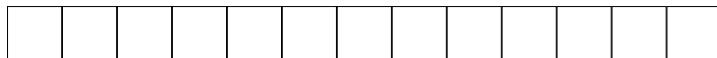


Figura 4.1: Espaço discreto e unidimensional dos autômatos celulares representado por um conjunto linear de células. Cada célula corresponde a uma região do espaço.

Cada célula dos **ACE** possuem um estado discreto, esse estado pode ser 0 ou 1 (Figura 4.2).

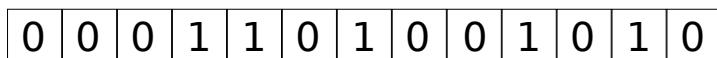


Figura 4.2: Estados discretos dos **ACE** representados com 0 e 1. Cada célula encontra-se em um estado.

Os autômatos celulares são modelos dinâmicos onde os estados discretos de suas células evoluem ao longo de um tempo também discreto (Figura 4.4). Essa evolução ocorre de acordo com um conjunto de regras (Figura 4.3). No caso dos **ACE**, essas regras possuem vizinhança 1 ($r=1$), o que indica que o estado de uma célula no tempo $t+1$ depende do estado da própria célula no tempo t e também dos estados das duas células vizinhas: uma a esquerda e uma a direita ($r=1$).

Apesar dos **ACE** serem o modelo mais simples de **AC**, os padrões produzidos por diferentes regras são interessantes. Wolfram [65] classifica esses padrões em quatro tipos: (1) uniformes, (2) repetitivos, (3) aleatórios e (4) complexos (Figura 4.5).

4.1.3 Autômatos celulares aplicados à predição de estruturas secundárias

Em 2007, Chopra e Bender [66] publicaram um método de predição de estruturas secundárias utilizando autômatos celulares. Neste trabalho, os autores utilizando um autômato celular com vizinhança cinco, ou seja, o estado de cada célula evolui de acordo com o estado da célula central e de cinco células vizinhas a direita e cinco

t	1 1 1	1 1 0	1 0 1	1 0 0
$t+1$	0	1	1	1

t	0 1 1	0 1 0	0 0 1	0 0 0
$t+1$	1	0	1	0

Figura 4.3: Exemplo de um conjunto de regras para um ACE. Os padrões de três células no tempo t ocasionam a mudança do estado da célula central no tempo $t+1$. A figura representa a “regra 110” dos ACE. Ao todo, os ACE possuem 2^8 (256) regras possíveis, pois há dois estados possíveis para cada um dos 8 elementos. As regras diferem entre si pelos estados $t+1$ dos elementos.

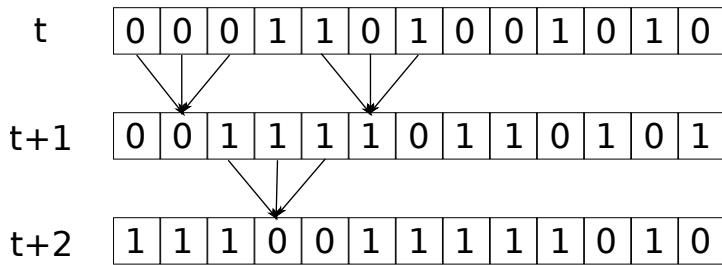


Figura 4.4: Esquema da evolução de um ACE. Cada célula no estado t evolui para o estado $t+1$ e depois para o estado $t+2$ através da aplicação das regras da figura 4.3.

a esquerda. Os estados de cada célula representam as probabilidades de cada resíduo ser hélice, fita e coil, e portanto, não são estados discretos. Para os estados iniciais do autômato, foram utilizadas as probabilidades definidas por Chou e Fasman [67]. A transição do autômato é definida por pesos para as probabilidades das 11 células no tempo t que influenciam na alteração das probabilidades no tempo $t+1$. Tais pesos foram otimizados utilizando um algoritmo genético. O autômato celular evolui por apenas duas gerações, pois, segundo os autores, um maior número de gerações reduzia a acurácia. A acurácia apresentada foi de aproximadamente 58% (Q_3).

O autômato celular desenvolvido por nós neste trabalho difere do publicado por Chopra e Bender [66] por utilizar estados discretos, vizinhança igual a um e um número superior de gerações durante a evolução. Acreditamos que tais alterações sejam mais adequadas para um método cujo um dos objetivos é fornecer informação sobre o processo de formação das estruturas secundárias proteicas.

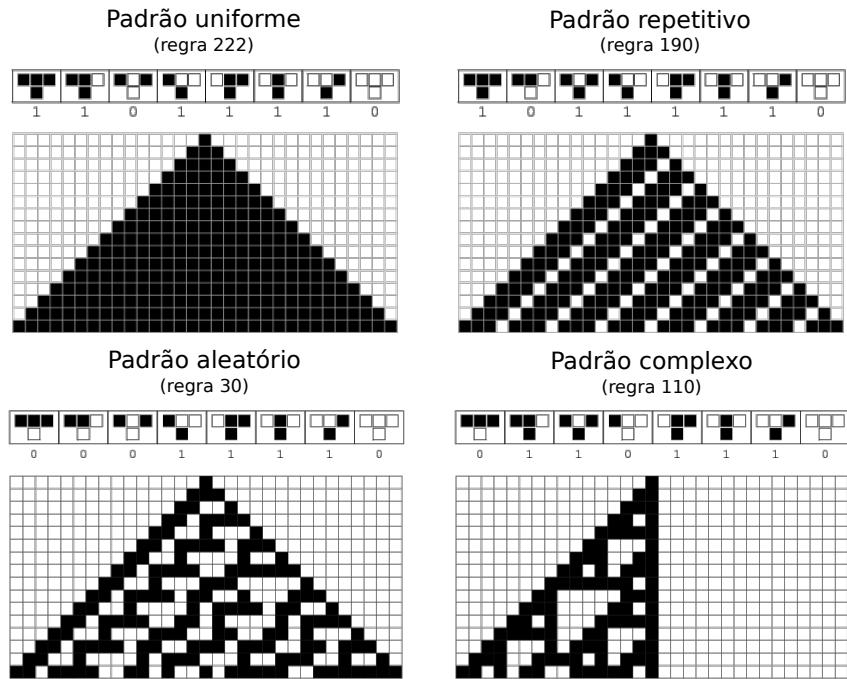


Figura 4.5: Exemplos de padrões produzidos por diferentes regras dos ACE.

4.2 MATERIAIS E MÉTODOS

4.2.1 Autômatos celulares

Neste trabalho foram utilizados autômatos celulares (AC) unidimensionais com vizinhança 1. Nesse aspecto, eles são semelhantes aos autômatos celulares elementares. Entretanto, a principal diferença ocorre nos estados possíveis para cada célula.

O conjunto de regras são formadas por elementos do tipo:

- $\boxed{a} \boxed{c} \boxed{p} \rightarrow \boxed{c'}$

Onde \boxed{c} é a célula central, \boxed{a} a célula anterior, \boxed{p} a célula posterior e $\boxed{c'}$ é a célula central no tempo imediatamente posterior ($t+1$).

O número de elementos no conjunto de regras depende do número de estados possíveis S para as células \boxed{a} , \boxed{c} e \boxed{p} . Como o número de estados é igual para as três células temos que o número de elementos no conjunto de regras é $n(S)^3$ ³.

O espaço do conjunto de regras possíveis depende também dos estados possíveis em $\boxed{c'}$. Considerando esses estados como S' , temos que o espaço do conjunto de regras possíveis é $n(S')^{n(S)^3}$ ³. É importante notar que no nosso modelo $S \neq S'$, enquanto que nos ACE (4.1.2) ambos são iguais a 2.

Primeiro iremos descrever os estados S , pois nesse trabalho testamos quatro modelos de autômatos celulares que diferem justamente pelos estados S utilizados.

Alguns estados são comuns entre os quatro modelos. Esses estados comuns correspondem aos aminoácidos e a um estado que indica o início/fim da cadeia polipeptídica. São eles:

- Estados S comuns entre os AC testados: {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, #}

onde # corresponde ao estado de início/fim da cadeia polipeptídica. Para os estados que representam aminoácidos, utilizamos o código de uma letra.

No primeiro modelo construído acrescentamos o menor número de estados possíveis para que o AC tivesse a capacidade de realizar a previsão de estruturas secundárias. Assim, nesse modelo nós acrescentamos apenas três novos estados, os quais, representam os elementos de estrutura secundária hélice, fita e coil. São eles, respectivamente:

- Estados S adicionados para o primeiro modelo: {*}, |, -}

Portanto, no primeiro modelo temos $n(S) = 24$.

Nos demais três modelos nós adicionamos uma informação que indica o contexto do elemento de estrutura secundária. Esse contexto está relacionado a uma característica do resíduo presente na mesma posição em que encontra-se o elemento de estrutura secundária.

O segundo modelo utiliza dois contextos para representar a hidrofobicidade. A separação dos resíduos polares e apolares seguiu a proposta por Rose e colaboradores [68]. São eles:

1. Polar (p) que representa o contexto de resíduos polares: D, E, G, H, K, N, P, Q, R, S, T, Y
2. Apolar (n) que representa o contexto de resíduos apolares: A, C, F, I, L, M, V, W

Assim, são adicionados 6 estados ao segundo modelo. Esses estados correspondem aos três elementos de estrutura secundária em dois contextos diferentes. São eles:

- Estados S adicionados para o segundo modelo: {*p}, {*n}, |p, |n, -p, -n}

Os quais representam, respectivamente: hélice polar, hélice apolar, fita polar, fita apolar, coil polar e coil apolar. Assim, nesse modelo temos $n(S) = 27$.

No terceiro modelo foram acrescentados dois novos contextos aos já utilizados no segundo modelo. Esses contextos representam os resíduos prolina e glicina. Consequentemente, os contextos utilizados no terceiro modelo são:

1. Polar (p) que representa o contexto de resíduos polares: D, E, H, K, N, Q, R, S, T, Y

2. Apolar (n) que representa o contexto de resíduos apolares: A, C, F, I, L, M, V, W
3. Prolina (P) que representa apenas o resíduo prolina
4. Glicina (G) que representa apenas o resíduo glicina

E os estados:

- Estados S adicionados para o terceiro modelo: $\{*\boxed{p}, \boxed{*n}, \boxed{*P}, \boxed{*G}, \boxed{|p}, \boxed{|n}, \boxed{|P}, \boxed{|G}, \boxed{-p}, \boxed{-n}, \boxed{-P}, \boxed{-G}\}$

Neste modelo temos $n(S) = 33$.

No quarto modelo foram adicionados outros dois contextos aos utilizados no terceiro modelo. Eles representam os resíduos carregados positivamente e negativamente. Então os contextos neste modelo são:

1. Polar (p) que representa o contexto de resíduos polares: H, N, Q, S, T, Y
2. Apolar (n) que representa o contexto de resíduos apolares: A, C, F, I, L, M, V, W
3. Prolina (P) que representa apenas o resíduo prolina
4. Glicina (G) que representa apenas o resíduo glicina
5. Positivos (+) que representa o contexto dos resíduos K e R
6. Negativos (-) que representa o contexto dos resíduos D e E

Apesar da histidina (H), em pH fisiológico, poder ser observada também no estado protonado (H^+) e portanto com carga positiva [1], nós optamos por categorizá-la de forma geral como polar (neutro). Esse padrão é adotado por outros métodos computacionais como por exemplo o Rosetta e é mencionado no trabalho de Kilambi e Gray [69]

Portanto, temos neste modelo a adição dos seguintes estados:

- Estados S adicionados para o terceiro modelo: $\{*\boxed{p}, \boxed{*n}, \boxed{*P}, \boxed{*G}, \boxed{|+}, \boxed{|-}, \boxed{|p}, \boxed{|n}, \boxed{|P}, \boxed{|G}, \boxed{|+}, \boxed{|-}, \boxed{|p}, \boxed{|n}, \boxed{|P}, \boxed{|G}, \boxed{|+}, \boxed{|-}\}$

E $n(S) = 39$.

Os estados possíveis de S' são apenas quatro em todos os modelos testados. São eles:

- Estados $S' : \{*, |, -, ?\}$

Esses estados representam, respectivamente, uma transição para o estado de hélice, uma transição para fita, transição para coil e a permanência no estado anterior, ou seja, representa a não mudança de estado entre t e $t+1$.

Nos três modelos onde foram utilizadas a informação sobre o contexto, a mesmo é adicionada após a transição de estado. Por exemplo, supondo um elemento de regra como:

- $\boxed{F} \boxed{V} \boxed{T} \rightarrow \boxed{I}$

O estado \boxed{V} no tempo t , irá ser alterado para \boxed{I} no tempo $t+1$. No entanto, considerando o contexto apolar do resíduo V, o estado da célula será alterado para $\boxed{|n|}$.

Após essa descrição dos estados é possível calcularmos o número de elementos da regra e o espaço de regras possíveis (Tabela 4.1).

AC	Contextos	S	Elementos por regra	Regras possíveis
1	-	24	$24^3 = 13824$	4^{13824}
2	Polar (p) Apolar (n)	27	$27^3 = 19683$	4^{19683}
3	Polar (p) Apolar (n) Prolina (P) Glicina (G)	33	$33^3 = 35937$	4^{35937}
4	Polar (p) Apolar (n) Prolina (P) Glicina (G) Positivos (+) Negativos (-)	39	$39^3 = 59319$	4^{59319}

Tabela 4.1: Número de elementos no conjunto de regras e número de regras possíveis para os quatro modelos testados.

4.2.2 Otimização do conjunto de regras

4.2.2.1 Algoritmo de estimação de distribuição

A busca por regras de um autômato celular que reproduzam um padrão específico, conhecido como problema inverso, é um problema de otimização. Na literatura, esse problema é normalmente abordado utilizando meta-heurísticas como algoritmos genéticos ou têmpera simulada (*simulated annealing*) [64]. Neste trabalho optamos por utilizar um Algoritmo de Estimação de Distribuição (EDA) que implementamos de forma distribuída através do modelo mestre-escravos. No modelo mestre-escravos, um processo mestre coordena o trabalho passando tarefas para os processos escravos executarem. Após a execução, os escravos enviam os resultados para os mestres.

O funcionamento do EDA distribuído na busca de regras para os autômatos celulares envolve os seguintes passos:

1. Mestre envia as probabilidades para os escravos

2. Escravos recebem as probabilidades e a utilizam para criar dois conjuntos de regras candidatas
3. Escravos utilizam as regras candidatas para evoluir dois autômatos celulares
4. Escravos calculam a acurácia da predição (*fitness*) dos dois autômatos celulares
5. Escravos enviam o conjunto de regra vencedor para o Mestre
6. O mestre acumula N conjunto de regras vencedoras e então atualiza as probabilidades
7. Retorna ao passo 1 ou termina se a geração final (G) do EDA foi atingida

A seguir descrevemos os passos em detalhes.

4.2.2.2 Probabilidades enviadas pelo mestre aos escravos

O mestre inicia o processo enviando um probabilidade uniforme (Figura 4.6) onde cada um dos quatro estados de transição possíveis tem probabilidade inicial de 25%. Isso ocorre para cada elemento do conjunto de regras com exceção de elementos cujo estado da célula central apresenta o estado de início/fim da cadeia polipeptídica (`#`). Nesses elementos a probabilidade da célula continuar no mesmo estado no tempo $t+1$ é igual a 100%. Consequentemente, todo elemento da regra que tiver o estado `#` na célula central terá a transição definida como `?` e permanecerão no estado `#`.

```
[ An ][ # ][ An ] -> { _ : 0.00, * : 0.00, | : 0.00, ? : 1.00 }
[ # ][ An ][ # ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ # ][ An ][ Vn ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ Sp ][ An ][ Wn ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ D- ][ An ][ _n ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ _p ][ An ][ *n ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ # ][ An ][ |n ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ # ][ An ][ D- ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ # ][ An ][ E- ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ # ][ An ][ Gp ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ *n ][ *p ][ Hp ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
[ # ][ An ][ K+ ] -> { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }
```

Figura 4.6: Exemplo de parte dos dados da probabilidade inicial enviada pelo mestre aos escravos. Os escravos construirão as regras candidatas de acordo com essa informação.

4.2.2.3 Construção das regras pelos escravos

Após os escravos recebem as probabilidades enviadas pelo mestre eles constroem dois conjuntos de regras candidatas utilizando tais probabilidades (Figura 4.7). Cada conjunto contém todos os elementos da regra necessários para a evolução do autômato.

Um conjunto de regras é construído da seguinte forma:

1. Para cada elemento do conjunto de regras é gerado um número aleatório entre 0 e 1
2. São construídos quatro intervalos utilizando as probabilidades do elemento. Ex:
 - a) Se o elemento [Sp][An][Wn] tem probabilidades { _ : 0.25, * : 0.25, | : 0.25, ? : 0.25 }. São construídos os intervalos: [0, 0.25], [0.25, 0.50], [0.50, 0.75], [0.75, 1.00]

A transição será definida pelo intervalo que contiver o número aleatório. Suponha que o número aleatório fosse 0.67, ele estaria contido no 3º intervalo, e consequentemente esse elemento da regra seria definido como uma transição para fita.



```
[ # ][ An ][ Hp ] -> [ An ]
[ # ][ An ][ Np ] -> [ |n ]
[ # ][ An ][ Qp ] -> [ *n ]
[ # ][ An ][ Sp ] -> [ |n ]
[ # ][ An ][ Tp ] -> [ |n ]
[ # ][ An ][ Yp ] -> [ _n ]
[ # ][ An ][ _p ] -> [ _n ]
[ # ][ An ][ *p ] -> [ _n ]
[ # ][ An ][ |p ] -> [ _n ]
```

Figura 4.7: Exemplo contendo elementos de uma regra candidata produzido pelo escravo.

4.2.2.4 Evolução dos autômatos celulares pelos escravos

Cada um dos dois conjuntos de regras gerados pelos escravos são utilizados para evoluir dois autômatos celulares por 100 passos. O estado inicial de cada autômato celular corresponde a concatenação de todas as sequências de aminoácidos das proteínas (Figura 4.8). Nessa concatenação, são adicionados o estado de início/fim (#) entre cada proteína e no início e no fim da concatenação. Esses estados # adicionados isolam a propagação de informação entre as proteínas durante a evolução do autômato.

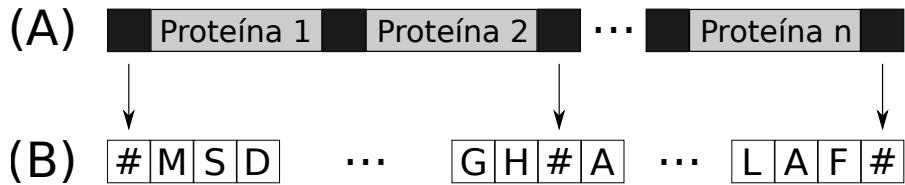


Figura 4.8: Concatenação da sequências de proteínas do conjunto de treinamento. (A) Esquema geral da concatenação. (B) Colocação de estados de ínicio/fim da sequência para impedir a troca de informação entre proteínas.

4.2.2.5 Cálculo da acurácia dos dois conjunto de regras pelo escravo

Após construção e a evolução dos dois autômatos pelo escravo são calculados a acurácia dos dois conjuntos de regras. A medida de acurácia é utilizada como *fitness* do conjunto de regras. Há, então, um torneio entre os dois conjuntos de regras e o vencedor é enviado para o mestre.

Inicialmente utilizamos como medida da acurácia o próprio Q_3 . Para o cálculo, o estado de estrutura secundária mais frequente ao longo da evolução do **AC** foi considerado como o resultado da predição para o resíduo.

Essa medida inicial foi utilizada para avaliar o desempenho dos quatro modelos de autômatos celulares testados.

Posteriormente, decidimos avaliar outras métricas de acurácia para tentar melhorar a capacidade preditiva do modelo mais promissor. Essas outras métricas testadas foram a *Class Balance Accuracy* (**CBA**) [70], o coeficiente de correlação de Matthews (**MCC**) e a entropia cruzada (**CE**). As duas primeiras métricas são eficientes para medir a acurácia em classes desbalanceadas [70].

CBA é definido como:

$$CBA = \frac{\sum_i^k \frac{c_{ii}}{\max(c_i, c_{\cdot i})}}{k} \quad (4.1)$$

Onde C^k é uma matriz de confusão e k representa as classes hélice, fita e coil, logo, $k = 3$. O “.” na equação representa todos os valores $k \neq i$.

		Observados		
		H	E	C
Preditos	H	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$
	E	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$
	C	$c_{3,1}$	$c_{3,2}$	$c_{3,3}$

Tabela 4.2: Matriz de confusão C^k para o cálculo do **CBA**.

MCC é definido como:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4.2)$$

e para aplicá-lo a um problema com múltiplas classes, nós utilizamos:

$$\begin{aligned} TP &= TP_H + TP_E + TP_C \\ TN &= TN_H + TN_E + TN_C \\ FP &= FP_H + FP_E + FP_C \\ FN &= FN_H + FN_E + FN_C \end{aligned}$$

E a entropia cruzada por resíduo é (CE_r):

$$CE_r = - \sum_i p(\hat{y}) \ln(y) \quad (4.3)$$

onde \hat{y} é probabilidade observada experimentalmente para estrutura secundária de um resíduo da proteína, a qual será 1 para o elemento de estrutura secundária atribuído. y será a frequência dos estados de estrutura secundária durante a evolução do AC. A CE final é a média das CE_r calculadas para cada resíduo que apresentou consenso na atribuição da estrutura secundária.

4.2.2.6 Atualização das probabilidades do EDA no mestre

Após o mestre receber $N/2$ (N população do EDA) conjunto de regras dos escravos as probabilidades dos elementos são atualizadas. Esse procedimento é realizado observando-se a frequência das transições para cada elemento entre as regras vencedoras recebidas pelo mestre. Por exemplo, se dentre as regras recebidas pelo mestre observarmos as seguinte frequência para o elemento $\boxed{A} \boxed{A} \boxed{A}$:

- $\boxed{A} \boxed{A} \boxed{A} \rightarrow \boxed{*} = 3000$
- $\boxed{A} \boxed{A} \boxed{A} \rightarrow \boxed{|} = 500$
- $\boxed{A} \boxed{A} \boxed{A} \rightarrow \boxed{-} = 1000$
- $\boxed{A} \boxed{A} \boxed{A} \rightarrow \boxed{?} = 500$

e considerando $N/2$ igual a 5000, temos que a nova probabilidade para esse elemento será:

- $[An][An][An] \rightarrow \{ _ : 0.20, * : 0.60, | : 0.10, ? : 0.10 \}$

Após as probabilidades de todos os elementos serem calculadas, elas começam a serem enviadas para os escravos para que esses iniciem a produção de novas regras candidatas.

4.2.2.7 Detalhes do procedimento

O método de otimização e os autômatos celulares foram implementados na linguagem Go. A comunicação entre o mestre e os escravos ocorrem por chamadas de procedimento remotos (RPC). Os programas foram executados no cluster de computação de alto desempenho EMU-2 (adquirido pelo “Programa de Equipamento Multusuário da FAPESP-2009”, 2009/53853-5, localizado no Centro Internacional de Pesquisa e Ensino do Hospital A. C. Camargo em São Paulo). O cluster contém 7 nós de processamento, cada um com 4 processadores AMD Opteron 6276 que possuem 16 cores (2,3GHz). A execução utilizando 320 cores de processamento demora aproximadamente 10 dias.

Parâmetros utilizados

- População do **EDA**: $N = 10000$
- Número de gerações do **EDA**: $G = 1000$
- Número de gerações do **AC**: $t = 100$

4.3 RESULTADOS

4.3.1 Seleção do modelo de autômato celular

A capacidade dos quatro modelos de autômatos celulares foram testadas num conjunto reduzido de proteínas. O objetivo deste teste foi avaliar se os estados utilizados em cada modelo de **AC** seriam capazes de predizer as estruturas secundárias com acurácia. Os resultados apresentados demonstram que o primeiro modelo, o qual não utiliza informação sobre o contexto, falhou em predizer estruturas menos frequentes como as fitas e coils (Figura 4.9).

Os outros três modelos, os quais utilizam informação sobre o contexto, obtiveram uma melhor acurácia na produção de estruturas secundárias. A acurácia aumentou a medida que aumentamos os tipos de contextos utilizados.

4.3.2 Análise da predição de estruturas secundárias

As análises da capacidade de predição de estruturas secundárias foram realizadas para o modelo quatro dos autômatos celulares testados, o qual utiliza os contextos polar, apolar, glicina, prolina, positivo e negativo.

Na otimização das regras do **AC** utilizando o subconjunto de treinamento nós utilizamos como funções de *fitness* as métricas **CBA**, **MCC** e **CE**. Essas métricas foram utilizadas pois notamos que o uso direto do Q_3 não era eficaz no treinamento com classes desbalanceadas.

Durante o treinamento por 1000 gerações do **EDA** não foram observadas convergências para a população de regras candidatas. En-

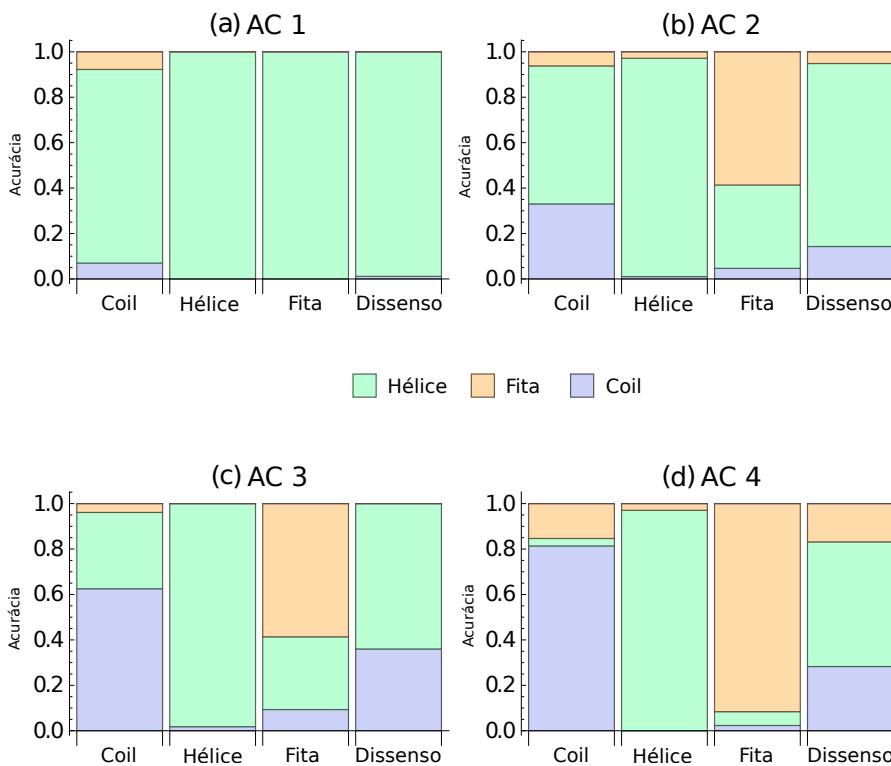


Figura 4.9: Acurácia dos modelos de AC testados em um conjunto reduzido de proteínas. O AC 4, o qual utiliza um maior número de contextos, apresentou maior capacidade de reproduzir a estrutura secundária.

tretanto, ao final da otimização o CBA estava aumentando a um taxa média de 0,00005 por geração, o MCC 0,00015 por geração e o CE reduzindo a uma taxa de 0,00002 por geração (Figura 4.10).

O AC treinado utilizando como função de *fitness* o CBA foi o que apresentou a melhor acurácia, tanto para o subconjunto de treinamento quanto para o de teste. Entretanto, a acurácia média do modelo foi de apenas 62,6% (Q_3) para o conjunto de teste (Figura 4.11). A entropia cruzada (CE) e o MCC obtiveram menor acurácia, 53,4% e 51,1%, respectivamente. Estes valores, medidos para os resíduos que apresentaram consenso na atribuição da estrutura secundária, são significativamente inferiores à acurácia apresentada pelas redes HK ($Q_3 \simeq 75\%$) e pelo método PSIPRED ($Q_3 \simeq 93\%$).

A acurácia na predição de hélices (Q_H) foi maior no modelo treinado utilizando a entropia cruzada, com valor próximo a 73,6%. O treinamento com CBA apresentou acurácia de 66,7% e o com MCC, acurácia de 60,5% (Figura 4.12).

A acurácia para hélices (Q_E) foi de 52,2% para o modelo treinado com CBA, 37,8% para o treinado com MCC e 21% para o treinado com CE. Como há um menor número resíduos atribuídos a estrutura secundária fita (Figura 2.3), isso demonstra que o CBA foi a métrica

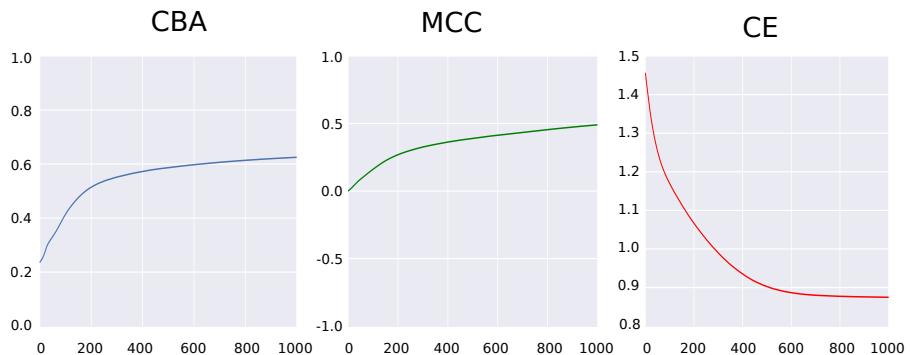


Figura 4.10: Valor médio observado durante a evolução do EDA por 1000 gerações utilizando as métricas **CBA**, **MCC** e **CE** como *fitness*.

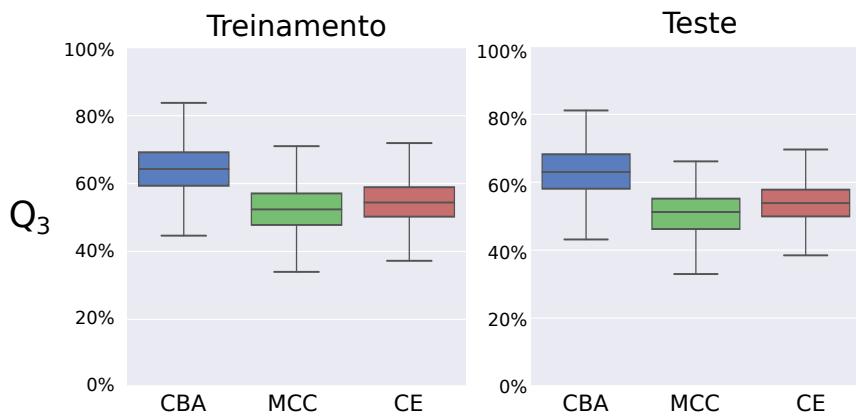


Figura 4.11: Distribuição da acurácia (Q_3) apresentada pelo **AC** 4 para proteínas do conjunto de treinamento e teste.

mais eficaz em nosso problema com classes desbalanceadas. Por outro lado, a entropia cruzada foi a menos eficaz.

A acurácia na predição de coil (Q_C) foi de 69,0% no modelo treinado utilizando **CBA**, 67,6% no que utilizou **CE** e 53,1% utilizando **MCC**.

Além da baixa acurácia, as estruturas secundárias preditas pelos autômatos celulares apresentaram comprimentos inferiores aos observados em estruturas secundárias de proteínas com estrutura resolvida (Figura 4.15). A mediana de hélices foi de apenas 4 resíduos enquanto que a mediana esperada seria entre 8 e 12 resíduos (Figura 2.6). Fitas e coils apresentam, respectivamente, medianas de 3 e 4 resíduos para o **AC** treinado com **CBA**. O esperado nestes casos seriam 5 resíduos para fitas e 4 resíduos para coils (Figuras 2.7 e 2.8).

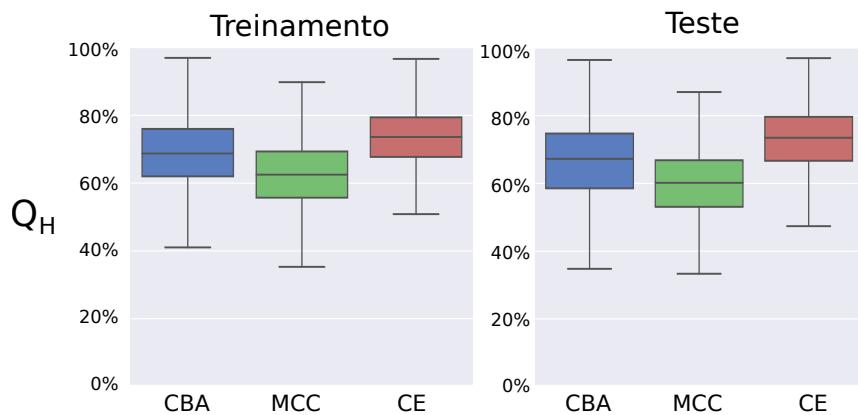


Figura 4.12: Distribuição da acurácia na predição de hélices (Q_H) apresentada pelo [AC 4](#) para proteínas do conjunto de treinamento e teste.

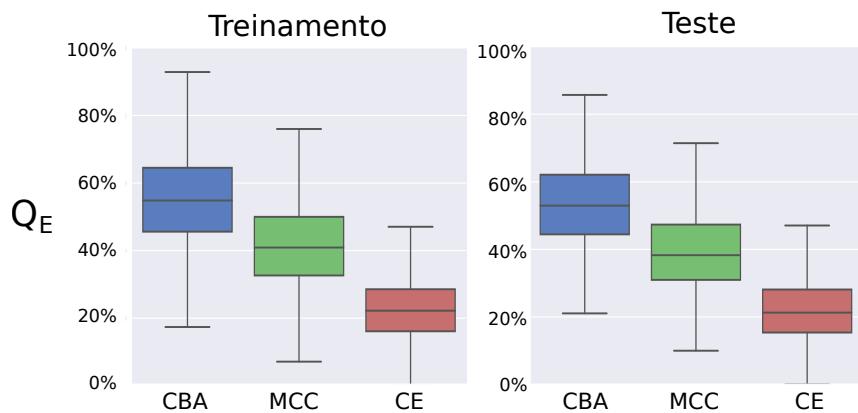


Figura 4.13: Distribuição da acurácia na predição de fitas (Q_E) apresentada pelo [AC 4](#) para proteínas do conjunto de treinamento e teste.

4.3.3 Exemplos de evolução do autômato celular para predizer a estrutura secundária

A observação da evolução do [AC 4](#) treinado com [CBA](#), o melhor modelo segundo nossos resultados, indica que ocorre uma rápida convergência ou estabilidade dos estados (Figuras 4.16 e 4.17). Ao propormos a utilização de autômatos celulares na predição de estruturas secundárias esperávamos observar a formação de padrões complexos como, por exemplo, alguns poucos resíduos iniciando a formação de uma hélice e os resíduos na vizinhança propagando essa informação. Um padrão similar a esta descrição pode ser observado na primeira hélice da Figura 4.17.

Entretanto, a baixa acurácia obtida pelo [AC 4](#) em conjunto com a estabilidade dos estados durante a evolução do [AC 4](#) não nos forne-

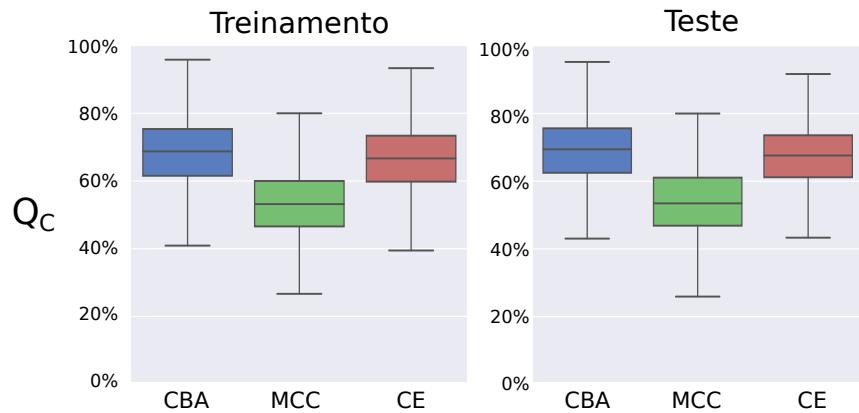


Figura 4.14: Distribuição da acurácia na predição de coils (Q_C) apresentada pelo [AC](#) 4 para proteínas do conjunto de treinamento e teste.

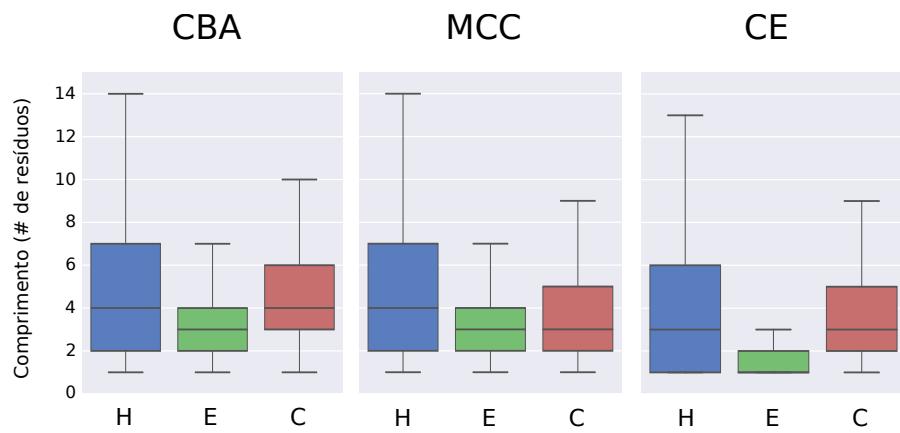


Figura 4.15: Distribuição do comprimento de estruturas secundárias preditas para o conjunto de treinamento. Independentemente da métrica utilizada na otimização das regras, o [AC](#) 4 produziu estruturas com comprimento inferior ao observado experimentalmente para as três classes de estruturas secundárias: hélice (H), fita (E) e coil (C).

cem informações que possam auxiliar numa melhor compreensão de processos de formação de estruturas secundárias.

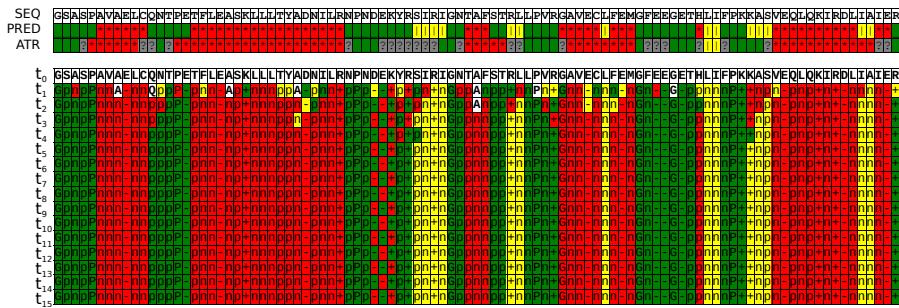


Figura 4.16: Evolução do AC4 com a regra otimizada por EDA utilizando CBA como função de fitness. O domínio PUB da proteína *Peptide-N(4)-(N-acetyl-beta-glucosaminyl) asparagine amidase* (PDB ID: 2CCQ) apresentou uma das maiores acurárias com $Q_3 = 90,23\%$, $Q_H = 84,31\%$, $Q_E = 100\%$ e $Q_C = 86,36\%$. Estados que representam hélices estão em vermelho, fitas em amarelo e coils em verde.

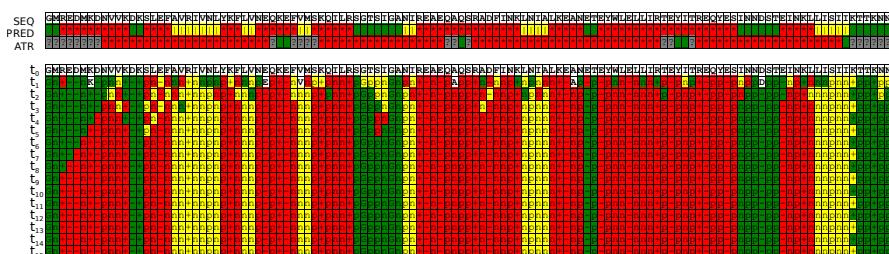


Figura 4.17: Evolução do AC4 com a regra otimizada por EDA utilizando CBA como função de fitness. A proteína sem função descrita (PDB ID: 2RLD) apresentou uma das piores acurárias com $Q_3 = 30\%$, $Q_H = 60,00\%$ e $Q_C = 0\%$. Estados que representam hélices estão em vermelho, fitas em amarelo e coils em verde.

5

REDES NEURAIS RESIDUAIS PROFUNDAS

5.1 INTRODUÇÃO

Neste capítulo implementamos e avaliamos os resultados de redes neurais residuais profundas, ou simplesmente redes residuais, aplicadas a predição de estruturas secundárias a partir da estrutura primária da proteína.

De modo similar aos autômatos celulares, a utilização de redes residuais foi motivada pela possibilidade de construirmos uma arquitetura que possibilitasse extrairmos informação sobre a formação de estruturas secundárias. Esse objetivo tornou nosso trabalho diferente de outros que utilizaram redes neurais profundas para a predição de estruturas secundárias. Em um desses trabalhos, publicado por Wang, Ma e Xu [71], os autores usaram redes neurais convolucionais profundas (não residuais) e, realizando a predição a partir de PSSM, obtiveram uma acurácia de aproximadamente 84% (Q_3). Na discussão comparamos ambos os modelos.

5.1.1 Redes neurais residuais

Redes neurais com mais de uma camada oculta de neurônios são chamadas de redes neurais profundas [72]. Devido ao maior número de camadas ocultas, as redes neurais profundas são capazes de integrar atributos de vários níveis de abstração de maneira hierárquica. Isso ocorre porque cada camada oculta de neurônios constrói atributos derivados a partir da informação contida nas camadas anteriores [73].

Dentre as arquiteturas de redes neurais profundas, as redes neurais convolucionais tem obtido grande sucesso em diversas aplicações relacionadas a imagens, como por exemplo, a classificação e a detecção de objetos em fotografias [74].

Nas redes convolucionais são utilizados neurônios que possuem conexões locais com a camada imediatamente anterior. Cada tipo de neurônio é então aplicado a camada anterior realizando uma espécie de varredura. Assim, um tipo de neurônio que responde a um determinado padrão local, irá responder a esse padrão independentemente de onde ele ocorra espacialmente. Isso difere das redes neurais totalmente conectadas (*Fully Connected*), como as utilizadas no capítulo 3, onde cada neurônio de uma camada está conectado a todos os neurônios da camada anterior.

Para uma explicação mais intuitiva, podemos pensar em uma rede neural convolucional cujo objetivo é identificar a ocorrência de um

rosto humano em fotografias. Nessa rede poderíamos ter um tipo de neurônio que seria responsável por identificar a presença de um olho, outro tipo neurônio para um nariz e outro tipo para a boca. Cada neurônio estaria conectado a uma pequena região da imagem, mas o conjunto de todos os neurônios do mesmo tipo estariam distribuídos cobrindo toda a imagem. Assim, neurônios do tipo que respondem a presença da boca emitiriam uma resposta somente na região da imagem onde há uma boca. O mesmo ocorreria para os demais tipos de neurônios. Na camada seguinte, um tipo de neurônio capaz de identificar faces humanas iria sinalizar a presença da mesma caso houvessem sinais de boca, nariz e olhos em regiões próximas. Essa rede exemplifica a hierarquia entre as camadas e a identificação de padrões mais complexos derivados de padrões mais simples. Caso fosse utilizada uma rede neural totalmente conectada, mesmo utilizando várias camadas ocultas, seria perdida a referência espacial. Isto significa que mesmo que houvessem neurônios capazes de identificar olhos, bocas e narizes, a rede não saberia interpretar se eles estariam posicionados numa face ou apenas como elementos dispersos na foto.

No exemplo anterior, por motivos didáticos, mencionamos tipos de neurônios capazes de identificar atributos específicos na imagem: olho, boca e nariz. Entretanto, é importante esclarecer que os neurônios não são pré definidos em relação aos atributos que eles identificam. Esse processo ocorre durante o treinamento das redes neurais e não costuma ser direcionado para atributos reais como no exemplo.

A capacidade de detectar padrões locais, combinada as múltiplas camadas das redes profundas, possibilita criarmos um modelo de predição com semelhanças importantes ao modelo de autômatos celulares. Essas semelhanças, as quais motivaram esse trabalho, são a identificação de padrões locais na sequência capazes de iniciar a formação e/ou propagarem a informação ao longo da proteína até a formação da estrutura secundária nativa.

Comparando ambos os métodos, redes neurais convolucionais profundas e autômatos celulares, podemos listar *a priori* algumas vantagens como:

- o modelo não exige uma pré determinação dos estados intermediários, como foi necessário com os autômatos celulares, os quais necessitam de estados discretos;
- as redes neurais possuem métodos de treinamento bem estabelecidos, eficazes e eficientes.

5.2 MATERIAIS E MÉTODOS

5.2.1 Arquitetura da rede residual

Dentre as arquiteturas de redes neurais convolucionais profundas, o modelo que acreditamos ser o mais adequado são o de redes neurais residuais. Esse modelo, proposto inicialmente por He e colaboradores [74], facilitou o treinamento de redes convolucionais profundas, permitindo a construção de redes com mais de 100 camadas de profundidade.

No modelo proposto por nós, a camada de entrada da rede utiliza a representação *one hot encoding*, onde cada aminoácido é representado por um vetor com 22 posições. Essas posições representam os 20 aminoácidos e mais duas posições que representam a ausência de um aminoácido anterior ou posterior a sequência de resíduos. A utilização de códigos para a ausência de aminoácidos tem como objetivo manter constante a somatória dos valores de entrada para as convoluções (Figura 5.1).

Com o objetivo de tornar mais didática a codificação de entrada, podemos comparar a entrada com uma imagem, o tipo de dado mais comum para redes convolucionais. As imagens coloridas (RGB) são representadas como *pixels*, onde cada *pixel* contém 3 valores, chamado de canais, que representam a intensidade das cores vermelho, verde e azul. O tamanho da imagem corresponde ao número de *pixels* que ela tem de largura e altura. Assim, uma imagem colorida de tamanho 800x600, tem dimensões 3x600x800 (*CHW - Channels, Height, Width*).

Na codificação da proteína, o número de canais é igual a 22 e cada um representa a ocorrência de um aminoácido ou a ausência deles como mencionado anteriormente. Uma das dimensões representa a cadeia polipeptídica e foi definida com tamanho 3000. Nessa dimensão, as primeiras 500 posições possuem o código de ausência de aminoácido anterior a proteína. A partir da posição 501, tem início os códigos que representam a sequência de resíduos da proteína. Após o término da sequência de resíduos, a matriz é preenchida com o código para ausência de aminoácidos posteriores a proteína. O tamanho foi pré-definido em 3000 para possibilitar que a cadeia polipeptídica seja longa em nosso conjunto de dados, a qual possui quase 2000 resíduos, tenha 500 códigos de ausência antes e depois. O tamanho de 500 foi intencionalmente superestimado durante a preparação dos dados para não limitar a exploração de arquiteturas diferentes como, por exemplo, redes com mais camadas de convolução ou neurônios de convolução com maior número de conexões. A outra dimensão da matriz é igual a 1, pois a cadeia polipeptídica é linear. Assim, a matriz de entrada possui dimensões 22x3000x1 (*CHW*).

Através da codificação inicial da sequência no formato *one hot encoding* é possível apenas inferir se os aminoácidos são iguais ou diferen-

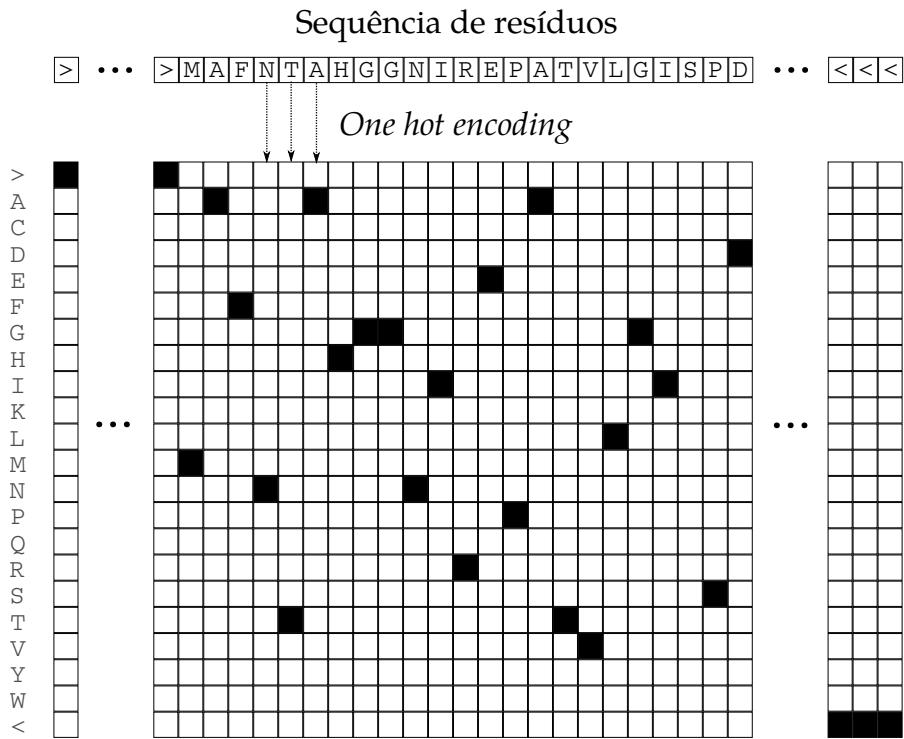


Figura 5.1: Codificação da sequência de aminoácidos da proteína no formato *one hot encoding*. Os códigos 'greater than' e 'less than' representam, respectivamente, a ausência de um aminoácido anterior e posterior a proteína.

tes. Consequentemente, um aspartato diferirá igualmente de um glutamato como de uma fenilalanina, ou seja, essa codificação não contém qualquer informação sobre similaridades estruturais e/ou físico-químicas.

Entretanto, ao invés de passarmos similaridades físico-químicas *a priori*, optamos por utilizar como segunda camada da rede (primeira camada oculta) uma convolução 1x1, que altera o número de canais de 22 para C , onde $C = \{8, 16, 32\}$. Essa camada tem como objetivo, aprender uma nova representação para os aminoácidos e portanto, similaridades entre eles que possam ser úteis para a predição da estrutura secundária (Figura 5.2).

A representação da segunda camada é mantida (residualmente) até a camada final, onde outra convolução 1x1 reduz o número de canais de C para 3. Os 3 canais finais estão relacionados aos 3 elementos de estrutura de secundária e, após a aplicação da função Softmax (Eq .5.1), representam a probabilidade de cada um dos 3 elementos, hélice, fita e coil, para cada resíduo da proteína.

$$f(x_i) = \frac{e^{x_i}}{\sum_j^N e^{x_j}} \quad (5.1)$$

É importante notar que nas camadas descritas até o momento, não há influência de resíduos vizinhos, dado que são utilizadas somente convoluções 1x1.

A influência da vizinhança é computada nos blocos da rede residual. O primeiro bloco recebe a saída da segunda camada e, após realizar seus cálculos, soma sua matriz de resultados a matriz de entrada. A matriz alterada pelo primeiro bloco, será a entrada para o segundo bloco, e assim sucessivamente (Figura 5.2). Após as alterações feitas pelos blocos, a camada final, descrita anteriormente, fará a convolução 1x1 de C para 3 canais. Como mencionado, os resultados da camada inicial são mantidos até a camada final, mas são alterados pelos blocos, ou seja, eles são mantidos apenas residualmente. Por essa características essa arquitetura de rede neural é chamada de rede residual.

5.2.2 Blocos

A composição dos blocos da rede residual é similar a proposta por Zagoruyko e Komodakis nas *Wide Residual Networks* [75]. Cada bloco contém duas funções de ativação não lineares do tipo Parametric Rectified Linear Unit (PReLU)[76], duas camadas de neurônios de convolução 3x1 e uma camada de *dropout* [77] para prevenir o sobreajuste (*overfitting*) (Figura 5.1).

A função PReLU é:

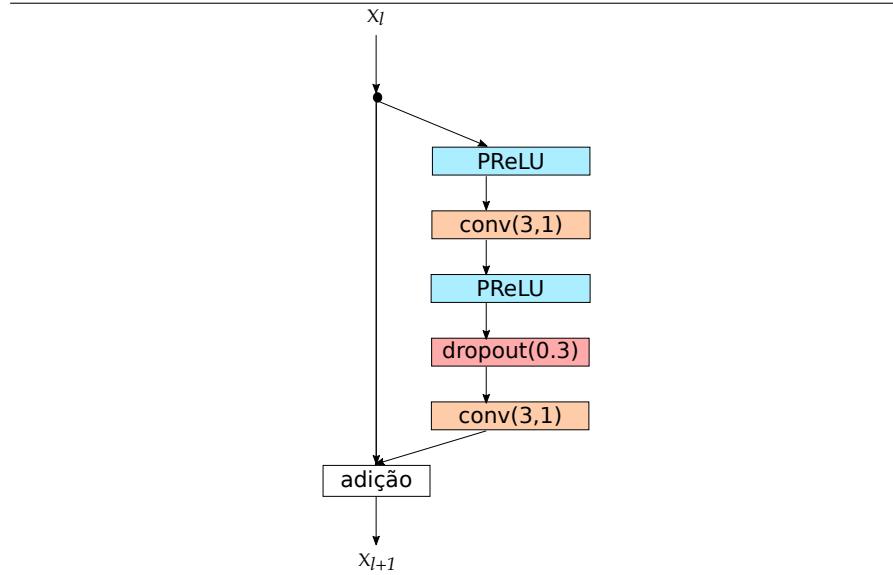
$$f(y_i) = \begin{cases} y_i & \text{se } y_i > 0 \\ \alpha_i y_i & \text{se } y_i \leq 0 \end{cases}$$

onde α_i é o coeficiente que controla a inclinação da parte negativa. Esse coeficiente é otimizado durante o treinamento. i é o índice do canal.

A camada de *dropout* atua bloqueando aleatoriamente algumas conexões da rede durante o treinamento. Isto é feito alterando os pesos das conexões para zero. A probabilidade de cada conexão ser bloqueada é indicada durante a construção do modelo, nós utilizamos uma probabilidade de 30%.

A camada de convolução 3x1 são as responsáveis por transmitir informação sobre a vizinhança do resíduos. Cada neurônio desta camada recebe valores referentes ao resíduo central e seus dois vizinhos e, após o cálculo, atribui o valor resultante na camada inferior e na posição correspondente ao resíduo central (Figura 5.3).

Algoritmo 5.1 Diagrama da configuração de bloco utilizada. Cada bloco possui duas funções de ativação PReLU, duas camadas de convolução 3×1 e uma camada de *dropout*. Cada bloco recebe a matriz de valores do bloco ou camada anterior e, após realizar seus cálculos, soma a matriz de resultados a matriz original.



5.2.3 Número de blocos e canais

O número de blocos define a profundidade da rede e, consequentemente, o tamanho da vizinhança, em número de resíduos, que irá influenciar na predição da estrutura secundária para cada resíduo.

Cada bloco possui duas camadas de convolução 3×1 e cada camada recebe a informação de dois vizinhos, um a esquerda e um a direita. Consequentemente, a utilização de 4 blocos determina que a predição da estrutura secundária para cada resíduo utilizará a informação contida numa janela de 17 resíduos da proteína ($4 \times 4 + 1$), sendo 8 resíduos anteriores e 8 resíduos posteriores. A rede com onze blocos utilizará a informação contida em uma janela de 45 resíduos ($11 \times 4 + 1$) e a com 21 blocos, a informação em uma janela de 85 resíduos ($21 \times 4 + 1$). Ou seja, a medida que o número de blocos aumenta, resíduos mais distantes na sequência proteica irão influenciar na predição da estrutura secundária local.

O número de canais tem relação com a capacidade de representar a informação local. Assim, os canais tem função similar ao número de neurônios da camada oculta no modelo de rede similar à de Holley e Karplus (Subseção 3.2.2) e, também, similar ao número de estados de transição do autômato celulares.

5.2.4 Treinamento

Cada rede neural foi treinada por 1000 épocas utilizando o algoritmo de aprendizado Adam [61] com taxa de aprendizado de 0.001. Uma época equivale a utilização de todas as proteínas do conjunto de treinamento.

As proteínas do conjunto de treinamento são utilizadas em lotes (*minibatches*) de 64, para as redes maiores, ou 128, para as menores. Essa variação foi necessária devido à memória disponível. O tempo médio de treinamento de cada rede foi de ~10 horas utilizando um computador com placa de vídeo (GPU) Nvidia GTX 1060 6GB. As redes foram implementadas usando o *framework* Pytorch (disponível em <http://pytorch.org/>).

5.3 RESULTADOS

5.3.1 Análise do treinamento

O treinamento das redes residuais demonstrou que tanto o aumento do número de blocos quanto o aumento do número de canais reduz o erro no conjunto de dados de treinamento. Entretanto, nas redes com 11 e 21 blocos, a utilização de 32 canais (ou filtros) tornou a rede suscetível ao sobreajuste (*overfitting*). Isso pode ser observado pela diferença entre a entropia cruzada no conjunto de treinamento e a entropia cruzada no conjunto de validação (Figuras 5.5 e 5.6).

A análise do treinamento indica que, dentre as redes residuais testadas, a que apresentou melhor desempenho foi a rede com 21 blocos e 16 canais.

5.3.2 Acurácia por tipo de estrutura secundária

Os resultados da acurácia por tipo de estrutura secundária indicam que o aumento do número de blocos, e consequentemente, a utilização de resíduos mais distantes na sequência para a predição da estrutura secundária, aumenta, sobretudo, a acurácia na predição de fitas.

Ao compararmos a rede residual b₂₁-c₁₆ (21 blocos e 16 canais) com a rede b₄-c₁₆ observamos um aumento de 2,8% na acurácia (Q_3) obtida no conjunto de teste. Respectivamente, Q_3 igual a 78,3% e 75,5%. No entanto, ao compararmos a acurácia por tipo de estrutura secundária observamos um aumento de 8,3% (68,0% e 59,7%) na predição de fitas, de 3,5% para a predição de coils (79,9% e 76,4%) e uma diminuição da acurácia na predição de hélices de 0,6% (81,9% e 82,5%) (Tabela 5.1).

A comparação com a rede neural similar a proposta por Holley e Karplus (Subseção 3.3.1) e com 64 neurônios na camada oculta (hk64)

mostra que a redes residuais com 11 e 21 blocos apresentam maior acurácia na predição. Quando comparamos especificamente a rede residual b21-c16 com a rede hk64 observamos uma acurácia (Q_3) 3,1% maior. A comparação da acurácia por tipo de estrutura secundária mostra uma acurácia 6,8% maior para fitas, 4,2% maior para coils e 0,1% menor para hélices.

5.3.3 Acurácia em relação ao métodos de atribuição

A rede residual com 21 blocos e 16 canais (b21-c16) apresentou a maior acurácia tanto para os resíduos com consenso na atribuição da secundária ($Q_3=78,3\%$) quanto para a estrutura secundária atribuída por cada um dos métodos de atribuição (Tabela 5.3).

Considerando que 74,4% dos resíduos apresentaram consenso entre os métodos de atribuição (ver capítulo 2.3) é possível inferir a acurácia da predição para as regiões de dissenso como aproximadamente 47,8%. Tal acurácia é semelhante à obtida através do modelo de Holley e Karplus (47,6%) e inferior à apresentada pelo PSIPRED para regiões de dissenso, estimada entre 60-65%.

$$0,744 * 0,783 + 0,256 * Q_3 \text{dissenso} \simeq 0,705$$

$$Q_3 \text{dissenso} \simeq 0,478$$

5.3.4 Distribuição da acurácia

Distribuição de acurácia da predição para as proteínas do conjunto de teste (Figura 5.7).

5.3.5 Distribuição do tamanho das estruturas secundárias preditas

A distribuição do comprimento das estruturas secundárias preditas pelas redes residuais (Figura 5.8) é semelhante a observada para o método PSIPRED (Figura 3.12) e a distribuição do comprimento das estruturas secundárias atribuídas pelos métodos DSSP, STRIDE, KAKSI e PROSS (Figuras 2.6, 2.7 e 2.8).

5.3.6 Similaridade entre aminoácidos

Como descrito na seção 5.2, a segunda camada da rede residual (ou primeira camada oculta) foi projetada para aprender uma representação dos resíduos contendo informação da similaridade entre aminoácidos. A codificação inicial não contém informação de qualquer similaridade entre aminoácidos, assim, acreditamos que a rede deverá aprender as características de cada um deles de acordo com a

influência dos mesmos na formação/predição da estrutura secundária.

O agrupamento (*clustering*) hierárquico da representação aprendida pelas redes residuais apresentaram características interessantes que podem ser observadas nas figuras 5.9, 5.10 e 5.11. A seguir explicitamos algumas dessas características.

As redes demonstraram um clara tendência em separar os aminoácidos prolina e glicina dos demais. Somente nas redes b₁₁-c₃₂ e b₂₁-c₃₂ que apresentaram sobreajuste a glicina se aproxima de outros resíduos. A distinção dos aminoácidos prolina e glicina é interessante devido as características estruturais dos mesmo. Na prolina, a maior restrição dos ângulos torcionais da cadeia principal e a ausência de um hidrogênio ligado ao nitrogênio restringem sua ocorrência em elementos de estrutura secundária. Na glicina, a ausência de cadeia lateral amplia os possíveis estados conformacionais reduzindo as restrições dos ângulos torcionais. Tais características são únicas desses dois resíduos, e portanto, diferentes dos demais 18 aminoácidos.

O agrupamento apresentou ainda a tendência de separar aminoácidos com cadeias laterais polares e apolares.

No grupo dos apolares ocorreu a formação de subgrupos como: (1) isoleucina (I) e valina (V); (2) fenilalanina (F), tirosina (Y) e triptofano (W); (3) cisteína (C), metionina (M) e leucina (L).

No grupo dos polares, há a formação de subgrupos como serina (S) e treonina (T) e de carregados positivamente, lisina (K) e arginina (R). Os carregados negativamente, aspartato (D) e glutamato (E) aparecem próximos em alguns grupos, mas em outros ocorre um agrupamento de aspartato (D) com asparagina (N) e de glutamato (E) com glutamina (Q).

5.3.7 Análise do processo de predição

Analisamos a predição da estrutura secundária de duas proteínas do conjunto de teste utilizando a rede residual b₂₁-c₁₆, a qual apresentou a melhor acurácia. Selecionei uma proteína entre as que obtiveram melhor acurácia e uma entre as com pior acurácia.

Entre as que obtiveram melhor acurácia está o domínio N-terminal da proteína RssB, o qual apresenta uma topologia “*Rossmann fold*” (Figura 5.13). Os estados internos da rede residual indicam que logo nos primeiros blocos da rede os resíduos que compõem as fitas apresentam uma maior probabilidade para essa estrutura secundária. Por outro lado, apenas poucos resíduos de algumas hélices possuem maior probabilidade para esse estrutura nos blocos iniciais. Entretanto, a medida que os blocos vão alterando as probabilidades, as hélices começam a serem preditas nas regiões corretas (Figura 5.12).

Entre proteínas com menor acurácia na predição está a RmlC de *Streptococcus suis* (PDB ID: 1NXM) que possui topologia “Jelly roll” (Figura 5.15).

As probabilidades extraídas após cada bloco da rede residual indicam vários resíduos com probabilidade de fitas. Interessante notar que a quarta fita atribuída é predita parcialmente até a penúltima camada, quando então passa a ser predita como coil. Algo similar ocorre com a décima e décima primeira fitas atribuídas, quando, nas últimas camadas ocorre uma mudança na predição de fita para hélice. Ao analisarmos a estrutura notamos que essas fitas preditas incorretamente formam a interface de dimerização, formando uma folha β com fitas da outra cadeia polipeptídica.

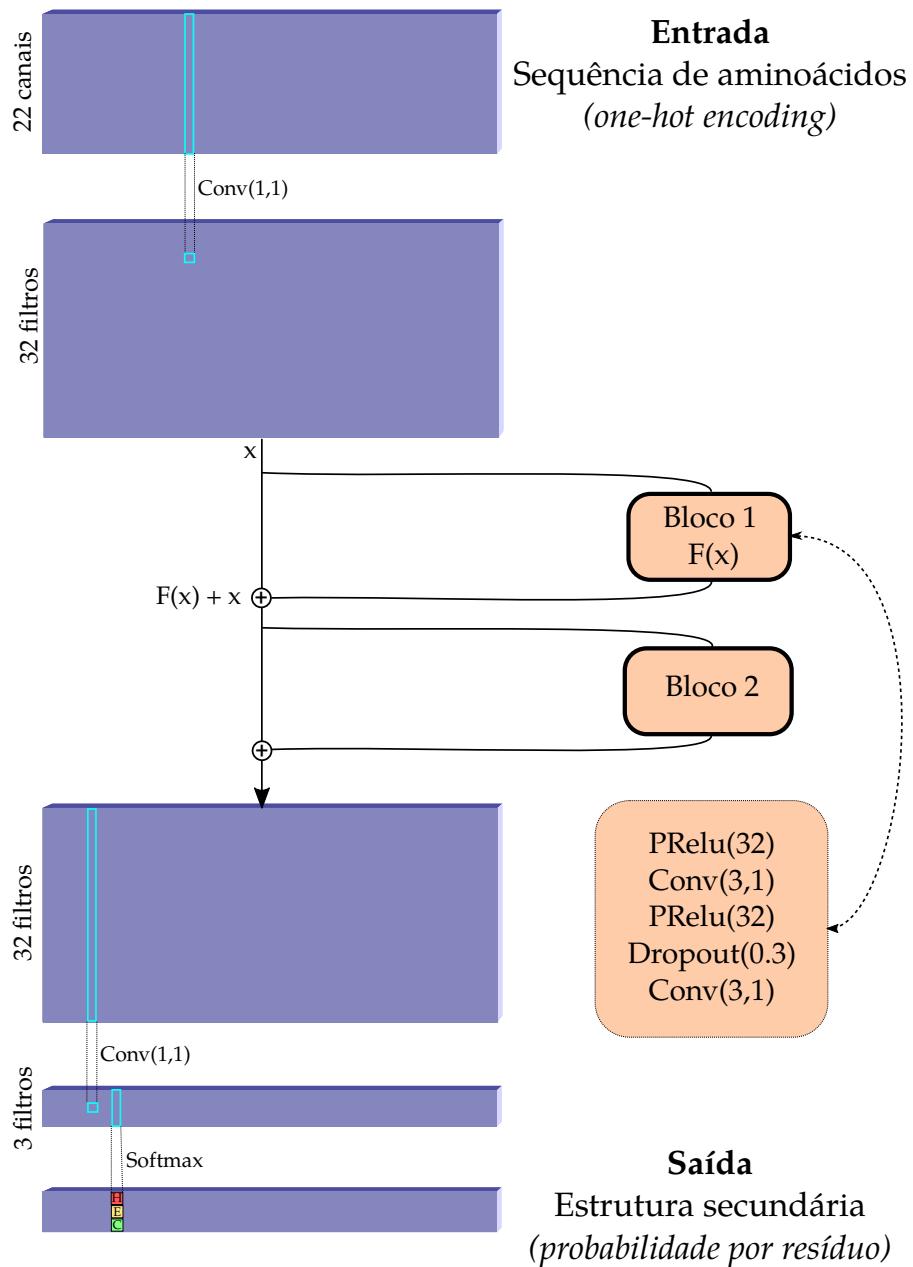


Figura 5.2: Diagrama da arquitetura da rede neural residual utilizada neste trabalho. A entrada da rede consiste na sequência de resíduos de uma proteína no formato *one-hot encoding*. A saída contém três probabilidades para cada resíduo, cada uma representa a probabilidade de uma estrutura secundária: hélice, fita ou coil. Para facilitar, apenas dois blocos estão representados e $C=32$ canais (ou filtros). No trabalho foram testadas redes com 4, 11 ou 21 blocos e 8, 16 ou 32 canais.

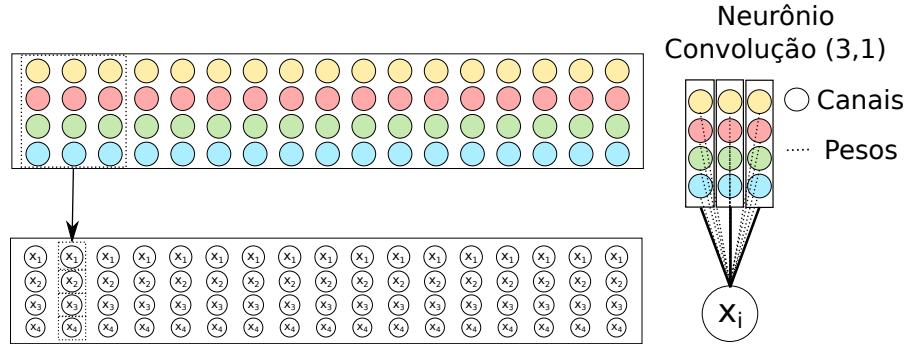


Figura 5.3: Esquema de um neurônio em uma camada de convolução 3x1. Cada neurônio utiliza os valores dos canais de 3 colunas, multiplica esses valores pelos pesos, soma, e atribui o resultado a um canal na posição correspondente. Os neurônios de um mesmo canal são idênticos, logo, o resultado irá variar de acordo com a entrada. Neurônios de canais diferentes tem pesos diferentes.

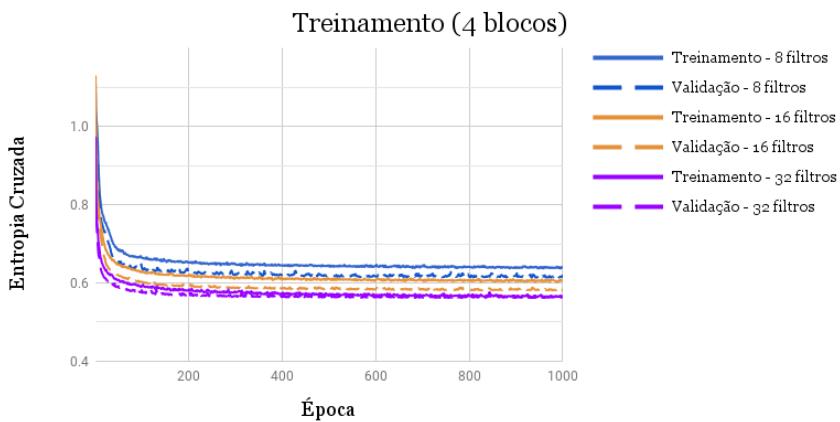


Figura 5.4: Aprendizado da rede residual com 4 blocos demonstrado pela redução da entropia cruzada ao longo de 1000 épocas. Os valores são para os dados do conjunto de treinamento e validação nas redes com 8, 16 e 32 canais (ou filtros).

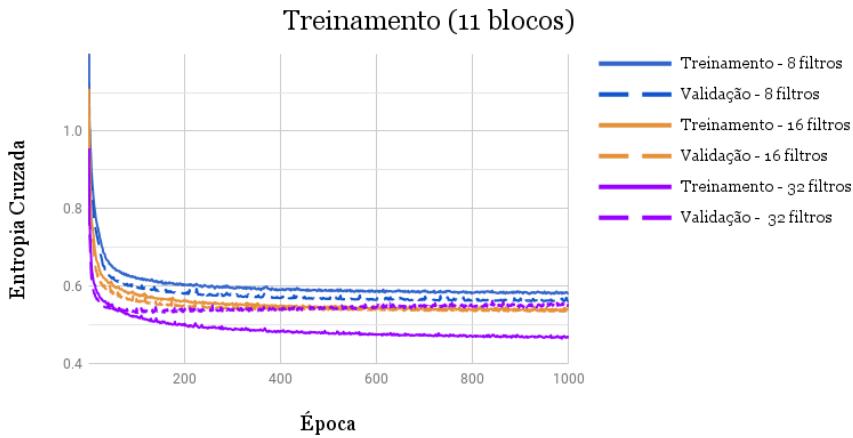


Figura 5.5: Aprendizado da rede residual com 8 blocos demonstrado pela redução da entropia cruzada ao longo de 1000 épocas. Os valores são para os dados do conjunto de treinamento e validação nas redes com 8, 16 e 32 canais (ou filtros). A rede com 32 canais apresenta sinais de sobreajuste.

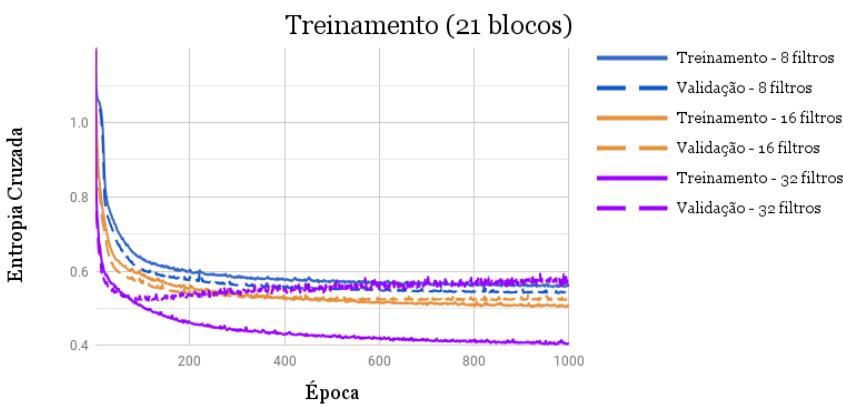


Figura 5.6: Aprendizado da rede residual com 21 blocos demonstrado pela redução da entropia cruzada ao longo de 1000 épocas. Os valores são para os dados do conjunto de treinamento e validação nas redes com 8, 16 e 32 canais (ou filtros). A rede com 32 canais apresenta fortes sinais de sobreajuste com a redução do erro no conjunto de treinamento e o aumento do erro no conjunto de validação.

CAMADA OCULTA (# de blocos)	DADOS	ENTROPIA CRUZADA	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)
8	TREINAMENTO	0,643	73,1			
	VALIDAÇÃO	0,621	73,9	74,7	63,6	78,5
	TESTE	0,626	73,7	74,7	63,3	78,3
16	TREINAMENTO	0,605	74,8			
	VALIDAÇÃO	0,581	75,7	82,2	59,9	76,4
	TESTE	0,59	75,5	82,5	59,7	76,4
32	TREINAMENTO	0,566	76,6			
	VALIDAÇÃO	0,563	76,7	83,9	62,1	76,0
	TESTE	0,571	76,3	84,1	61,8	75,8
8	TREINAMENTO	0,582	75,9			
	VALIDAÇÃO	0,562	76,7	83,9	59,0	77,7
	TESTE	0,571	76,5	83,9	59,3	77,7
16	TREINAMENTO	0,54	77,7			
	VALIDAÇÃO	0,545	77,4	78,9	68,6	80,3
	TESTE	0,549	77,3	79,1	68,3	80,3
32	TREINAMENTO	0,465	81,1			
	VALIDAÇÃO	0,554	78,1	82,3	68,1	78,4
	TESTE	0,563	77,6	82,1	67,7	78,1
8	TREINAMENTO	0,561	76,8			
	VALIDAÇÃO	0,546	77,5	81,2	64,5	80,2
	TESTE	0,554	77,2	81,3	64,1	79,9
21	TREINAMENTO	0,503	79,5			
	VALIDAÇÃO	0,525	78,6	82,2	68,3	80,0
	TESTE	0,531	78,3	81,9	68,0	79,9
32	TREINAMENTO	0,405	83,6			
	VALIDAÇÃO	0,559	78,6	85,2	64,7	78,1
	TESTE	0,570	78,0	84,9	64,3	77,8

Tabela 5.1: Acurácia apresentada pelas redes residuais em resíduos com consenso na atribuição da estrutura secundária.

CAMADA OCULTA	TREINAMENTO			TESTE				
	BLOCOS	CANAIS	FONTE	CONSENSO	DSSP	STRIDE	KAKSI	
				(%)	(%)	(%)	(%)	
4	8	CONSENSO		73,7	67,0	66,7	66,4	68,3
	16	CONSENSO		75,5	68,1	68,1	67,8	69,0
	32	CONSENSO		76,3	68,8	68,9	68,5	69,5
11	8	CONSENSO		76,5	68,8	69,0	68,9	69,9
	16	CONSENSO		77,3	69,9	69,9	69,7	71,2
	32	CONSENSO		77,6	70,0	70,2	69,9	70,9
21	8	CONSENSO		77,2	69,7	69,8	69,5	71,0
	16	CONSENSO		78,3	70,6	70,8	70,5	71,7
	32	CONSENSO		78,0	70,2	70,4	70,3	71,0

Tabela 5.3: Acurácia (Q_3) das redes residuais em relação a estrutura secundária atribuída por diferentes métodos computacionais.

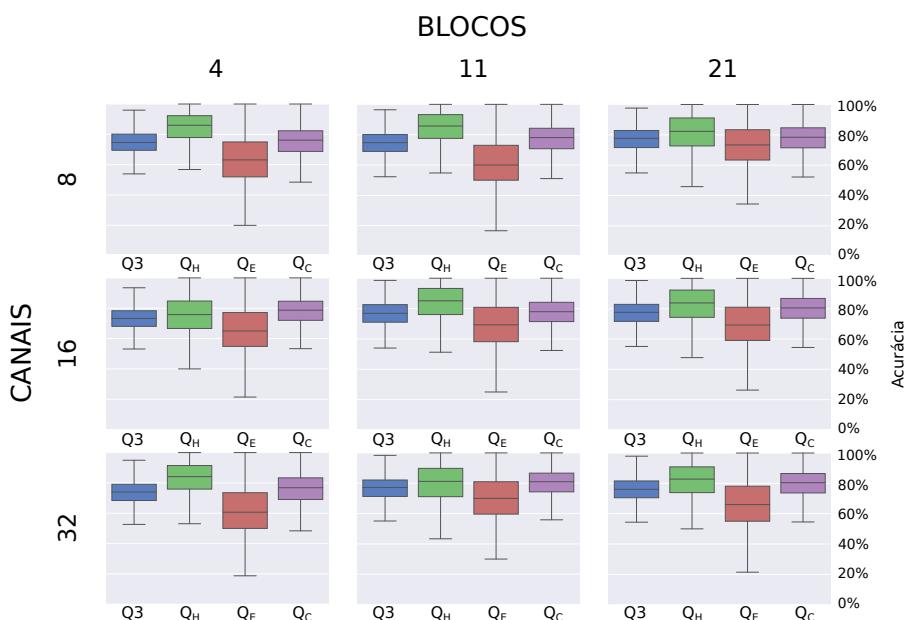


Figura 5.7: Distribuição de acurácia da predição para as proteínas do conjunto de teste.

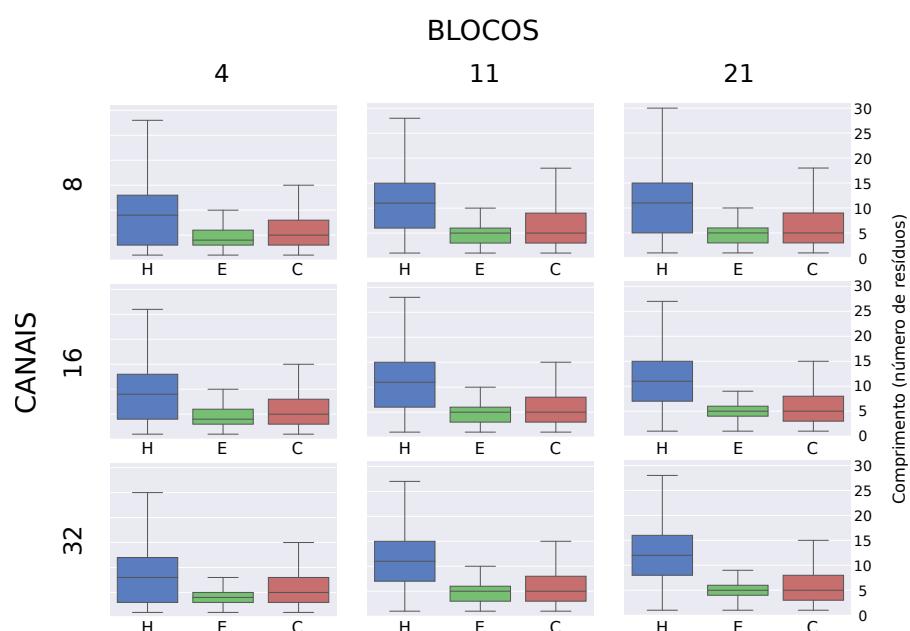


Figura 5.8: Distribuição do comprimento das estruturas secundárias preditas pelas redes residuais. As hélices, fitas e coils preditos apresentam comprimento similar ao observado experimentalmente.

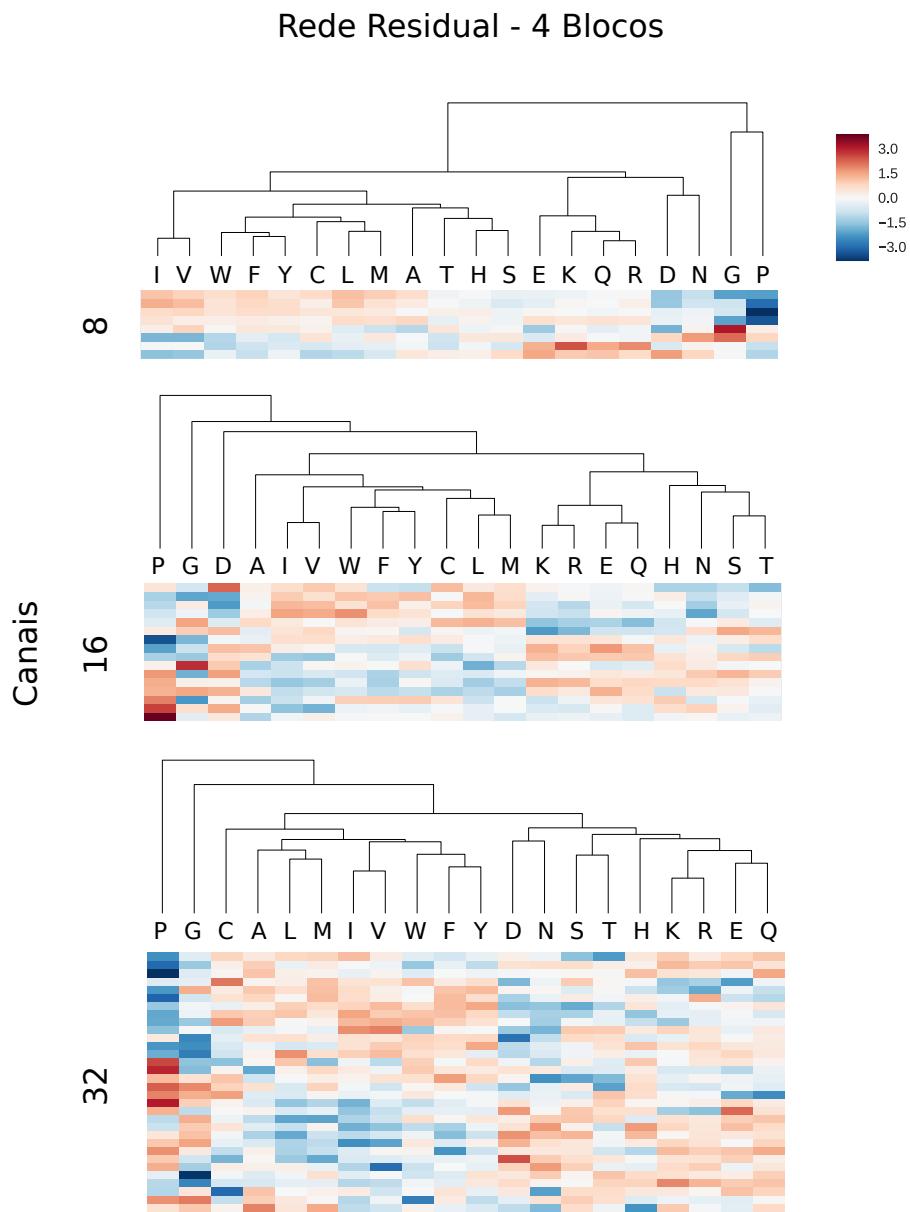


Figura 5.9: Agrupamento da representação dos aminoácidos aprendida pela primeira camada oculta da rede residual com 4 blocos.

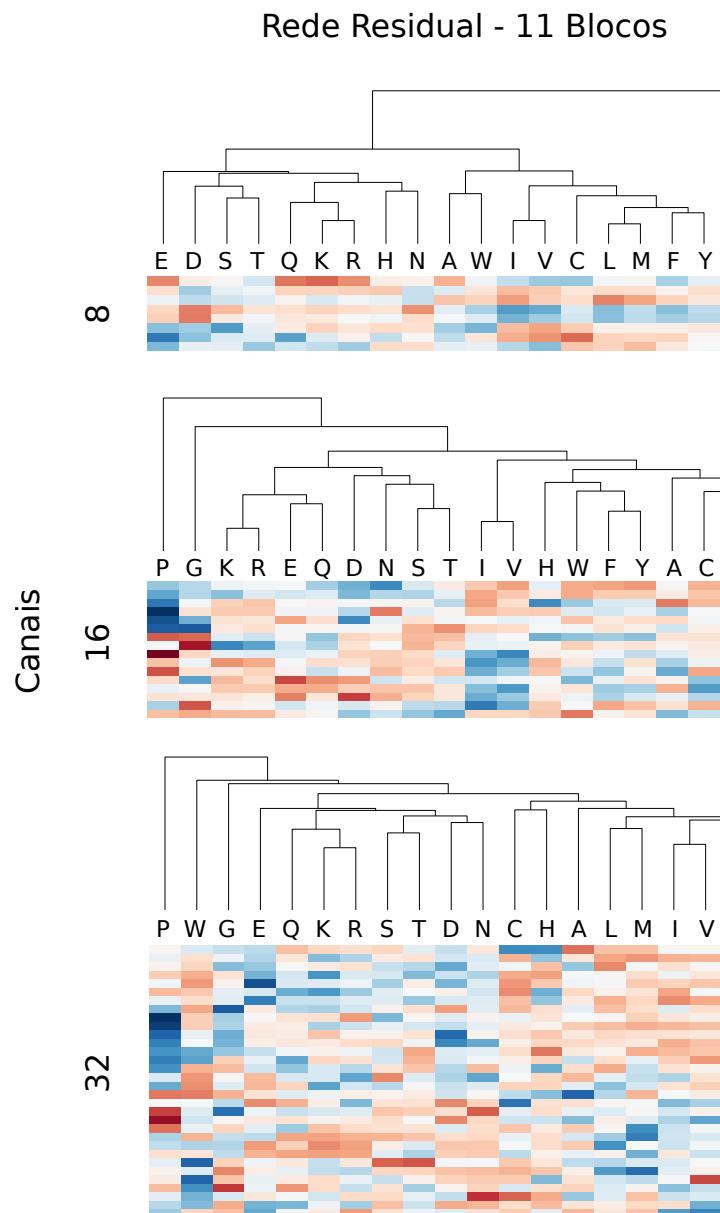


Figura 5.10: Agrupamento da representação dos aminoácidos aprendida pela primeira camada oculta da rede residual com 11 blocos.

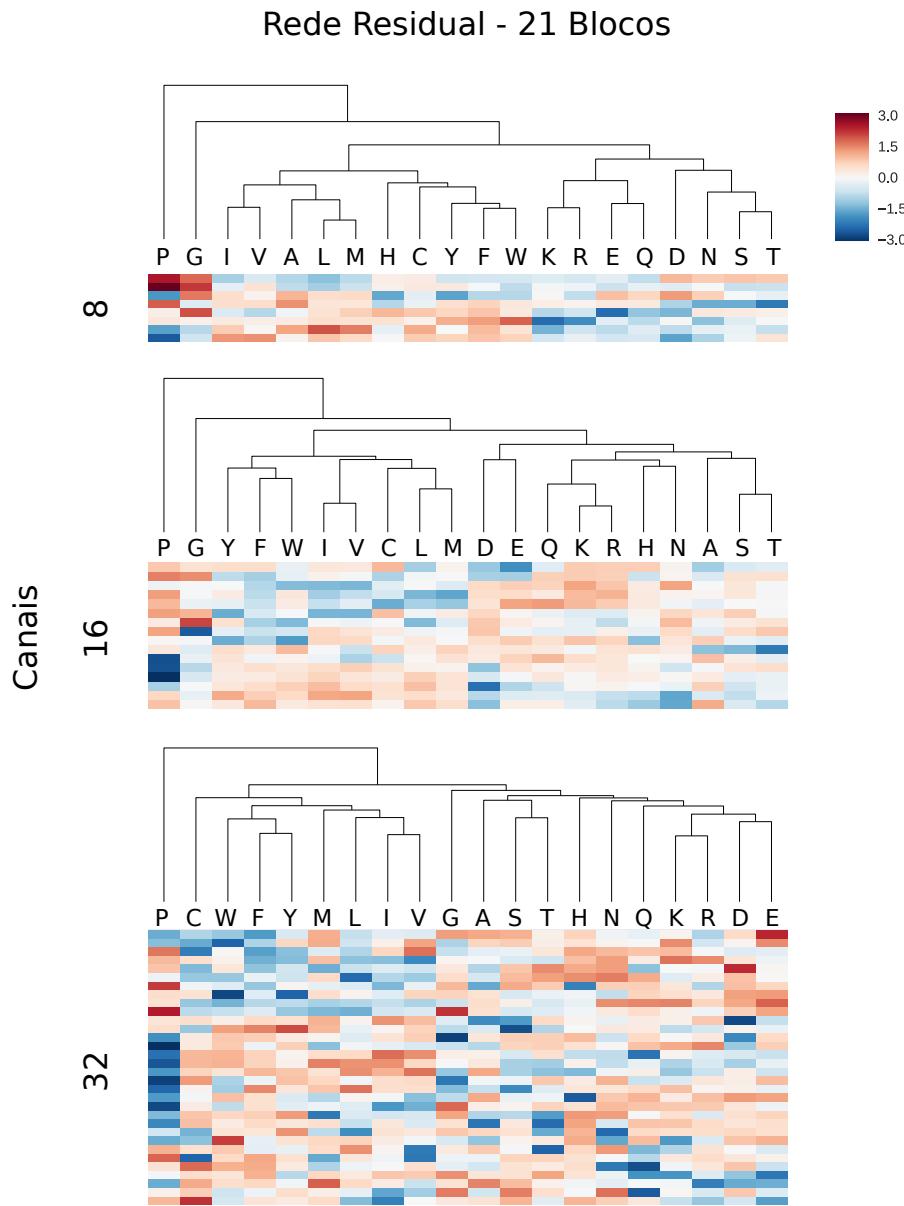


Figura 5.11: Agrupamento da representação dos aminoácidos aprendida pela primeira camada oculta da rede residual com 21 blocos.

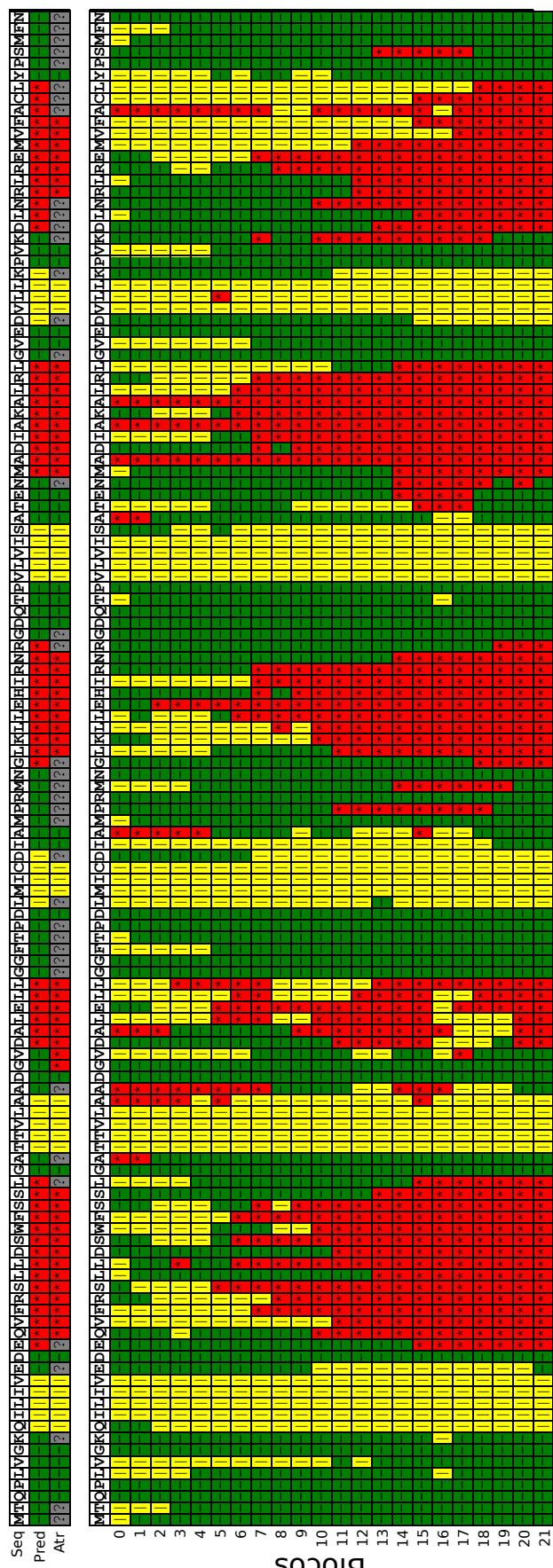


Figura 5.12: Estados internos da rede residual durante a predição do domínio N-terminal da proteína RssB (PDB ID: 3EOD). As três primeiras linhas correspondem, respectivamente, a sequência de resíduos, a estrutura secundária previda e a estrutura secundária atribuída. Nas demais linhas temos a estrutura secundária com maior probabilidade em cada um dos 21 blocos. Hélices são representadas em vermelho, fitas em amarelo e coils em verde.

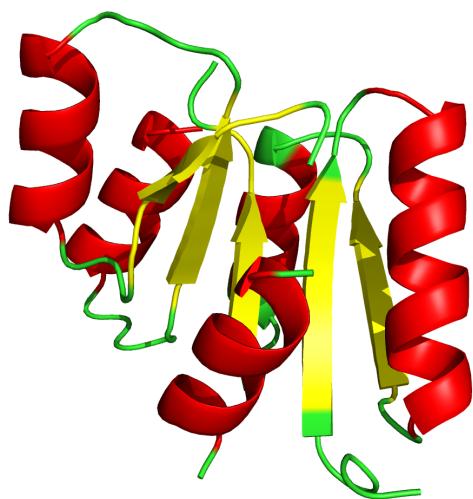


Figura 5.13: Domínio N-terminal da proteína RssB (PDB ID: 3EOD). As cores representam a estrutura secundária predita com maior probabilidade. Resíduos preditos como hélices são representadas em vermelho, fitas em amarelo e coils em verde.

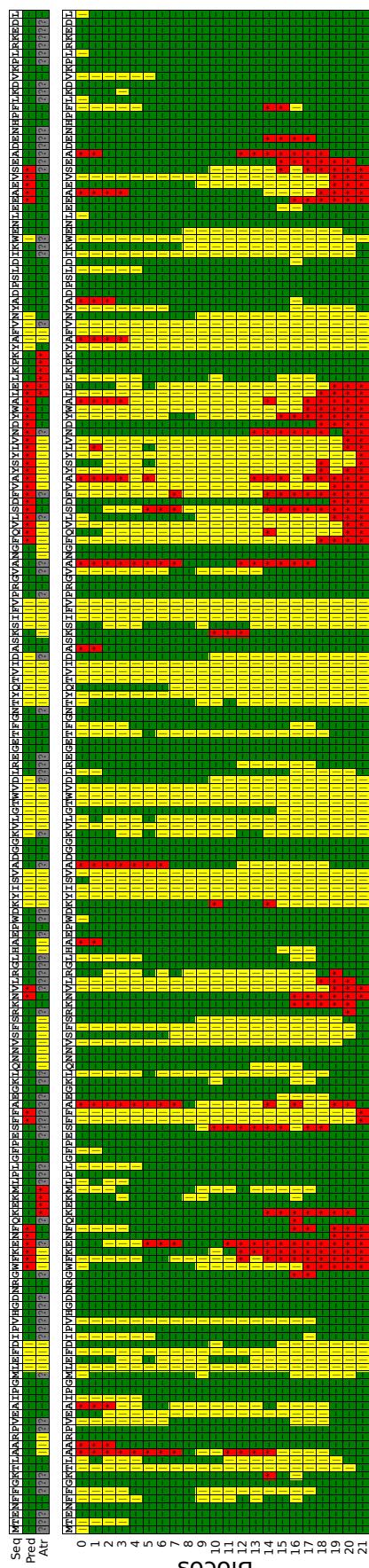


Figura 5.14: Estados internos da rede residual durante a predição da proteína RmlC de *Streptococcus suis* (PDB ID: 1NXM). As três primeiras linhas correspondem, respectivamente, a sequência de resíduos, a estrutura secundária predita e a estrutura secundária atribuída. Nas demais linhas temos a estrutura secundária com maior probabilidade em cada um dos 21 blocos. Hélices são representadas em vermelho, fitas em amarelo e coils em verde.

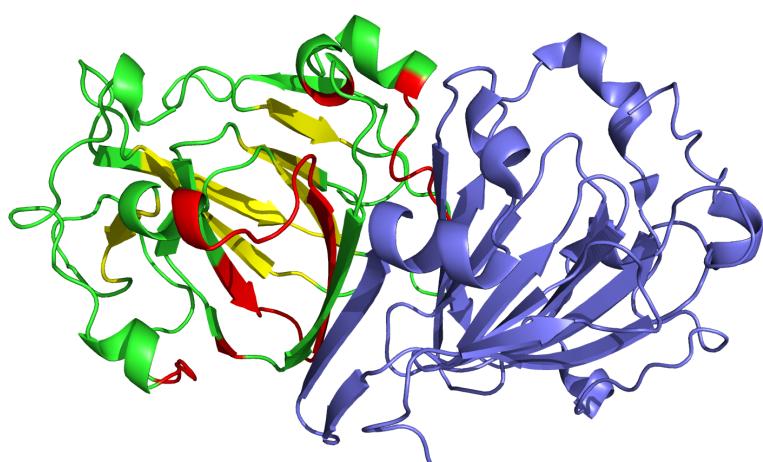


Figura 5.15: Proteína RmlC de *Streptococcus suis* (PDB ID: 1NXM). As cores representam a estrutura secundária predita com maior probabilidade. Resíduos preditos como hélices são representadas em vermelho, fitas em amarelo e coils em verde. Em azul é representada a outra cadeia do homodímero.

Parte III
CONCLUSÃO

6

DISCUSSÃO

6.1 MÉTODOS DE ATRIBUIÇÃO DE ESTRUTURA SECUNDÁRIA

Uma das etapas iniciais no desenvolvimento de métodos de predição de estrutura secundária é a preparação dos dados de referência para serem usados tanto no treinamento quanto no teste, onde a acurácia é calculada. Parte dessa preparação consiste na atribuição da estrutura secundária em proteínas com estrutura atômica resolvida. Em geral, trabalhos da literatura costumam utilizar o método DSSP como padrão. Entretanto, as diferenças observadas entre as atribuições de estrutura secundária efetuadas por diferentes métodos computacionais indicam que essa etapa pode ser importante para o desenvolvimento de preditores com maior acurácia.

Entre os quatro métodos de atribuição analisados, há a ocorrência de uma variação considerável entre as estruturas secundárias atribuídas a cada resíduo. Como esperado, a maior similaridade é observada entre os métodos DSSP e STRIDE, uma vez que o último é inspirado no primeiro [42, 44]. Entretanto, a similaridade entre os demais métodos encontra-se na faixa entre 83% e 85%. A porcentagem de resíduos que apresentaram consenso entre os quatro métodos foi de apenas 74.43%.

Ao longo deste trabalho nós exploramos a utilização dos dados de cada um dos quatro métodos citados e sua influência no treinamento de métodos de predição. Nossos resultados indicam que a utilização apenas de resíduos que apresentam consenso na atribuição resulta em preditores com acurácia maior ou, ao menos, semelhante ao melhor método de atribuição. Neste contexto, melhor método de atribuição significa o que produziu os dados que resultaram no preditor com maior acurácia.

6.2 MÉTODOS DE REFERÊNCIA PARA A PREDIÇÃO DE ESTRUTURA SECUNDÁRIA

Os resultados obtidos com a rede neural similar a de Holley e Karplus (HK), a qual realiza a predição a partir da sequência de resíduos, e o PSIPRED [59], que realiza a predição a partir da PSSM, indicam que o ganho na acurácia da predição ao se utilizar informações da conservação evolutiva pode ser de até 15%. A rede HK forneceu ainda uma acurácia de referência para compararmos com nossos modelos, os quais também realizam a predição a partir da sequência de resíduos.

O fato do treinamento das redes neurais similares a de Holley e Karplus (HK) utilizando somente os dados de resíduos que apresentaram consenso na atribuição da estrutura secundária ter apresentado acurácia similar ou superior em relação ao treinamento com dados de qualquer um dos quatro métodos de atribuição indicou que apenas o consenso pode ser utilizado para o treinamento de métodos de predição estrutura secundária.

A diferença observada no comprimento das estruturas atribuídas em relação às estruturas preditas pela rede neural HK, não foi observada para o PSIPRED [59]. Acreditamos que o fator responsável por produzir estruturas com comprimentos similares ao experimental seja a utilização de uma segunda rede neural. Essa segunda rede neural recebe como entrada três valores para cada resíduo da janela. Tais valores representam os três tipos de estruturas secundárias. Assim, a segunda rede neural não recebe informação direta dos aminoácidos, mas apenas das estruturas secundárias. Consequentemente, sua função é construir estruturas mais próximas as observadas experimentalmente, removendo ou conectando estruturas muito curtas como hélices com apenas um resíduo.

6.3 AUTÔMATOS CELULARES PARA A PREDIÇÃO DE ESTRUTURAS SECUNDÁRIAS

Ao propormos a utilização de autômatos celulares para a predição de estruturas secundárias nosso objetivo era observar indícios de como a estrutura secundária poderia se organizar a partir da sequência de aminoácidos. Entretanto, durante o desenvolvimento do método, diversos problemas foram sendo enfrentados, alguns dos quais, não haviam sido previstos por nós.

O primeiro obstáculo era a dificuldade em se resolver o problema inverso dos autômatos celulares. Esse desafio já era de nosso conhecimento por ser algo discutido na literatura. Ganguly e colaboradores [64] escreveram:

The inverse problem of deducing the local rules from a given global behavior is extremely difficult. There have been some efforts, with limited success, to build the attractor basin according to a given design specification (...) most popular methodology to address the inverse problem of mapping the global behavior to local CA rules are based on evolutionary computation techniques namely genetic algorithms and simulated annealing.

Esse problema foi enfrentado por nós com a implementação de um algoritmo de estimativa de distribuição (EDA) de forma distribuída, possibilitando a utilização de infraestrutura de computação de alto desempenho. Consequentemente, isso aumentou a eficiência no pro-

cesso de otimização das regras assim como possibilitou o aumento no número de proteínas em nosso conjunto de treinamento.

Outro problema observado foi a definição de estados do **AC**. Nossa abordagem inicial foi utilizar cada resíduo como um estado e adicionar quatro estados: um para hélice, um para fita, um para coil e mais um para início/fim da sequência. Acreditávamos que caso não obtivéssemos sucesso em encontrar regras com esses estados a origem do problema seria o enorme espaço de regras possíveis, o qual já havíamos calculado. Com isso, chegamos a planejar uma possível redução do número de estados, agrupando aminoácidos com características estruturais e físico-químicas similares em um mesmo estado.

Entretanto, durante o trabalho, notamos que os quatro estados adicionados não seriam suficientes. Ao observarmos que uma célula, ao evoluir de um estado representando um aminoácido para um estado que representasse uma estrutura secundária, perdia toda a informação sobre o resíduo originalmente naquela posição da sequência, percebemos que precisaríamos acrescentar mais estados.

A solução proposta para isso foi manter parcialmente a informação do resíduo através do uso do que chamamos “contexto”. Iniciamos com apenas dois contextos, polar e apolar. Acreditamos que esses dois contextos auxiliariam o **AC** na formação de hélices e fitas com características anfipáticas. Com o uso de contextos observamos uma melhora na capacidade preditiva do **AC**. Isso nos levou a explorar outras características dos resíduos como contextos. Primeiro, tornamos glicinas e prolinas dois novos contextos e então, adicionamos outros dois, para resíduos carregados, positivos e negativos.

Os contextos melhoraram a capacidade dos ACs na formação de estruturas secundárias, no entanto, também tornaram a busca por regras ainda mais difícil.

A definição de uma métrica para a acurácia e a forma de como calculá-la na evolução do autômato também nos trouxeram dificuldades. A melhor métrica testada foi o **CBA** (*Class Balance Accuracy*) [70], o qual foi desenvolvido para problemas com classes desbalanceadas, como é caso das estruturas secundárias. Em relação a forma de calculá-la para o **AC**, haviam diversas possibilidades, cada uma com vantagens e desvantagens. Poderíamos predefinir um número finito de evoluções do **AC** e calcular a acurácia na última geração. No entanto, observamos que o **AC** podia apresentar um padrão de oscilação entre estados durante a evolução. Testamos solucionar este problema calculando a acurácia em um intervalo de gerações até finalmente decidirmos utilizar todas as gerações do **AC**.

É possível que essa última decisão tenha conduzido o autômato celular a produzir padrões quase uniformes como os observados. Padrões onde as células evoluem para algum estado de estrutura secundária logo na primeira geração e, na maioria das vezes, permanecem até a última geração.

Apesar dos resultados não atingirem nossas expectativas, a motivação inicial de utilizar autômatos celulares para observarmos, dinamicamente, a predição das estruturas secundárias, nos levaram a buscar métodos que tivessem essa mesma capacidade. Encontramos esse potencial em redes neurais residuais profundas.

6.4 REDES NEURAIS RESIDUAIS

6.4.1 Representação do aminoácidos

Ao propormos a utilização de uma camada de neurônios de convolução 1×1 como primeira camada oculta da rede neural esperávamos que ela aprendesse a qual a melhor representação para os aminoácidos. Como essa representação estava relacionada apenas a tarefa de predizer estruturas secundárias, não sabíamos como os aminoácidos seriam agrupados. Interessantemente, o agrupamento dessas representações demonstrou que a rede é capaz de aprender similaridades estruturais e físico-químicas dos resíduos.

6.4.2 Configuração dos blocos da rede residual

A proposta original para as redes residuais utilizam uma função de ativação não linear do tipo Rectified Linear Unit (**ReLU**) [78] ao longo do caminho residual [74]. Dentro do bloco são utilizados, em sequência, uma convolução 3×3 , uma camada de *batch normalization* [79], um ReLU, outra convolução 3×3 e outra camada de *batch normalization* (Figura 6.1 a). Posteriormente He e colaboradores testaram diversas variações dos blocos e notaram que a alteração da ordem das camadas e a mudança da função de ativação do caminho residual para dentro do bloco melhoraram o desempenho da rede residual (Figura 6.1 b).

No entanto, alguns testes preliminares na fase de construção do nosso modelo demonstravam uma queda de desempenho com a utilização da camada de *batch normalization* (resultados não mostrados). Acreditamos que o motivo disso seja a influência das regiões com códigos que representam a ausência de resíduos possam ter no código dos aminoácidos e, uma vez que as cadeias polipeptídicas tem tamanhos diferentes para proteínas diferentes, essa influência possivelmente não é constante.

Por outro lado, a retirada da camada de *batch normalization* tornava as redes testadas mais suscetível ao sobreajuste (*overfitting*). A solução encontrada de regularização foi utilizar um camada de *dropout* [77] dentro do bloco, como utilizado por Zagoruyko e Komodakis nas *Wide Residual Networks* [75].

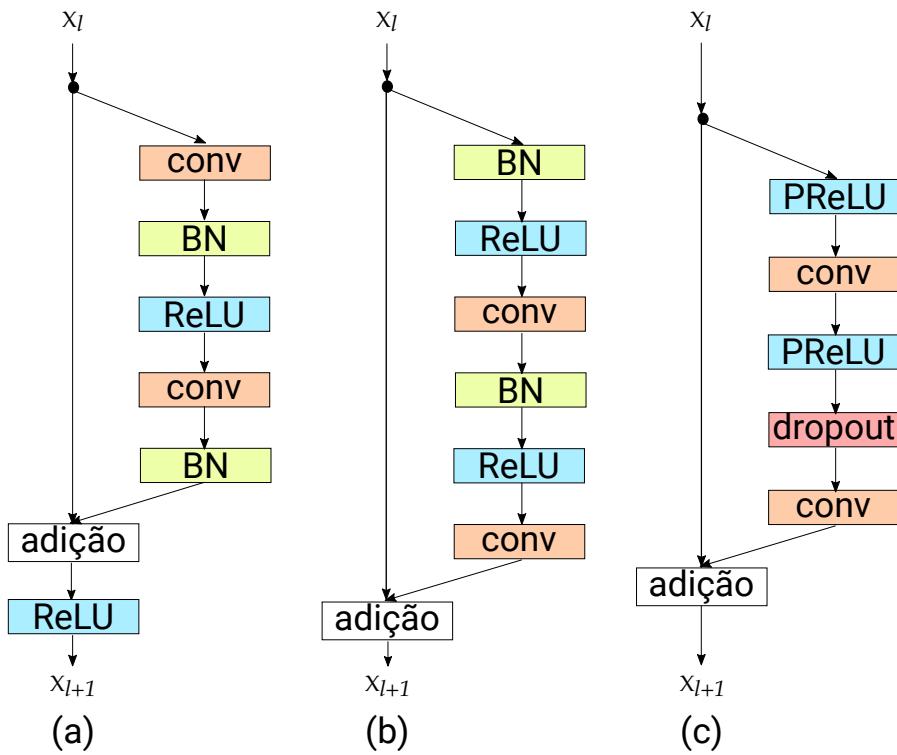


Figura 6.1: Configurações de blocos para redes residuais. (a) Configuração original proposta por He e colaboradores [74], (b) Variação também proposta por He e colaboradores [80], (c) Bloco proposto por Zagoruyko e Komodakis que substitui as duas camadas de *batch normalization* por uma de *dropout* [75].

6.4.3 Acurácia dos modelos de predição

Os modelos de predição de estruturas secundárias desenvolvidos neste trabalho, os quais utilizam apenas a informação contida na estrutura primária, apresentaram acurácia inferior aos modelos que utilizam PSSM. Entretanto, acreditamos que o modelo baseado em redes residuais apresenta propriedades que precisam ser consideradas durante a comparação.

Uma delas é o número de parâmetros do modelo. Os parâmetros são as variáveis que precisam ser otimizadas durante o processo de aprendizado ou treinamento. Nesse aspecto, a rede residual que apresentou a melhor acurácia ($b_{21}-c_{16}$) possui um número baixo de parâmetros quando comparado a outros modelos (Tabela 6.1). Outra rede residual, a b_4-c_{16} , devido ao uso de 4 blocos, utiliza para a predição da estrutura secundária de cada resíduo a informação equivalente a uma janela de 17 resíduos na sequência. Informação esta similar a utilizada pelas redes HK. Entretanto, com apenas ~6 parâmetros ela atinge uma acurácia similar as redes HK(32), com 32 neurônios na camada oculta e ~36 mil parâmetros, e HK(64) com ~72 mil parâmetro. Isso indica que nossa arquitetura de rede residual foi mais eficiente.

Outro fator relevante é a não utilização de PSSM como entrada. Como nosso objetivo sempre envolveu a construção de um modelo capaz de fornecer informações que pudessem ser relacionadas ao processo físico de formação das estruturas secundárias e, consequentemente, ao enovelamento, acreditamos que o modelo deveria se basear apenas na informação contida na estrutura primária da proteínas. No entanto, futuramente poderemos adaptar um modelo similar que utilize PSSM. Nesse ponto, é relevante mencionar que todos os 5 métodos testados por Wang e colaboradores [71] e que utilizam PSSM, entre eles o PSIPRED e a rede neural convolucional profunda desenvolvida por eles, tem uma perda de até 10% de acurácia quando o alinhamento para construção da PSSM contém poucas proteínas. Para esses casos, os 5 métodos apresentaram acurácia em torno de 74% [71].

Modelo	Entrada	Parâmetros	Informação sobre o processo de predição	Acurácia Q_3
Rede HK (32)	Sequência	~36 mil	Não	75,17%
Rede HK (64)	Sequência	~72 mil	Não	75,27%
PSIPRED	PSSM	~82 mil	Não	93%
Autômato Celular (4)	Sequência	~59 mil ^(a)	Sim	62,6%
ResNet (b4-c16)	Sequência	~6 mil	Sim	75,5%
ResNet (b21-c16)	Sequência	~32 mil	Sim	78,3%
DCNN [71]	PSSM	~500 mil	Não	~84% ^(b)

Tabela 6.1: Comparação de métodos de predição de estruturas secundárias.

(a) Consideramos como parâmetros do autômato celular (AC4) o número de elementos no conjunto de regras. Nos demais modelos os parâmetros correspondem ao número de pesos que são otimizados durante o treinamento. (b) A acurácia da rede neural convolucional profunda (DCNN) corresponde ao dado do artigo [71]. Para os demais casos, a acurácia foi calculada utilizando nossos dados e com os resíduos que apresentam consenso na atribuição da estrutura secundária.

7

CONCLUSÕES E PERSPECTIVAS

Neste trabalho testamos dois modelos de predição de estruturas secundárias com potenciais para fornecer informações detalhadas sobre o processo de predição a partir da sequência de aminoácidos das proteínas.

O modelo utilizando autômatos celulares apresentou menor acurácia em relação à outros métodos de predição de estruturas secundárias. O processo de predição dessas estruturas, observado durante a evolução do autômato, apresentou padrões de baixa complexidade, o que não é ideal para estudos da formação de estruturas proteicas. No entanto, acreditamos que uma melhor definição dos estados que representam estruturas secundárias e alterações no mecanismo de otimização das regras, o qual inclui uma melhor função de *fitness*, poderão produzir melhores resultados.

Por outro lado, o modelo desenvolvido utilizando redes neurais residuais profundas apresentou maior acurácia que outros métodos de predição a partir da estrutura primária. A extração de informações da rede residual durante o processo de predição indicou ser possível determinarmos quais regiões da sequência tem maior influência na predição de cada elemento da estrutura secundária. Ainda não temos evidências suficientes para afirmar que há uma relação significativa entre o processo de predição e o processo físico de formação dessas estruturas. Futuramente planejamos comparar dados experimentais sobre as etapas iniciais do enovelamento de proteínas com os resultados da predição utilizando nosso método baseado em redes neurais residuais profundas. Nesse sentido, a observarmos neste trabalho que as redes residuais aprenderam como agrupar aminoácidos e que esses agrupamentos tem relação com características estruturais e físico-químicas pode ser um ótimo indício de que a relação entre o processo de predição da rede residual possa ser relacionada ao processo de formação das estruturas proteicas.

REFERÊNCIAS BIBLIOGRÁFICAS

- (1) Gu, J. e Bourne, P. E., *Structural Bioinformatics*, Google-Books-ID: 4H_ai7ivRIcC; John Wiley & Sons: 2009; 1095 pp. (ver pp. 5, 52).
- (2) Pauling, L., Corey, R. B. e Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* 37, 205–211 (ver p. 5).
- (3) Pauling, L. e Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America* 37, 251–256 (ver p. 5).
- (4) Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry* 34, 167–339 (ver p. 5).
- (5) Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. e Phillips, D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* 181, 662 (ver p. 5).
- (6) Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)* 181, 223–230 (ver p. 5).
- (7) Rose, G. D., Fleming, P. J., Banavar, J. R. e Maritan, A. (2006). A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences* 103, 16623–16633 (ver pp. 6, 9).
- (8) Levinthal, C. (1968). Are There Pathways For Protein Folding? *Extrait du Journal de Chimie Physique* 65 (ver p. 6).
- (9) Levinthal, C. em, Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois, ed. por Debrunner, J. e Munck, E., University of Illinois Press: 1969, pp. 22–24 (ver pp. 6, 7).
- (10) Cellmer, T., Bratko, D., Prausnitz, J. M. e Blanch, H. (2005). Protein-folding landscapes in multichain systems. *Proceedings of the National Academy of Sciences* 102, 11692–11697 (ver p. 6).
- (11) Ben-Naim, A. (2012). Levinthal's question revisited, and answered. *Journal of Biomolecular Structure & Dynamics* 30, 113–124 (ver p. 7).
- (12) Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. e Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) — round x. *Proteins* 82, 1–6 (ver p. 7).

- (13) Baker, D. e Sali, A. (2001). Protein structure prediction and structural genomics. *Science (New York, N.Y.)* 294, 93–96 (ver p. 7).
- (14) Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society Interface* 5, 387–396 (ver pp. 7, 8).
- (15) Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* 9, 40 (ver p. 7).
- (16) Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. e Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology* 383, 66–93 (ver p. 7).
- (17) Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. e Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure* 29, 291–325 (ver p. 8).
- (18) Dunbrack, R. L. (2006). Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology* 16, 374–384 (ver p. 8).
- (19) Overington, J., Johnson, M. S., Sali, A. e Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proceedings. Biological Sciences* 241, 132–145 (ver p. 8).
- (20) Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544 (ver p. 8).
- (21) Coulson, A. F. W. e Moult, J. (2002). A unifold, mesofold, and superfold model of protein fold use. *Proteins* 46, 61–71 (ver p. 8).
- (22) Kolodny, R., Pereyaslavets, L., Samson, A. O. e Levitt, M. (2013). On the universe of protein folds. *Annual Review of Biophysics* 42, 559–582 (ver p. 8).
- (23) Giri, R., Morrone, A., Travaglini-Allocatelli, C., Jemth, P., Brunori, M. e Gianni, S. (2012). Folding pathways of proteins with increasing degree of sequence identities but different structure and function. *Proceedings of the National Academy of Sciences of the United States of America* 109, 17772–17776 (ver p. 9).
- (24) Dill, K. A. e Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology* 4, 10 (ver p. 9).
- (25) Dill, K. A., Ozkan, S. B., Shell, M. S. e Weikl, T. R. (2008). The Protein Folding Problem. *Annual review of biophysics* 37, 289–316 (ver p. 9).
- (26) Dill, K. A. e MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science (New York, N.Y.)* 338, 1042–1046 (ver p. 9).

- (27) Baldwin, R. L. e Rose, G. D. (1999). Is protein folding hierachic? II. Folding intermediates and transition states. *Trends in Biochemical Sciences* 24, 77–83 (ver p. 9).
- (28) Baldwin, R. L. e Rose, G. D. (1999). Is protein folding hierachic? I. Local structure and peptide folding. *Trends in Biochemical Sciences* 24, 26–33 (ver p. 9).
- (29) Daggett, V. e Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends in Biochemical Sciences* 28, 18–25 (ver p. 9).
- (30) Englander, S. W. e Mayne, L. (2017). The case for defined protein folding pathways. *Proceedings of the National Academy of Sciences* 114, 8253–8258 (ver p. 9).
- (31) Eaton, W. A. e Wolynes, P. G. (2017). Theory, simulations, and experiments show that proteins fold by multiple pathways. *Proceedings of the National Academy of Sciences* 114, E9759–E9760 (ver p. 9).
- (32) Englander, S. W. e Mayne, L. (2017). Reply to Eaton and Wolynes: How do proteins fold? *Proceedings of the National Academy of Sciences* 114, E9761–E9762 (ver p. 9).
- (33) Harper, E. T. e Rose, G. D. (1993). Helix stop signals in proteins and peptides: the capping box. *Biochemistry* 32, 7605–7609 (ver p. 9).
- (34) Aurora, R., Srinivasan, R. e Rose, G. D. (1994). Rules for alpha-helix termination by glycine. *Science (New York, N.Y.)* 264, 1126–1130 (ver p. 9).
- (35) Aurora, R., Creamer, T. P., Srinivasan, R. e Rose, G. D. (1997). Local interactions in protein folding: lessons from the alpha-helix. *The Journal of Biological Chemistry* 272, 1413–1416 (ver p. 9).
- (36) Colloc'h, N. e Cohen, F. E. (1991). Beta-breakers: an aperiodic secondary structure. *Journal of Molecular Biology* 221, 603–613 (ver p. 9).
- (37) Betancourt, M. R. e Skolnick, J. (2004). Local propensities and statistical potentials of backbone dihedral angles in proteins. *Journal of Molecular Biology* 342, 635–649 (ver p. 10).
- (38) Otaki, J. M., Tsutsumi, M., Gotoh, T. e Yamamoto, H. (2010). Secondary structure characterization based on amino acid composition and availability in proteins. *Journal of Chemical Information and Modeling* 50, 690–700 (ver p. 10).
- (39) Abagyan, R. e Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *Journal of Molecular Biology* 235, 983–1002 (ver p. 10).

- (40) Pedersen, J. T. e Moult, J. (1997). Protein folding simulations with genetic algorithms and a detailed molecular description. *Journal of Molecular Biology* 269, 240–259 (ver p. 10).
- (41) Srinivasan, R. e Rose, G. D. (1999). A physical basis for protein secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* 96, 14258–14263 (ver pp. 10, 18, 19).
- (42) Kabsch, W. e Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637 (ver pp. 13, 14, 16, 91).
- (43) Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. e Mornon, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Engineering* 6, 377–382 (ver p. 13).
- (44) Frishman, D. e Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23, 566–579 (ver pp. 13, 15, 91).
- (45) Zhang, Y. e Sagui, C. (2015). Secondary structure assignment for conformationally irregular peptides: comparison between DSSP, STRIDE and KAKSI. *Journal of Molecular Graphics & Modelling* 55, 72–84 (ver p. 13).
- (46) Martin, J., Letellier, G., Marin, A., Taly, J.-F., de Brevern, A. G. e Gibrat, J.-F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC structural biology* 5, 17 (ver p. 17).
- (47) Aurora, R. e Rose, G. D. (1998). Helix capping. *Protein Science: A Publication of the Protein Society* 7, 21–38 (ver p. 25).
- (48) Doig, A. J. e Baldwin, R. L. (1995). N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Science : A Publication of the Protein Society* 4, 1325–1336 (ver p. 25).
- (49) Farzadfar, F., Gharaei, N., Pezeshk, H. e Marashi, S.-A. (2008). Beta-sheet capping: signals that initiate and terminate beta-sheet formation. *Journal of Structural Biology* 161, 101–110 (ver p. 25).
- (50) Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. e Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389–3402 (ver pp. 29, 32, 33, 36).
- (51) McCulloch, W. S. e Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5, 115–133 (ver p. 29).
- (52) Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review*, 65–386 (ver p. 29).

- (53) Minsky, M. L. e Papert, S. A., *Perceptrons: Expanded Edition*; MIT Press: Cambridge, MA, USA, 1988 (ver p. 30).
- (54) Rumelhart, D. E., Hinton, G. E. e Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533 (ver p. 30).
- (55) Qian, N. e Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* 202, 865–884 (ver pp. 30, 32).
- (56) Sejnowski, T. e Rosenberg, C. (1987). Parallel networks that learn to pronounce {E}nglish text. *Complex Systems* 1, 145–168 (ver p. 30).
- (57) Holley, L. H. e Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the United States of America* 86, 152–156 (ver pp. 31, 32, 35).
- (58) Chandonia, J. M. e Karplus, M. (1996). The importance of larger data sets for protein secondary structure prediction with neural networks. *Protein Science : A Publication of the Protein Society* 5, 768–774 (ver pp. 31, 32).
- (59) Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292, 195–202 (ver pp. 32, 34, 36, 91, 92).
- (60) Henikoff, S. e Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* 89, 10915–10919 (ver p. 32).
- (61) Kingma, D. P. e Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (ver pp. 36, 71).
- (62) Mitchell, M., *Complexity: A Guided Tour*; Oxford University Press, Inc.: New York, NY, USA, 2009 (ver p. 47).
- (63) Wolfram, S. (1984). Universality and complexity in cellular automata. *Physica D: Nonlinear Phenomena* 10, 1–35 (ver p. 47).
- (64) Ganguly, N., Sikdar, B. K., Deutsch, A., Canright, G. e Chaudhuri, P. P. A Survey on Cellular Automata., 2003 (ver pp. 48, 53, 92).
- (65) *A New Kind of Science*; Wolfram Media Inc.: Champaign, Illinois, US, United States, 2002 (ver p. 48).
- (66) Chopra, P. e Bender, A. (2007). Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. *In Silico Biology* 7, 87–93 (ver pp. 48, 49).
- (67) Chou, P. Y. e Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry* 13, 222–245 (ver p. 49).

- (68) Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. e Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science (New York, N.Y.)* 229, 834–838 (ver p. 51).
- (69) Kilambi, K. P. e Gray, J. J. (2012). Rapid calculation of protein pKa values using Rosetta. *Biophysical Journal* 103, 587–595 (ver p. 52).
- (70) Mosley, L. A balanced approach to the multi-class imbalance problem., tese de dout., Iowa State University, 2013 (ver pp. 56, 93).
- (71) Wang, S., Peng, J., Ma, J. e Xu, J. (2016). Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports* 6, 18962 (ver pp. 65, 96, 97).
- (72) Mhaskar, H., Liao, Q. e Poggio, T. A. em *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.* 2017, pp. 2343–2349 (ver p. 65).
- (73) Goodfellow, I., Bengio, Y. e Courville, A., *Deep Learning*; The MIT Press: 2016 (ver p. 65).
- (74) He, K., Zhang, X., Ren, S. e Sun, J. em *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778 (ver pp. 65, 67, 94, 95).
- (75) Zagoruyko, S. e Komodakis, N. (2016). Wide Residual Networks. *arXiv:1605.07146 [cs]* (ver pp. 69, 94, 95).
- (76) He, K., Zhang, X., Ren, S. e Sun, J. em *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE Computer Society: Washington, DC, USA, 2015, pp. 1026–1034 (ver p. 69).
- (77) Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. e Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1929–1958 (ver pp. 69, 94).
- (78) Nair, V. e Hinton, G. E. em *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress: USA, 2010, pp. 807–814 (ver p. 94).
- (79) Ioffe, S. e Szegedy, C. em *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, JMLR.org: Lille, France, 2015, pp. 448–456 (ver p. 94).
- (80) He, K., Zhang, X., Ren, S. e Sun, J. (2016). Identity Mappings in Deep Residual Networks. *arXiv:1603.05027 [cs]* (ver p. 95).