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

Candidato: Gabriel Bianchin de Oliveira
Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Hélio Pedrini

Universidade Estadual de Campinas

17 de Outubro de 2019

Sumário

Sumário

Introdução

- Proteínas
- Sequenciamento de proteínas cada vez mais simples
- Alto custo na determinação de características
- Métodos laboratoriais e tecnologias de informação
- Protein Data Bank (PDB)

Sumário

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos Específicos

- Estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos Específicos

- Estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas
- Aquisição da base de dados PDB

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos Específicos

- Estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas
- Aquisição da base de dados PDB
- Proposição de métodos para a classificação de estruturas secundárias e de funções de proteínas

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos Específicos

- Estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas
- Aquisição da base de dados PDB
- Proposição de métodos para a classificação de estruturas secundárias e de funções de proteínas
- Realização de experimentos

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos Específicos

- Estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas
- Aquisição da base de dados PDB
- Proposição de métodos para a classificação de estruturas secundárias e de funções de proteínas
- Realização de experimentos
- Avaliação e comparação do método proposto com outras abordagens

Objetivos

Objetivo Geral

Desenvolver uma metodologia capaz de predizer estruturas secundárias e funções de proteínas

Objetivos Específicos

- Estudo das abordagens utilizadas para a predição de estruturas secundárias e funções de proteínas
- Aquisição da base de dados PDB
- Proposição de métodos para a classificação de estruturas secundárias e de funções de proteínas
- Realização de experimentos
- Avaliação e comparação do método proposto com outras abordagens
- Publicação dos resultados

Sumário

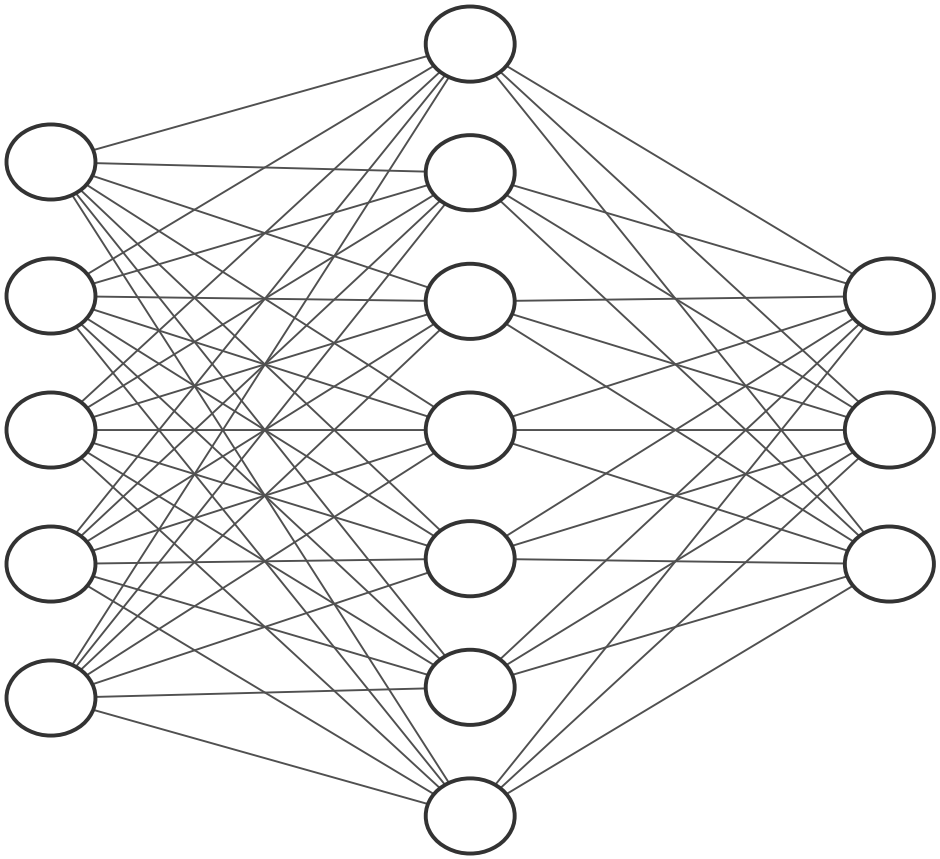
Estruturas Secundárias das Proteínas

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

Funções das Proteínas

- Funções definidas pelas estruturas 3D
- Proteínas com mesma origem evolutiva
- Sequência de aminoácidos, estruturas 3D e interação
- Ontologia Genética

Redes Neurais



Redes Neurais Recorrentes

- Análise de dados temporais ou espaciais
- Redes unidirecionais ou bidirecionais
- LSTM e GRU

Redes Neurais Convolucionais

- Informações locais e globais
- Camadas de convolução, *pooling* e totalmente conectadas

Trabalhos Relacionados

Predição de Estruturas Secundárias

- Lin et al. (2016) [1]
  - Rede convolucional com *shift-and-stitch*
  - Bases: CullPDB, CB513 e 4prot
- Hattori et al. (2017) [2]
  - Rede recorrente bidirecional com LSTM
  - Bases: CullPDB e CB513
- Fang et al. (2018) [3]
  - Rede convolucional com *inceptions*
  - Bases: CullPDB, JPRED, CASP10, CASP11, CASP12, CB513 e PDB

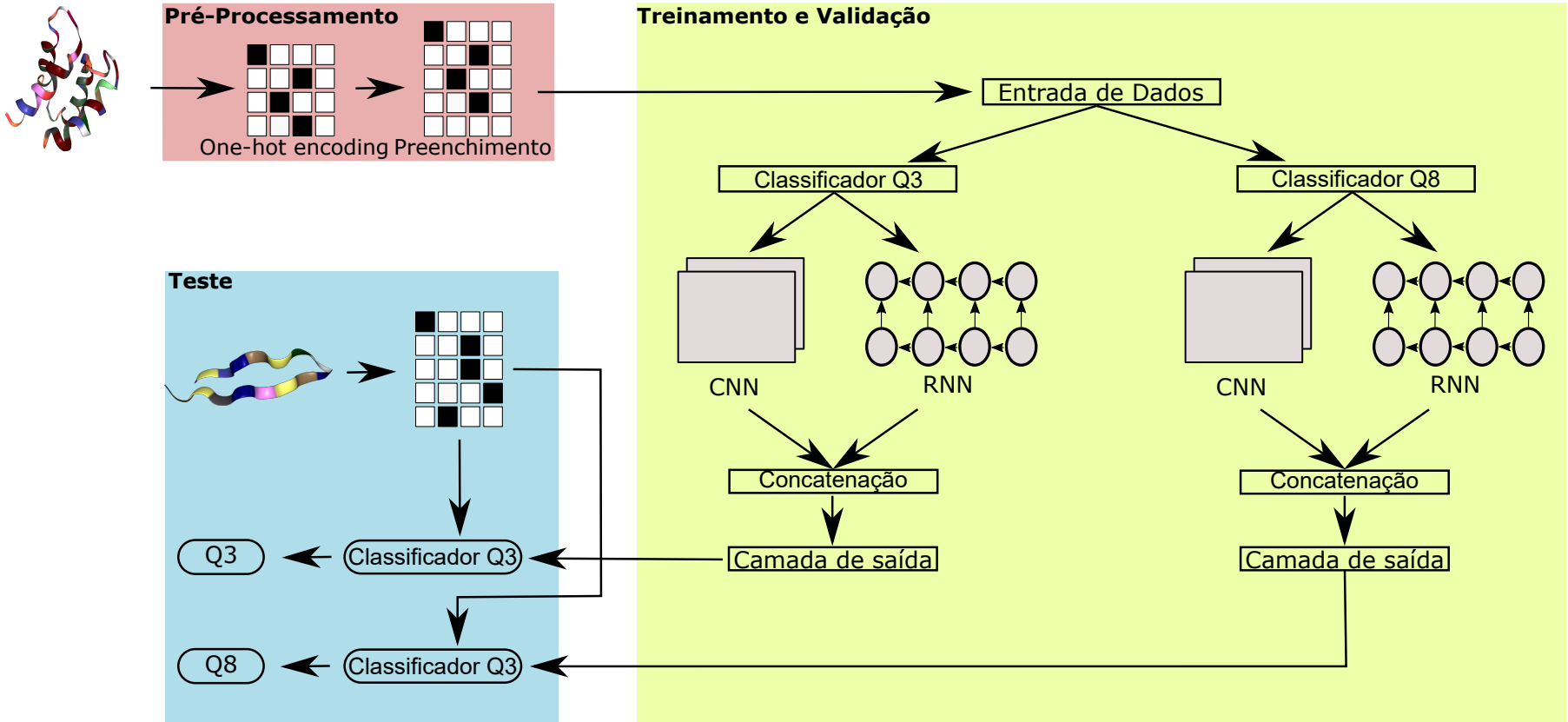
Trabalhos Relacionados

Predição de Funções

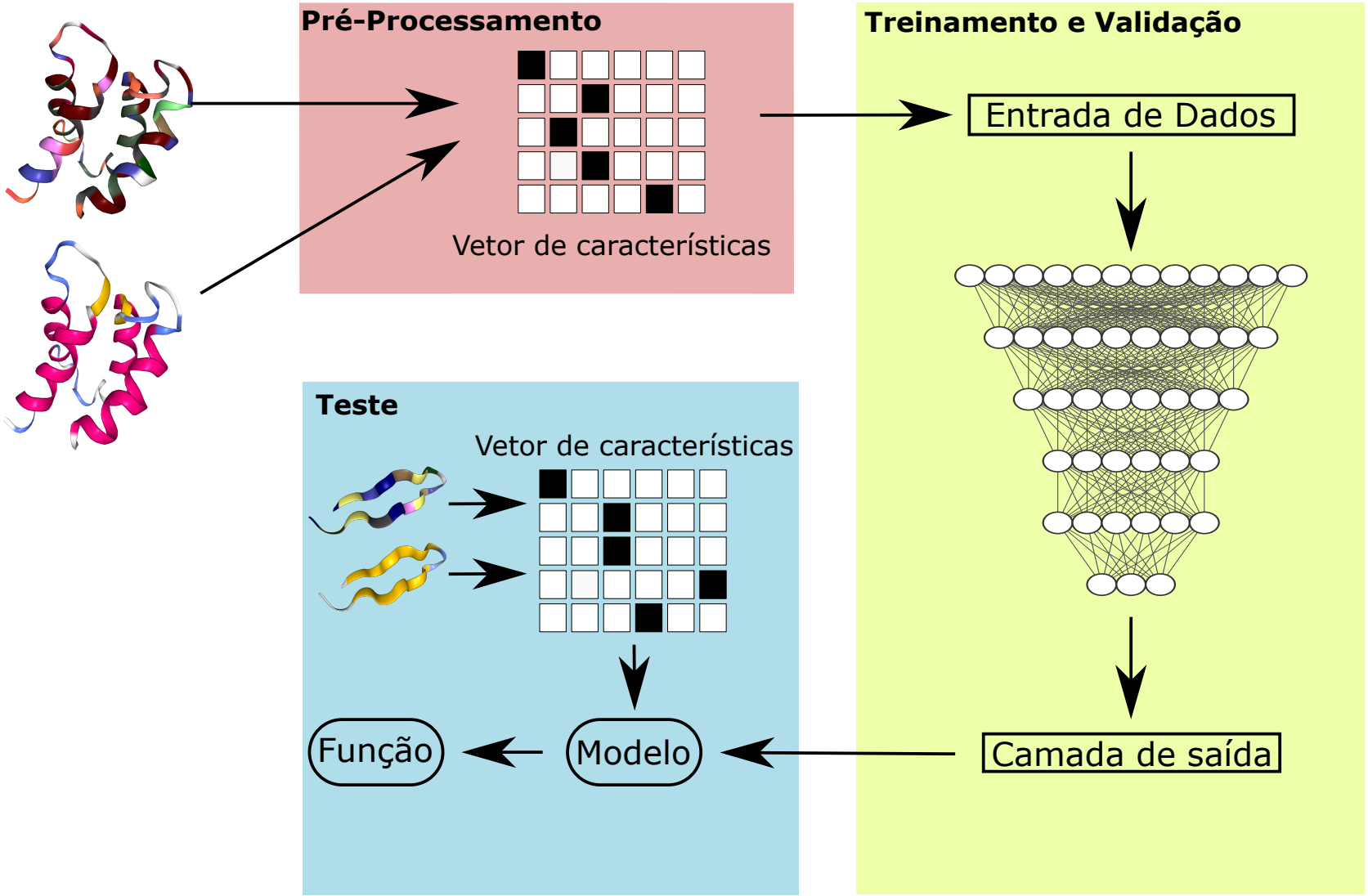
- Nadzirin e Firdaus-Raih (2012) [4]
  - Funções de proteínas do PDB
  - Anotações da base UniProtKB e ferramentas BLAST e DALI
- Roy et al. (2010) [5]
  - Modelo com proteínas com estruturas similares
  - Comparação entre a estrutura formada com proteínas com funções conhecidas do PDB
- Zhang et al. (2019) [6]
  - Rede densa com aprendizado profundo
  - Utiliza sequência de aminoácidos e interações entre proteínas
  - Bases: base derivada do trabalho de Kulmanov et al. [7] e CAFA3

Sumário

Metodologia – Predição de Estruturas Secundárias



Metodologia – Predição de Funções



Base de Dados – Protein Data Bank (PDB)

- Estruturas 3D das proteínas
- Atualização semanal
- Versão de 03/05/2019

Métricas de Avaliação

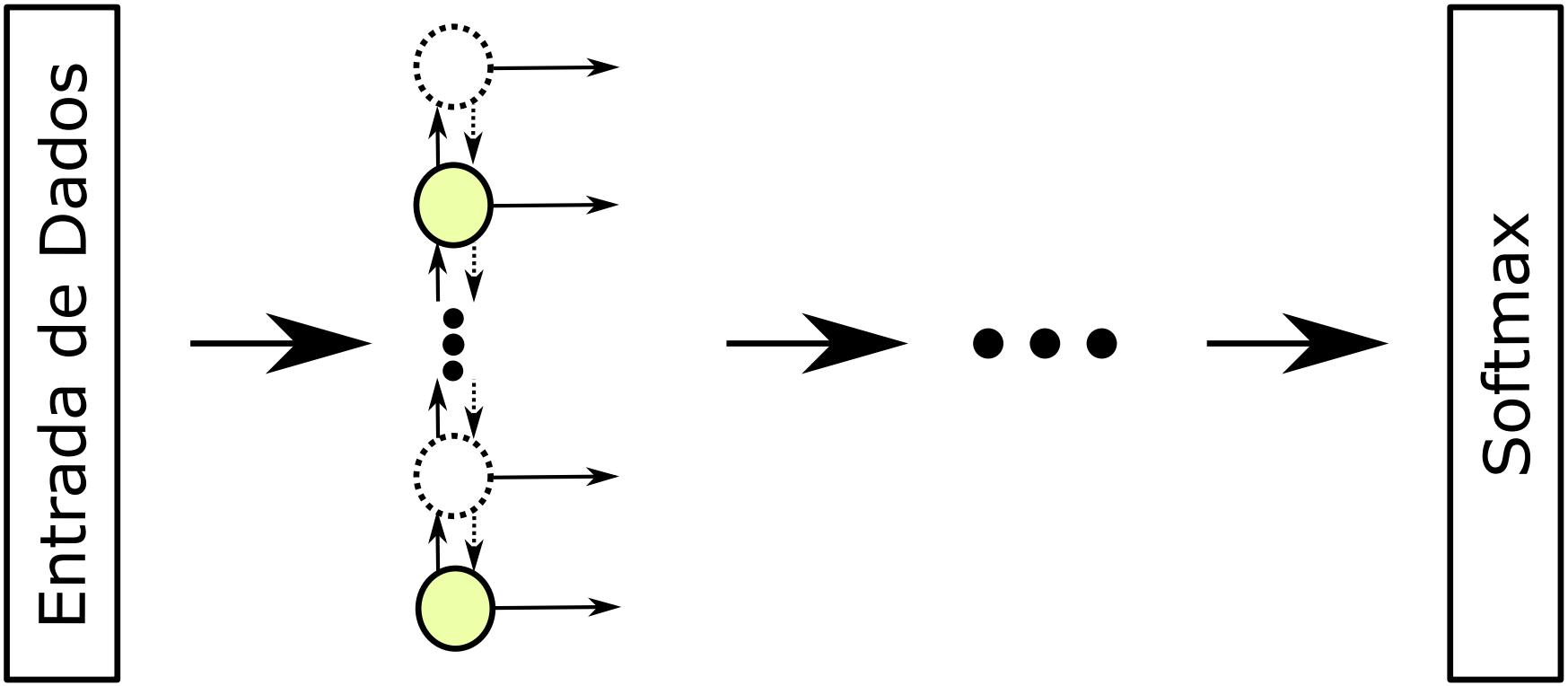
- Acurácia
- Precisão
- Revocação
- Taxa F1
- Métricas balanceadas

Sumário

Resultados Preliminares

- 2.000 proteínas do PDB
- Proteínas com até 700 aminoácidos
- Proteínas no sentido normal e invertido
- Redes unidirecionais, bidirecionais e combinação de redes
- Unidades LSTM e GRU
- De uma a dez camadas

Resultados Preliminares



Resultados Preliminares

Melhores Resultados Obtidos

- Combinação de redes bidirecionais
- Redes com 5 camadas para classificação Q3
- Redes com 4 camadas para classificação Q8

Q3	Acurácia	Precisão	Revocação	F1
5 camadas	0,71	0,73	0,73	0,72

Q8	Acurácia	Precisão	Revocação	F1
4 camadas	0,38	0,57	0,60	0,57

Resultados Preliminares

- GRU apresentou resultados melhores que LSTM
- Redes bidirecionais foram superiores às redes unidirecionais tanto na classificação Q3 quanto na Q8
- Combinação de classificadores resultou em melhores resultados

Dados das Estruturas Secundárias do PDB

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

Dados das Estruturas Secundárias do PDB

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

Sumário

Plano de Trabalho e Cronograma de Execução

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

Referências I

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

[2] Leandro Takeshi Hattori, Cesar Manuel Vargas Benitez, and Heitor Silverio Lopes.
A deep bidirectional long short-term memory approach applied to the protein secondary structure prediction problem.
In *4th IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE, 2017.

Referências II

[3] Chao Fang, Yi Shang, and Dong Xu.
MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction.
Proteins: Structure, Function, and Bioinformatics, 86(5):592–598, 2018.

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

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

Referências III

[6] Fuhao Zhang, Hong Song, Min Zeng, Yaohang Li, Lukasz Kurgan, and Min Li.
Deepfunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions.
Proteomics, pages 1–7, 2019.

[7] Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf.
DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier.
Bioinformatics, 34(4):660–668, 2017.

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

Candidato: Gabriel Bianchin de Oliveira
Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Hélio Pedrini

Universidade Estadual de Campinas

17 de Outubro de 2019