REPRESENTAÇÃO, CLASSIFICAÇÃO E INTERPRETAÇÃO DE SEQUÊNCIAS PROTEICAS DO VÍRUS DA DENGUE

LEONARDO RODRIGUES DE SOUZA

REPRESENTAÇÃO, CLASSIFICAÇÃO E INTERPRETAÇÃO DE SEQUÊNCIAS PROTEICAS DO VÍRUS DA DENGUE

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Juan Gabriel Colonna
Coorientador: Prof. Dr. Felipe Gomes Naveca

Manaus
Março de 2021

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Souza, Leonardo Rodrigues de
S729r Representação, classificação e interpretação de sequências proteicas do vírus da dengue / Leonardo Rodrigues de Souza . 2021
93 f.: il. color; 31 cm.

Orientador: Juan Gabriel Colonna
Coorientador: Felipe Gomes Naveca
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Dengue. 2. Proteínas. 3. Matriz de Co-ocorrência. 4. Classificação. 5. Interpretação. I. Colonna, Juan Gabriel. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Representação, classificação e interpretação de sequências proteicas do vírus da dengue"

LEONARDO RODRIGUES DE SOUZA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Juan Gabriel Colonna - PRESIDENTE

Prof. Eduardo Freire Nakamura - MEMBRO INTERNO

Profa. Elloá Barreto Guedes da Costa - MEMBRO EXTERNO

Manaus, 26 de Março de 2021

Agradecimentos

Ao meu orientador prof. Dr. Juan Gabriel Colonna que, além de colaborar com suas indispensáveis pontuações e ideias, auxiliou no meu desenvolvimento como aluno e pesquisador.

Ao meu co-orientador prof. Dr. Felipe Gomes Naveca que apresentou o problema deste trabalho e contribuiu de forma imprescindível na minha jornada de aprendizado sobre sequências biológicas e o vírus da dengue.

A minha colega de pesquisa Joseana Mendes Comodaro que esclareceu diversas dúvidas sobre temas relacionados a genética e também colaborou com a coleta dos dados utilizados no presente trabalho.

A todos meus professores de disciplina de pós graduação por apresentarem o conhecimento necessário para o desenvolvimento dessa pesquisa.

A secretaria e co-ordenação do curso por sempre estarem disponíveis para solucionar problemas pontuais da pós-graduação.

Aos amigos de pós-graduação pelo incentivo e diálogos frutíferos sobre o tema dessa pesquisa e abordagem proposta.

A minha família pelo apoio e força para seguir em frente no curso de pós graduação.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) pelo financiamento do presente trabalho - Código de Financiamento 001.

A Samsung Electrônica da Amazônia Ltda pelo financiamento parcial do presente trabalho através de convênio nº 003/2019, firmado com o ICOMP/UFAM, conforme previsto no Art. 48 do decreto nº 6.008/2006 nos termos da Lei Federal nº 8.387/1991.

Resumo

Introdução: O vírus da dengue é responsável por causar uma infecção muito comum em alguns países da América Latina e do Oeste do Pacífico, desencadeando diversos sintomas, tais como, febre, dor de cabeça, náuseas, vômitos e dores musculares. Os níveis da infecção podem ser divididos em: febre, febre hemorrágica e síndrome de choque, sendo os dois últimos casos associados a fatalidades. As causas que levam os hospedeiros a desenvolverem casos graves da infecção não são completamente conhecidas. No entanto, as proteínas que constituem o material genético do vírus da dengue são uma potencial fonte para extração de informação, um exemplo disso são as características presentes nessas que permitem diferenciar o vírus entre subclasses de sorotipos e genótipos, além de conter informações filogenéticas. Portanto, é aceitável assumir que essas estruturas guardem características capazes de elevar a compreensão sobre a dengue severa.

Métodos: O desafio de trabalhar com proteínas é a dificuldade de capturar características de interesse, visto que estas ocorrem na forma de padrões em pequenas regiões funcionais espalhadas dentro da sequência. Diante disso, representações de proteínas em estruturas onde padrões possam ser facilmente acessados passa a ser uma alternativa viável para o tratamento de dados deste tipo. Nesta pesquisa, propomos uma metodologia para identificar padrões em proteínas da dengue associados a dengue severa em hospedeiros humanos. O método baseia-se na representação de proteínas da dengue em matrizes de co-ocorrências de códons. Os algoritmos *Random Forests* (RF) e *Convolutional Neuural Network* (CNN) são empregados na classificação das matrizes rotuladas como dengue clássica/severa. Posteriormente, os classificadores são interpretados pelo método *SHAP Values* que, por sua vez, evidência quais co-ocorrências aumentam a probabilidade de dengue severa na amostra. Os resultados das interpretações são agrupados em gráficos de importância que permitem evidenciar os padrões de co-ocorrência de códons associadas a dengue severa.

Resultados: Classificamos de forma independente cada uma das dez proteínas da dengue. Os experimentos utilizando a RF alcançaram resultados AUC que variam entre

0.70 e 0.83. Os melhores resultados foram obtidos a partir da classificação de matrizes da proteína E em 25 resultados (cinco experimentos com cinco *folds* de validação cruzada cada), atingindo um AUC de 0.83 ± 0.02 com 95% de intervalo de confiança. Os testes estatísticos de *Levene*, *Shapiro-Wilk*, *ANOVA* e *Tukey* foram utilizados para testar se as médias das métricas calculadas nos 25 resultados eram diferentes entre as proteínas, com isso, constatou-se que os resultados da proteína E são estatisticamente distintos dos resultados das outras proteínas, dando indícios de que a proteína E caracteriza melhor a dengue severa.

Conclusão: Por meio do método proposto, conseguimos novas evidências sobre o desenvolvimento da dengue severa, associando-a diretamente a padrões frequentes de co-ocorrência de códons. Nosso método permitiu encontrar a existência de co-ocorrências elevadas na proteína E que podem estar associadas ao desencadeamento da dengue severa no hospedeiro. Além disso, em explorações mais granulares, observamos grupos de co-ocorrências que aumentam a probabilidade de dengue severa para os distintos sorotipos. Esses resultados podem desempenhar um papel importante na proposta de novos tratamentos, assim como ser alvo de debate sobre novas teorias referentes ao desenvolvimento de dengue severa em hospedeiros humanos.

Palavras-chave: Dengue, Proteínas, Matriz de Co-ocorrência, Classificação, Interpretação.

Abstract

Introduction: The dengue virus is responsible for causing a very common infection in some Latin America and the Western Pacific countries, triggering several symptoms, such as fever, headache, nausea, vomiting and muscle pain. The infection levels can be divided into: fever, hemorrhagic fever and shock syndrome, the last two cases being associated with fatalities. The causes that lead hosts to develop severe infection cases are not completely known. However, the proteins that make up the dengue virus genetic material are a potential source for extracting information, an example of which are the characteristics present in those that allow differentiating the virus between serotypes and genotypes subclasses, in addition to containing phylogenetic information. Therefore, it is acceptable to assume that these structures have characteristics capable of raising the severe dengue understanding.

Methods: The challenge of working with proteins is the difficulty of capturing interest characteristics, since they occur in patterns forms in small functional regions scattered in sequence. Therefore, proteins representations in structures where patterns can be easily accessed becomes a viable alternative for data treatment of this type. In this research, we propose a methodology to identify patterns in dengue proteins associated with severe dengue in human hosts. The method is based on dengue proteins codon co-occurrence matrices representation. The *Random Forests* (RF) and *Convolutional Neuural Network* (CNN) algorithms are used to classify matrices labeled as classic/severe dengue. Subsequently, the classifiers are interpreted by *SHAP Values* method, which, in turn, shows which co-occurrences increase severe dengue probability in the sample. The interpretations results are grouped into importance plots that make it possible to highlight the codon co-occurrence patterns associated with severe dengue.

Results: We independently classify each dengue proteins. Experiments using RF achieved AUC results ranging from 0.70 to 0.83. The best results were obtained from the protein E matrices classification in 25 results (five experiments with five cross-validation *folds* each), reaching an AUC of 0.83 ± 0.02 with 95% interval trust. The statistical tests of *Levene*, *Shapiro-Wilk*, *ANOVA* and *Tukey* were used to test whether

the metrics averages calculated in the 25 results were different between proteins, thus, it was found that the results of protein E are statistically different from other proteins results, giving evidence that protein E best characterizes severe dengue.
Conclusion: Through the proposed method, we obtained new evidence on severe dengue development, directly associating it with frequent codon co-occurrence patterns. Our method made it possible to find the existence of high co-occurrences in protein E that may be associated with the severe dengue onset in the host. In addition, in more granular explorations, we observed co-occurrences groups that increase the severe dengue likelihood for those different four serotypes. These results may play an important role in proposing new treatments, as well as being the subject of debate on new theories regarding the development of severe dengue in human hosts.

Keywords: Severe dengue, proteins, co-occurrence matrices, Classification, Interpretation.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A dengue é um patógeno viral com incidência global transmitido por mosquitos *Aedes Aegypti*. O vírus desenvolve no hospedeiro humano uma infecção viral que em sua forma clássica, causa síndrome febril sem risco clínico elevado, acompanhada de dores de cabeça, olhos, músculos e articulações, náusea, vomito e erupções cutâneas. Entretanto, casos de dengue severa em que o hospedeiro desenvolve febre hemorrágica ou síndrome de choque ocorrem com determinada frequência. Pacientes com quadros severos de dengue podem apresentar dificuldade respiratória, sangramentos graves, dores abdominais severas, vômitos frequentes, retenção de líquido e fadiga. Essa combinação de sintomas torna a dengue severa potencialmente fatal. A identificação antecipada da infecção aliada a tratamentos adequados podem reduzir as chances de fatalidade [Shope and Meegan, 1997, WHO, 2009, 2011].

Estima-se que cerca de 390 milhões de pessoas sejam infectadas por dengue todos os anos ao redor do mundo, desses, aproximadamente 96 milhões desenvolvem casos com algum nível de gravidade clínica, resultando em uma estimativa anual de 25 mil mortes [Bhatt et al., 2013]. No Brasil, embora sejam realizadas campanhas anuais pelo Governo Federal para prevenção contra dengue, que promovem o combate ao mosquito *Aedes Aegypti* e melhoria de condições de saneamento, observam-se níveis elevados da infecção. O Ministério da Saúde do Brasil calculou quase 1 milhão de casos de dengue no ano de 2020 [CNN Brasil, 2020].

O vírus da dengue possui quatro sorotipos que geram respostas imunes diferentes no organismo infectado. Dessa forma, torna-se uma infecção reincidente, pois o processo de defesa do organismo infectado criará anticorpos capazes de combater somente os vírus de um sorotipo, permanecendo vulnerável a infecções causadas por outros sorotipos. Atualmente, a infecção viral da dengue é tratada através de remédios para combater os sintomas causados pela infecção.

A estrutura genética do vírus é composta por uma sequência de RNA simples que transcrevem 10 proteínas (C, M, E, NS1, NS2A, NS2B, NS3, NS4 e NS5) que integram o vírion (Figura 1.1) e que podem ser interpretadas como subsequências do RNA completo. Cada proteína é responsável por uma tarefa específica. A proteína E é responsável pelo reconhecimento e entrada na célula a ser infectada [Kuhn et al., 2002], enquanto que a proteína NS1 se encarrega da replicação do RNA e ajuda na formação de imunocomplexos [Mackenzie et al., 1998, Avirutnan et al., 2006]. A proteína NS2A é importante para patogênese viral, enquanto que as proteínas NS2B e NS3 desempenham um papel importante na protease viral [Chambers et al., 1989, Clum et al., 1997, Xie et al., 2013]. A proteína NS4A está associada à proteína M através de regiões internas e executa o rearranjo da membrana [Miller et al., 2007]. Por fim, a proteína NS5 burla o sistema de resposta imune inata do organismo infectado e também auxilia na formação da capa de RNA [Ray et al., 2006, Laurent-Rolle et al., 2010].

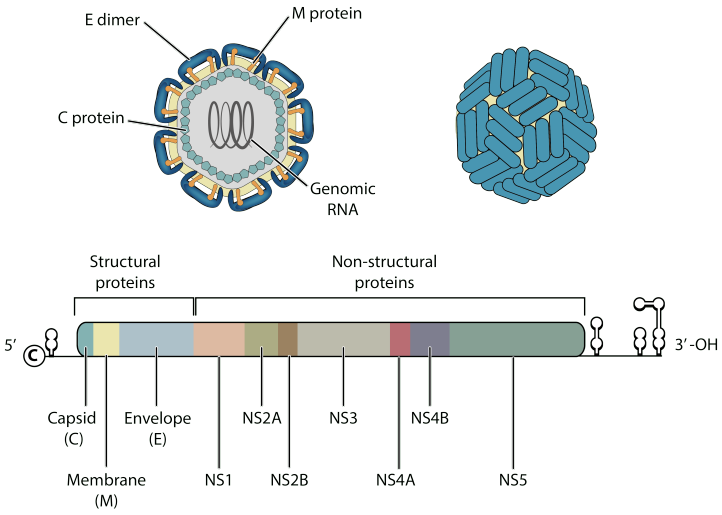


Figura 1.1: **Acima** Vírion da dengue com suas regiões especificadas. **Abaixo** RNA viral completo com cores únicas para cada proteína. As regiões não codificantes 5'UTR e 3'UTR também podem ser observadas.

Apesar de existirem teorias [Halstead, 1970, Rosen, 1977, Palacios Serrano et al., 2001] que tentam explicar o desenvolvimento de dengue severa e geram especulações sobre o assunto, as causas que levam hospedeiros humanos a esse acontecimento ainda são desconhecidas. Portanto, pesquisas que desempenhem o papel de elevar o conhecimento atual sobre os processos que causam a dengue severa são de importância científica e interesse social. O RNA da dengue apresenta-se como uma fonte potencial de informação para exploração de funções biológicas, padrões genéticos e particularidades existentes nos vírus que possam estar associadas a dengue severa.

Diante disso, este trabalho é motivado pela necessidade da elaboração de um método capaz de representar sequências proteicas da dengue, tornando-as utilizáveis por métodos de aprendizagem de máquina, para que características estruturais capazes de descrever dengue severa possam ser observadas. De maneira geral, espera-se que o método atenda esses requisitos e que também possa ser utilizado em outros problemas que envolvam representação e classificação de sequências biológicas.

1.1 Motivação

Segundo o *World Mosquitoe Program* (WMP), o vírus da dengue pode estar presente em qualquer lugar que o mosquito *Aedes Aegypti* exista [WMP, 2020]. O programa estima que aproximadamente 141 países ao redor do mundo são afetados pelo vírus e mais de 40% da população mundial possui risco de ser infectada. De acordo com relatórios fornecidos pelo WMP, todos os anos cerca de 390 milhões de pessoas são infectadas pelo vírus, das quais 500 mil apresentam um caso agravado da infecção. Do total de casos agravados estima-se que 25 mil pessoas morrem todos os anos por conta de complicações da dengue severa. Adicionalmente, a falta de um tratamento específico resulta em diversos tratamentos voltados para amenizar os sintomas e manter o volume de fluidos corporais, além da prevenção de surtos através do controle do mosquito transmissor, o que pode não ser eficiente em alguns cenários, aumentando ainda mais a gravidade do problema.

Dados públicos revelam que em 2019 o Brasil passou pela segunda maior epidemia de dengue já registrada [Fiocruz, 2020]. Apenas em 2019, a dengue ocasionou a morte de 754 pessoas, enquanto que em 2015 foram registrados 986 casos fatais. Esses resultados são os piores desde 1998, quando os números de mortes por dengue começaram a ser registrados [Folha UOL, 2020]. Uma das principais preocupações está associada a capacidade de reinfecção pelo vírus, pois, acredita-se que a infecção prévia torne os hospedeiros propensos a desenvolverem quadros graves de dengue em infecções posteriores [Singhi et al., 2007, Reich et al., 2013].

Estudos clínicos e estatísticos tentam correlacionar características genéticas de determinados vírus da dengue com o grau da infecção no hospedeiro [Halstead, 1970, Rosen, 1977, Vaughn et al., 2000, Palacios Serrano et al., 2001, Halsey et al., 2012]. No entanto, de acordo com a nossa revisão de literatura, não existem pesquisas que abordam proteínas da dengue, tampouco trabalhos que empregam técnicas de representação, classificação e interpretação oriundas do campo de aprendizagem de máquina. Diante disso, pesquisas que sejam capazes de compreender a estrutura desses organismos

e identificar previamente casos graves de dengue antes do desenvolvimento completo da infecção são de interesse médico, por auxiliar em discussões sobre novos tratamentos, científico, por disponibilizar métodos que podem ser aplicados para desdobrar interpretações sobre outros tipos de doenças e governamental, por ser uma pesquisa com tema base uma infecção que afeta drasticamente a população brasileira todos os anos, resultando em despesas governamentais exacerbadas que podem ser amenizadas com pesquisas desse cunho.

1.2 Definição do problema

A extração de informações, relações e funções presentes em proteínas auxiliam diversos campos de estudo composto por múltiplas tarefas, incluindo problemas como classificação de famílias, alinhamento, modelagem tridimensional e previsão de estrutura sequencial [Asgari, 2015, Korf et al., 2003, Brunk et al., 2018, Wang et al., 2016]. Não obstante, essas informações também podem indicar funções biológicas que desempenham um papel específico para o comportamento de um organismo. Por exemplo, algumas funções e mutações genéticas podem fazer que alguns vírus de uma mesma classe sejam mais agressivos ao hospedeiro [Kim et al., 2016a,b, Yang et al., 2017].

A dengue tem seu RNA representado por uma cadeia de caracteres retirados de um alfabeto específico conhecido como tabela IUPAC (*International Union of Pure and Applied Chemistry*) [IUPAC-IUB, 1970]. Sendo assim, o RNA da dengue possui aproximadamente 10.700 caracteres (nucleotídeos) que transcrevem três proteínas estruturais: C, M e E; e sete proteínas não estruturais: NS1, NS2A, NS2B, NS3, NS4A, NS4B e NS5 [Mackenzie et al., 1998, Kuhn et al., 2002]. O RNA também possui as regiões reguladoras 5'UTR e 3'UTR que não traduzem proteínas [Mackenzie et al., 2004, Perera and Kuhn, 2008]. Tarefas específicas dos processos virais são atribuídos a essas proteínas, tais como replicação viral e defesa contra resposta imune do hospedeiro. Em uma ótica mais granular, cada proteína é formada por subsequências de nucleotídeos de tamanhos iguais a 3, essas subsequências traduzem um aminoácido e são denominadas códons. Portanto, excluindo as regiões reguladoras, uma sequência de RNA completo da dengue possui pouco mais que 3300 códons, conforme a Figura 1.2.

Apesar dos vírus da dengue apresentarem regularidade estrutural em suas sequências de RNA, assim como qualquer organismo biológico seus materiais genéticos também possuem variações sutis que os tornam distintos uns dos outros, algumas dessas podendo ser tão significantes a ponto de gerar grupos de sorotipos e genótipos. Diante disso, podemos levantar a hipótese da existência de variações no RNA do vírus capazes

de caracterizar o grau de agressividade ao organismo infectado [Westaway and Blok, 1997, Lindenbach et al., 2013].

Portanto, primeiramente desejamos classificar amostras de cada proteína da dengue de acordo com sua severidade para que, por fim, possamos explorar separadamente os resultados obtidos para cada proteína à procura de padrões de códons que se destaquem em amostras associadas a dengue severa.

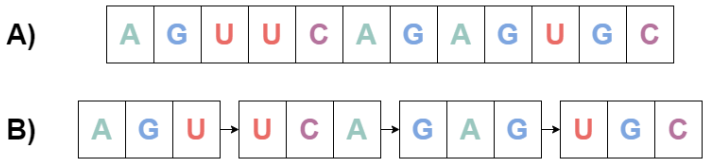


Figura 1.2: A) Exemplo da representação sequencial de proteína; B) A mesma proteína com seus códons em evidência.

1.3 Justificativa

O presente projeto tem por motivação o aprofundamento no conhecimento de características estruturais do RNA da dengue que possam estar associadas à severidade da infecção. A medida que os anos se passam, a total compreensão do conjunto de eventos que levam um humano a desenvolver dengue severa segue obscuro, restando apenas poucas teorias que levantam hipóteses sobre o desenvolvimento de infecção grave.

As teorias mais aceitas são: a teoria de Rosen [1977] relaciona formas mais graves de dengue à virulência da cepa infectante, de modo que cepas muito virulentas fornecem condições mais sérias; a teoria de Halstead [1970] relaciona as formas graves da infecção às reinfecções por diferentes sorotipos da dengue; a teoria mais recente, proposta por pesquisadores cubanos [Palacios Serrano et al., 2001], combina as teorias de virulência da cepa e das reinfecções como fator de risco para casos graves de dengue. O fato das razões que proporcionam infecções graves em alguns hospedeiros não ser completamente compreendida viabilizou a escolha do tema dessa pesquisa, pois, pesquisas que revelem novas informações sobre esse assunto possuem significância para área médica, cientifica e social, pois podem colaborar com o desenvolvimento de novas drogas, disponibilizar conhecimentos que podem ser aplicados na problemática de outros organismos virais semelhantes e contribuir na elaboração de medidas de proteção e controle do vírus.

Podemos observar com frequência em nossas revisões de literatura que diversos autores empregam métodos robustos de Processamento de Linguagem Natural (PLN) para etapa de representação das sequências biológicas, dessa forma aprendendo boas

representações para grandes conjuntos de dados e, constantemente, obtendo ótimos resultados de classificação [Asgari, 2015, Asgari et al., 2019, Le et al., 2019a, Ho et al., 2019]. Infelizmente, um dos problemas de nossa pesquisa é a quantidade de amostras, inviabilizando a utilização de modelos de PLN diante da alta dimensão vetorial que essas representações exigem para codificação de informações estruturais das sequências. Além disso, a complexidade algébrica por trás desses métodos tornam necessário a utilização de quantidades significativas de amostras para que associações entre as partes que integram as sequências sejam caracterizadas corretamente pelos vetores de representação.

Logo, como a base de dados deste trabalho não é suficientemente grande, fez-se necessário a exploração de outros caminhos para representação de sequências. Matrizes de co-ocorrências tem sido utilizadas para coleta de estatísticas de dados variados, especialmente dados de imagem e texto. Na análise de imagens médicas, as matrizes de co-ocorrência são empregadas para mensurar texturas de imagens. No campo de Processamento de Linguagem Natural (NLP), as co-ocorrências podem fornecer indícios de relações semânticas entre palavras em um corpo de texto. As matrizes não exigem cálculos complexos para sua geração e são capazes de codificar relações funcionais importantes para os processos biológicos de proteínas que podem ser encontrados na forma de padrões idênticos de co-ocorrências presentes em diferentes regiões da sequência Lee et al. [2013, 2014]. Portanto, surge a oportunidade do emprego de matrizes de co-ocorrências para tarefa de representação de sequências Carr and De Miranda [1998], Zhang et al. [2017], Brochier et al. [2019], Abdel-Nasser et al. [2019], Pennington et al. [2014].

Além disso, não foram encontradas na literatura abordagens que empreguem técnicas de aprendizagem de máquina para solução da tarefa de classificação desse problema, tornando esse um caminho ainda inexplorado. De maneira geral, independente do problema de classificação associada a sequências biológicas, não foram encontrados na literatura trabalhos que empreguem métodos de interpretação de modelos como *SHAP Values* e *LIME* para exploração de padrões genéticos. Dessa forma, a utilização da interpretação de modelos agrega valor de compreensão do classificador ao apresentar combinações que levem a tomada de decisão e, consequentemente, elevam o conhecimento do problema.

1.4 Metodologia

Conduzida pela pesquisa realizada por Asgari [2015], foi realizada uma revisão da literatura para avaliar os trabalhos de computação/bioinformática que abordam a tarefa de classificação de sequências biológicas em geral. Para obter um panorama de abordagens recentes, foram selecionados trabalhos entre os anos de 2011 e 2019. A revisão será apresentada no decorrer deste trabalho.

Na revisão, é possível identificar que, de maneira geral, para alcançar o objetivo de classificação os métodos realizam uma etapa de pré-processamento dos dados sequenciais, denominada etapa de representação onde as sequências são codificando em vetores/matrizes de números reais capazes de expressar informações biológicas presentes na sequência bruta. Alguns autores fazem associação de sequências biológicas com linguagens naturais, possibilitando o emprego de métodos desenvolvidos no campo da teoria formal da linguagem para a representação de sequências e, dessa forma, obter vetores de números reais (*embeddings*) capazes de caracterizar uma sequência inteira [Asgari, 2015, Asgari et al., 2019, Ho et al., 2019]. No entanto, outros autores seguem o caminho da codificação padrões de subsequências (códons, nucleotídeos e aminoácidos) em estruturas matriciais e, dessa forma, destacando-se padrões de difícil acesso [Zeng et al., 2016, Pan and Shen, 2018b, Sharma et al., 2019].

Neste trabalho comparamos a capacidade de representação dos métodos de *embeddings* proposto por Asgari [2015] e matrizes de co-ocorrência de códons para nossa base de dados. Serão apresentados resultados de classificação de dengue severa através de representações obtidas por métodos de PLN e representações através da codificação de padrões de códons em matrizes de co-ocorrências.

Também analisamos a capacidade de caracterização de dengue severa de cada proteína. Para que isso seja possível, cada sequência bruta de RNA deve ser alinhada por um método especifico, dessa forma, viabilizando a extração da subsequência de RNA que representam as 10 proteínas da dengue. Para isso, foram realizados testes de hipótese estatísticos para avaliar a capacidade de cada proteína em caracterizar dengue severa.

Em seguida, as matrizes de co-ocorrência da proteína com melhores resultados de classificação foram selecionadas para interpretação. O método de interpretação local *SHAP Values* [Lundberg and Lee, 2017] foi selecionado para essa tarefa, portanto, as interpretações realizadas pelo método são independentes para cada amostra. Por fim, as co-ocorrências que elevam a probabilidade de dengue severa foram ranqueadas de acordo com a significância atribuída pelo interpretador, gerando-se então gráficos de interpretação.

Logo, nosso método é dividido em 6 etapas, sendo elas: i) alinhar a sequência do RNA viral e segmentá-la por proteína para que estas sejam exploradas de forma independente; ii) normalizar e *tokenizar*[1] sequências para padronização e obtenção de códons das proteínas, respectivamente; iii) gerar matrizes de co-ocorrências de códons que servirão como dados de treino para o classificador; iv) classificar as matrizes de co-ocorrências de códons; v) realizar testes estatísticos para avaliar a proteína com melhor caracterização de dengue severa; vi) apresentar de forma gráfica as interpretações ranqueadas obtidas pelo *SHAP Values*, evidenciando características de co-ocorrências de códons que possam estar associadas ao desenvolvimento de dengue severa.

1.5 Objetivos

O objetivo principal deste trabalho é revelar e selecionar a proteína da dengue com maior capacidade de caracterização de dengue severa e, adicionalmente, gerar interpretações para o classificador da proteína selecionada, ranqueando os padrões que elevem a probabilidade de dengue severa nas amostras.

Devemos atingir os seguintes objetivos específicos para alcançar o objetivo geral deste trabalho:

1. Coletar e organizar uma base de dados com sequências de RNA do vírus da Dengue rotuladas de acordo à severidade da infecção do hospedeiro;

2. Elaborar um método para representação de proteínas que considere as interações entre pares de códons;

3. Selecionar, implementar e refinar o classificador com melhor performance sob as representações de proteínas;

4. Realizar testes estatísticos para definir a proteína com maior capacidade de caracterização de dengue severa;

5. Interpretar as decisões do classificador;

6. Apresentar as diferenças estruturais nos padrões significantes para dengue severa comparadas com dengue clássica.

tokenizar[1]: Processo de dividir uma sequência de caracteres em subsequências contíguas.

1.6 Contribuições

As contribuições alcançadas neste trabalho são:

1. Disponibilização de uma base de dados de sequências proteicas da dengue com rótulos de severidade, possibilitando sua utilização em outras pesquisas. As amostras de proteínas das bases não foram encontradas de forma integral em nenhuma base pública, fazendo necessário que pesquisas exaustivas fossem realizadas;

2. Um método aplicável a outros problemas de mesma finalidade. Por exemplo, a identificação de particularidades genômicas no RNA de Sars-Cov2 que levam ao desenvolvimento a quadros clínicos que incluem dificuldade respiratória, perda de fala e movimento e pressão no peito;

3. Uma metodologia de interpretação que permite extrair evidências de características genômicas estruturais distintas entre amostras de proteínas de dengue clássica e severa;

4. Evidências da proteína com maior capacidade de caracterização de dengue severa, ou seja, quando comparada com outras proteínas, essa deve permitir que a dengue severa seja identificada com maior facilidade.

1.7 Estrutura do trabalho

Os capítulos deste trabalho estão organizados da seguinte maneira: a fundamentação teórica dos métodos utilizados para as tarefas de representação, classificação e interpretação de proteínas da dengue são apresentadas no Capítulo 2; o Capítulo 3 discute pesquisas relacionadas ao processo de representação e classificação de sequências biológicas em geral; a solução proposta é apresentada no Capítulo 4; os resultados dos experimentos realizados podem ser encontrados no Capítulo 5 e; por fim, as considerações finais desta pesquisa são apresentadas no Capítulo 6.

Capítulo 2

Fundamentação teórica

Este capítulo expõe informações necessárias para a compreensão de seções das abordagens propostas. Neste capítulo são apresentados conceitos sobre organismos virais, sequências biológicas, proteínas e códons. São exibidas ferramentas para aquisição das amostras que constituem as bases de dados. Finalmente, são apresentados os conceitos sobre métodos de representação de sequências, Redes Neurais Artificiais e Convolutivas, *Random Forests* e *SHAP Values*.

2.1 O vírus da dengue e seu RNA

Os vírus pertencem a uma classe de seres simples, basicamente formados por uma cápsula proteica que envolve o material genético (RNA e/ou DNA). Eles podem infectar organismos vivos e desencadear reações indesejadas em seus hospedeiros. A dengue pertence a família de vírus *flaviviridae*, em que os vetores são essencialmente artrópodes. Ademais, os vírus da dengue são separáveis em quatro grupos sorologicamente distintos (DENV1, DENV2, DENV3 e DENV4), ou seja, em grupos que apresentam respostas diferentes para anticorpos semelhantes, fazendo com que as chances de reinfecções causadas pelo vírus seja maior [Lindenbach et al., 2013]. Portanto, diferenças nas estruturas genômicas dos vírus de cada sorotipo são suficientes para que a resposta imune do hospedeiro seja diferente em cada infecção [Westaway and Blok, 1997].

Todos organismos biológicos possuem materiais genéticos com informações que os caracterizam e coordenam seu desenvolvimento, funcionamento e transmissão de características hereditárias [Durbin et al., 1998, Alberts et al., 2002]. Esses materiais genéticos podem ser representados por uma sequência de caracteres retirados de um alfabeto específico conhecido como tabela iUPAC (international Union of Pure and Applied Chemistry) [IUPAC-IUB, 1970]. Dessa forma, $\forall g_i \in A : G = \{g_1 g_2 g_3 ... g_{n-1} g_n\}$,

onde G é um genoma, A é o alfabeto e g_i é um caractere que pode representar um nucleotídeo ou aminoácido. A representação dos materiais genéticos mediante símbolos discretos permite que esses também possam ser denominados, de forma geral, sequências biológica ou, de formas específicas, sequência proteica, sequência de RNA/DNA ou cadeia de nucleotídios.

O material genético da dengue consiste em uma sequência de RNA composta por aproximadamente 10,700 nucleotídeos (10,700 pares de bases – bp), onde, uma subsequência de aproximadamente 10,200 bp é traduzida em 3 proteínas estruturais: capsídeo (C), membrana (M) e envelope (E) e; em 7 proteínas não estruturais (NS): NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5. A subsequência traduzida é denominada sequência codificante (*Coding Sequence* – CDS) e contém as regiões codificantes de genes, ou seja, proteínas e moléculas de RNA. Duas subsequências encontradas nas extremidades do RNA não transcrevem proteínas e são conhecidas como regiões reguladoras 5'UTR e 3'UTR, somando aproximadamente 500 bp nucleotídeos. Portanto, as subestruturas elementares da dengue estão dispostas na seguinte sequência {5'UTR, C, M, E, NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5, 3'UTR}, conforme apresentado na figura 1.1 (Seção 1).

2.2 Códons

Sequências biológicas são compostas por enormes cadeias de caracteres, o que dificulta a observação e extração de padrões internos. Para contornar esse problema, uma técnica bastante empregada é a divisão da sequência em múltiplas subsequências, tornando possível a coleta de padrões e associações entre subsequências capazes de representar características da sequência completa, tais como, expressões gênicas, regiões funcionais e identificadores de espécies em amostras meta-genômicas [Welch et al., 2009, Gustafsson et al., 2004, Perry and Beiko, 2010]. Na literatura, os tamanhos das subsequências frequentemente utilizados são $1, 2, 3$, e 4, pois representam, respectivamente, as seguintes estruturas biológicas: nucleotídeos, dinucleotídeos, aminoácidos (códons) e tetranucleotídeos [Perry and Beiko, 2010, Yakovchuk et al., 2006, Karlin, 1998, Hershberg and Petrov, 2008].

Um códon é composto por uma subsequência de três nucleotídeos que codificam determinado aminoácido ou região que indica começo ou fim da tradução do RNA/DNA. Os códons da sequência de RNA da dengue são combinações com repetição dos nucleótidos adenina (A), citosina (C), guanina (G) e uracila (U). Apesar das combinações entre nucleotídeos gerar um total de 64 códons, apenas 20 aminoáci-

dos são codificados. Essa assimetria é chamada de degenerescência do código genético e caracteriza a capacidade de um aminoácido ser codificado por mais de um códon [Lehmann and Libchaber, 2008]. A Tabela 2.1 representa as combinações que geram os 20 aminoácidos existentes.

		2 base			
		U	**C**	**A**	**G**
1 base	**U**	UUU - Fenilalanina (F)	UCU - Serina (S)	UAU - Tirosina (Y)	UGU - Cisteína (C)
		UUC - Fenilalanina (F)	UCC - Serina (S)	UAC - Tirosina (Y)	UGC - Cisteína (C)
		UUA - Leucina (L)	UCA - Serina (S)	UAA - Ocre	UGA - Opala
		UUG - Leucina (L)	UCG - Serina (S)	UAG - Âmbar	UGG - Triptofano (W)
	C	CUU - Leucina (L)	CCU - Prolina (P)	CAU - Histidina (H)	CGU - Arginina (R)
		CUC - Leucina (L)	CCC - Prolina (P)	CAC - Histidina (H)	CGC - Arginina (R)
		CUA - Leucina (L)	CCA - Prolina (P)	CAA - Glutamina (Q)	CGA - Arginina (R)
		CUG - Leucina (L)	CCG - Prolina (P)	CAG - Glutamina (Q)	CGG - Arginina (R)
	A	AUU - Isoleucina (I)	ACU - Treonina (T)	AAU - Asparagina (N)	AGU - Serina (S)
		AUC - Isoleucina (I)	ACC - Treonina (T)	AAC - Asparagina (N)	AGC - Serina (S)
		AUA - Isoleucina (I)	ACA - Treonina (T)	AAA - Lisina (K)	AGA - Arginina (R)
		AUG - Metionina (M)	ACG - Treonina (T)	AAG - Lisina (K)	AGG - Arginina (R)
	G	GUU - Valina (V)	GCU - Alanina (A)	GAU - Ácido aspártico (D)	GGU - Glicina (G)
		GUC - Valina (V)	GCC - Alanina (A)	GAC - Ácido aspártico (D)	GGC - Glicina (G)
		GUA - Valina (V)	GCA - Alanina (A)	GAA - Ácido glutâmico (E)	GGA - Glicina (G)
		GUG - Valina (V)	GCG - Alanina (A)	GAG - Ácido glutâmico (E)	GGG - Glicina (G)

Tabela 2.1: Todos 64 possíveis códons no RNA da dengue e suas respectivas transcrições em aminoácidos. Âmbar, Ocre e Opala são códons que marcam o fim da tradução de RNA, enquanto que os códons UUG (L), CUG (L), AUU (I), AUG (M) e GUG (V) marcam o início da tradução.

Agora, com as sequências biológicas representadas por subsequências, podemos concluir que o dicionário de subestruturas das sequências será composto pelas subsequências exclusivas. O algoritmo de tokenização pode ser utilizado para tarefa de obtenção de subsequências e dicionário. Esses algoritmo será explicado a seguir.

A tokenização é uma abordagem eficiente capaz de revelar regiões conservadas na sequência que desempenham papel funcional fundamental [Daraselia et al., 2004, Hatzivassiloglou et al., 2001]. A Figura 2.1 apresenta o comportamento do algoritmo de tokenização (1) sobre uma sequência de nucleotídeos. As sequências originais X e o tamanho k da subsequência são passados para o algoritmo. Em seguida, o algoritmo particiona cada sequência original $x_i \in X$ em um conjunto M_i de múltiplas subsequências. Por fim, retorna o dicionário D e o conjunto S de subsequências para cada sequência original. A complexidade deste algoritmo é $O(|P| \times |X|)$.

2.3 Representação de sequências biológicas

Tarefas recorrentes em análise sequencial envolvem classificação de sequências em algum nível, como por exemplo, a descoberta de processos evolutivos de determinado

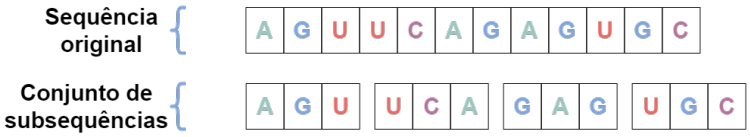


Figura 2.1: Divisão de uma sequência em múltiplas subsequências através do algoritmo de tokenização.

Algoritmo 1 *Tokenizer*

Entrada 1: $X \leftarrow$ Sequências biológicas
Entrada 2: $k \leftarrow$ Tamanho da subsequência
Saída 1: Dicionário de subsequências
Saída 2: Conjunto de subsequências para cada sequência
$l \leftarrow 1$
$D \leftarrow \{\}$
$S \leftarrow \{\}$
para cada sequência x em X **faça**
    $M \leftarrow \{\}$
    para $i \leftarrow 1$ até X.tamanho **faça**
        se $x[l : (l + k - 1)] \notin D$ **então**
            $D \leftarrow x[l : (l + k - 1)]$
        fim se
        $M.insere(x[l : (l + k - 1)])$
        $l \leftarrow l + k$
    fim para
    $S.insere(M)$
fim para
Retorna D (Dicionário de subsequências) e S (Conjunto de subsequências para cada sequência)

organismo através da classificação de suas sequências biológicas em famílias de organismos conhecidos [Saidi et al., 2012]. Por muito tempo a classificação de sequências se deu por alinhamento sequencial e busca por similaridades entre novas sequências e sequências conhecidas e alinhadas. Essa abordagem possui o grave problema de sequências homólogas, onde, caso a sequência de interesse não apresente sequências homólogas, torna-se inutilizável [Debroas et al., 2009].

Esse e outros problemas tornaram evidente a necessidade da compreensão aprofundada sobre estruturas, processos e funções que podem ocorrer nas sequências biológicas, fazendo com que diversos autores utilizassem abordagens mais eficientes para representação estrutural dessas sequências (Seção 3.1). Nesse ponto, métodos e técnicas de Processamento de Linguagem Natural (PLN) surgem como alternativa diante da sua capacidade de quantificar relações entre itens pertencentes a uma estrutura, por

exemplo, quantificar relações entre palavras de uma frase.

As atividades exercidas no campo de PLN consistem, resumidamente, em empregar um conjunto de abordagens matemáticas, estatísticas e computacionais na análise de texto. A definição mais aceita de PLN diz que a área consiste em uma gama de técnicas motivadas teoricamente para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística, com o objetivo de obter processamento de linguagem semelhante ao processamento humano para uma variedade de tarefas ou aplicativos [Liddy, 2001].

Duas características de sequências biológicas que justificam o emprego de métodos de PLN para o pré-processamento de sequências são: a estrutura da sequência, com nucleotídeos/aminoácidos retirados de um alfabeto específico (Seção 2.1) e; sequências biológicas seguem determinada ordem lógica na distribuição de suas subestruturas, assim como frases e textos [Asgari, 2015, Dongardive and Abraham, 2016, Islam et al., 2018].

Em PLN uma matriz de coocorrência é uma formulação tabular de ocorrências conjuntas de palavras em uma janela de contexto, permitindo que valores de probabilidades condicionais para ocorrência de palavras sejam gerados. A matriz de co-ocorrências permite encontrar relações entre palavras distantes em uma frase, mas que pertencem a uma mesma janela de contexto. Essas características as tornam ferramentas úteis para encontrar relações e padrões de palavras capazes de caracterizar um evento.

Sejam as frases "eu gosto de café" e "eu gosto de jogar vídeo game", temos o conjunto de palavras P = {eu, gosto, de, café, jogar, vídeo, game }, portanto, a matriz de co-ocorrências de palavras terá linhas e colunas que representam as palavras do conjunto P. Os possíveis conjuntos de palavras para uma janela de contexto de tamanho três são:

1. **Frase 1** {(eu, gosto, de), (gosto, de, café)}
2. **Frase 2** {(eu, gosto, de), (gosto, de, jogar), (de, jogar, vídeo), (jogar, vídeo, game) }

Observe que as palavras "gosto" e "de" coocorrem quatro vezes. A tabela 2.2 apresenta o restante das coocorrências de palavras das frases. Matrizes de co-ocorrências são boas ferramentas para coletar estatísticas globais e, consequentemente, revelar padrões globais que ocorrem em um corpo de texto [Pennington et al., 2014].

As matrizes de co-ocorrência são simétricas e intercambiáveis, isso significa que as co-ocorrências (eu, gosto) e (gosto, eu) tem o mesmo valor e o mesmo papel sintático

	eu	gosto	de	café	jogar	vídeo	game
eu	0	2	2	0	0	0	0
gosto	2	0	4	1	1	0	0
de	2	4	0	1	2	1	0
café	0	1	1	0	0	0	0
jogar	0	1	2	0	0	1	1
vídeo	0	0	1	0	1	0	1
game	0	0	0	0	1	1	0

Tabela 2.2: Matriz de co-ocorrências de palavras das frases "Eu gosto de café."e "Eu gosto de jogar vídeo game.".

na frase. Isso torna possível o redimensionamento das matrizes de co-ocorrência em estruturas matemáticas de dimensão menor por podermos representar toda matriz apenas pela sua matriz triangular superior/inferior.

A aplicação de matrizes de co-ocorrências também se expande para o campo da bioinformática, por exemplo, em sequencias proteicas evidências de relações funcionais importantes para os processos biológicos de proteínas podem ser encontradas quando padrões idênticos de co-ocorrências de aminoácidos estão presentes em diferentes regiões [Lee et al., 2013, 2014]. Para nosso problema, as matrizes de co-ocorrência podem ser empregadas para estruturar de forma tabular padrões de co-ocorrência de códons nas sequências do vírus da dengue.

2.4 Algoritmo Quick Hull

O método *Quick Hull*, proposto por Barber et al. [1996], pertence a uma família de métodos computacionais capazes de calcular o casco convexo de pontos em um espaço vetorial. Em outras palavras, *Quick Hull* é capaz de gerar um envelope para um conjunto finito de pontos, como é possível observar na Figura 2.2. Assume-se inicialmente que um conjunto de $d + 1$ pontos estão posicionados de tal maneira que seu envelope seja um complexo simples de vértices e arestas. Cada aresta inclui um conjunto de vértices, um conjunto de arestas vizinhas e um hiperplano. As duas operações geométricas utilizadas pelo método são: hiperplano orientado por pontos e distância entre ponto e hiperplano. Barber et al. [1996] utilizam as operações de distância para determinar se um ponto se encontra dentro ou fora do envelope.

O Teorema 1 para a criação do envelope de pontos diz que: Seja H um envelope em $\mathbb{R}^d$ e p um ponto em $\mathbb{R}^d - H$. Então F é uma aresta de $envelope(p \cup H)$, se e somente se,

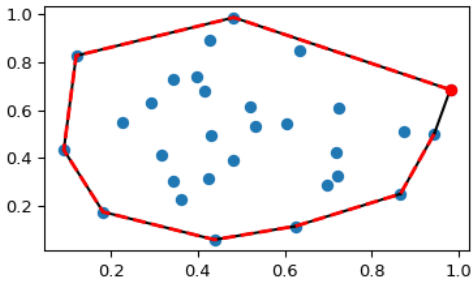


Figura 2.2: Envelope do subconjunto denso de pontos em um espaço bidimensional. Fonte: Scipy

Teorema 1 *F é uma aresta de H e p está abaixo de F; ou F não é uma aresta de H, seus vértices são p e os vértices de um envelope H com uma aresta incidente abaixo de p e a outra aresta acima de p.*

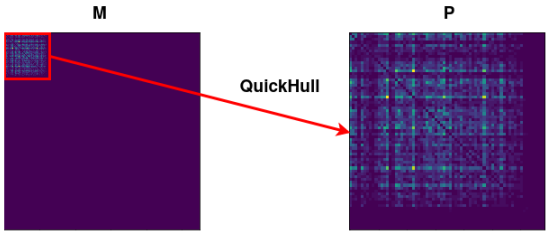


Figura 2.3: Exemplo da utilização do algoritmo *Quick Hull* em uma matriz esparsa para redução de dimensão e esparsidade.

Neste trabalho utiliza-se diversas matrizes de co-ocorrências de códons como entrada de dados para classificadores. No entanto, estas matrizes apresentam áreas esparsas que, consequentemente, atrapalha o desempenho dos classificadores. Por isso, o algoritmo *Quick Hull* será utilizado neste trabalho para gerar novas matrizes de co-ocorrências menos esparsas. Para isso, uma matriz de coocorrência M será tratada como um plano, tal que cada valor de M diferente de 0 é considerando um ponto existente no plano. Um envelope de pontos será gerado para M, posteriormente uma nova matriz P menor e menos esparsa é obtida com os pontos internos do envelope de M, como podemos visualizar na Figura 2.3.

2.5 Redes Neurais Artificiais

Redes Neurais Artificiais (*Artificial Neural Networks* – ANN) são modelos de aprendizagem de máquina não lineares capazes de aprender a reconhecer padrões em dados. As ANN são formadas por camadas de funções de ativação f(ou neurônios) não-lineares e pesos w, de modo que as funções f podem ser interpretadas como regressões que geram a entrada de dados para os próximos neurônios da rede, conforme apresentado na Figura 2.4. Essas arquiteturas comumente são caracterizadas por grafos, tal que, os nós e arestas representam neurônios e pesos, respectivamente. Essa forma de visualização das ANN torna a compreensão de sua estrutura e fluxo de dados mais amigável.

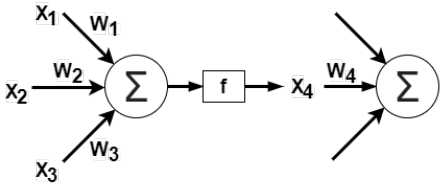


Figura 2.4: Representação de um neurônio artificial. A função $\Sigma(\cdot)$ quantifica a soma do produto de cada peso com sua respectiva entrada acrescido de um viés β, logo, $\Sigma = x_1w_1 + x_2w_2 + x_3w_3 + \beta$. A função de ativação f calcula a saída de x_4 através de $f(\Sigma(\cdot))$. Esse processo se repete para todos neurônios.

O objetivo das ANN é obter, por meio de diversas transformações não lineares das entradas, um conjunto de pesos capaz de modelar corretamente dados com separação não trivial. Para tal, os pesos da rede devem ser atualizados de modo a minimizar a função de custo do modelo. Esse processo é conhecido como treinamento e é realizado pelo algoritmo de *backpropagation* (propagação retrograda).

O algoritmo *backpropagation* baseia-se no conceito de gradientes para atualizar os pesos da rede. Esses gradientes são obtidos através das derivadas parciais da função de custo em relação aos pesos do modelo, em seguida os gradientes são utilizados para ajustar os pesos de modo que esse ajuste reflita na função de custo alcançar seu mínimo global, e dessa forma, fazendo com que o erro de predição seja o menor possível [Goodfellow et al., 2016].

Em uma explicação mais técnica, os gradientes são obtidos derivando a função de custo em relação à função de ativação, por conta disso, essas últimas necessitam ter derivada, ou seja, devem ser funções não lineares. Os algoritmos que calculam o gradiente são conhecidos como otimizadores e movem os gradientes iterativamente em sentido a direção mais íngreme do gráfico tridimensional da função de custo [Goodfellow

et al., 2016, Deng et al., 2014]. É possível observar na Figura 2.5 a descida de gradientes em uma superfície, em redes neurais, essa superfície é, na verdade, um hiperplano.

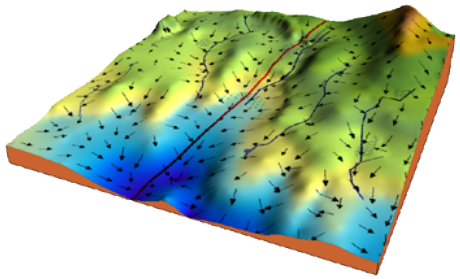


Figura 2.5: Movimentação de gradientes a procura do ponto de mínimo global de uma função de custo

A estrutura de uma ANN básica é composta por uma camada específica para entrada de dados na rede e uma camada de saída para gerar resultados, sejam eles de classificação ou regressão. Outras estruturas de ANN mais complexas, denominadas ANN profundas, possuem uma ou mais camadas em níveis internos do grafo da rede, também chamadas de camadas ocultas. A quantidade de camadas ocultas torna as redes profundas mais complexas que redes rasas permitindo que estas aprendem abstrações de alto nível dos dados de entrada. As primeiras camadas permitem que o modelo aprenda características simples, e a medida que a rede fica profunda são aprendidas caraterísticas mais complexas, desse modo, o modelo aprende informações complexas a partir de informações mais simples. Portanto, torna-se desnecessária a especificação explicita de informação para que o modelo atinja seu objetivo, pois ele aprende através de experiência [Goodfellow et al., 2016, Rosenblatt, 1961].

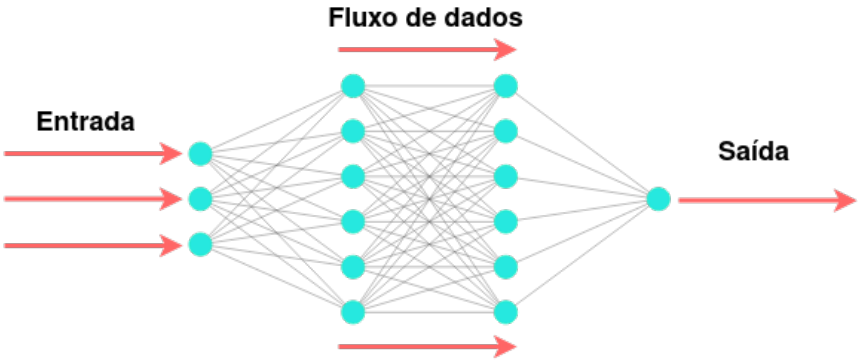


Figura 2.6: Exemplo de *feed-foward* com quatro camadas.

Um tipo de ANN bastante comuns são as redes *feed-fowards*, representadas por um grafo acíclico. Neste tipo de rede as informações fluem em uma única direção, para frente, passando pelos neurônios de entrada e, por fim, pelos neurônios de saída [Schmidhuber, 2015]. A Figura 2.6 contém um exemplo de grafo de rede *feed-foward*, onde, os nós em cor azul representam os neurônios e as arestas em cor cinza, os pesos. Além disso, o sentido do fluxo de dados é representado pelas setas vermelhas.

As Redes Neurais Convolutivas (*Convolutional Neural Network* – CNN) são um tipo específico de *feed-fowards* para classificação de dados com estrutura tabular e, diferente das *feed-fowards* clássicas que empregam a operação de soma de produto entre pesos e entradas nos seus neurônios, estas utilizam a operação matemática de convolução entre as matrizes de entrada e a matriz de *kernel*. Uma CNN básica pode ser representada como uma pilha de camadas de operações convolutivas e de agrupamento (*pooling*) [LeCun et al., 2015].

A convolução pode ser resumida como uma função linear que quantifica a soma do produto entre duas funções. Sejam f e g duas funções, a função h denota a convolução $f \otimes g$, tak que, a área da interseção entre $f(x)$ e $g(\alpha - x)$ expressa a convolução no instante α [Gonzales and Woods, 2002], conforme a equação abaixo,

$$h(x) = f(x) \otimes g(x) = \int_{-\infty}^{\infty} f(\alpha)g(x - \alpha)\partial\alpha. \tag{2.1}$$

As CNN utilizam o conceito de campos receptivos locais, inspirando-se no funcionamento do córtex visual humano e na quantidade de células responsáveis pela detecção de luz que este possui. Os campos receptivos locais (*kernels*) de camadas convolutivas armazenam informações sobre regiões menores da matriz de entrada e as repassam para uma matriz denominada mapa de características. Esse processo é realizado até a última camada de convolução.

Os *kernels* são aplicados na imagem de entrada com sobreposição, dessa forma, permitindo a cobertura de toda imagem por um único *kernel*. Isso permite que a convolução entre imagem e *kernel* resultem em mapas que conservam correlações espaciais entre os *pixels* da imagem original. Diferentes mapas de recursos podem ser aprendidos por diferentes *kernels*, portanto, em uma única camada de convolução podem ser obtidos diferentes mapas. A quantidade de *kernels* por camada também é conhecida como quantidade de filtros.

Os mapas de recursos conservam abstrações da matriz de dados passada para operação de convolução. Em camadas iniciais os mapas aprendem características básicas e bastante descritivas das imagens de entrada, como por exemplo, cantos, bordas e profundidade. Ressalta-se que quanto mais filtros, mais abstrações aprendidas e, di-

ante da quantidade de ruídos e informações irrelevantes nos dados de entrada brutos, é esperado que nas camadas iniciais existam poucos filtros como forma de prevenir o aprendizado de abstrações de ruídos e informações tendenciosa e evitar que essas sejam passadas para o restante da rede.

Como forma de reduzir ainda mais a possibilidade de inserção de ruído na rede, as CNN empregam uma operação adicional após os mapas de recursos, essa operação é denominada *pooling*. As camadas de *pooling* geram matrizes significativamente menores que os mapas de recursos, essas matrizes são compostas por agrupamentos espaciais de regiões dos mapas. Por substituir os mapas por estatísticas resumidas de saídas próximas, preservando informações relacionadas à tarefa e removendo detalhes irrelevantes, as operações de *pooling* tornam o problema menos complexo, fazendo com que as CNN sejam mais profundas e utilizem menos parâmetros que outras *feed-fowards* que poderiam resolver problemas com entrada tabular [Goodfellow et al., 2016, Boureau et al., 2010].

As operações de *pooling* frequentemente empregadas em CNN atuais são *pooling* de média (*average pooling*) e *pooling* de máximo (*max pooling*). Essas operações são empregadas em vizinhanças retangulares dos mapas, tal que, o *pooling* de média calcula a média da vizinhança e repassa esse valor à matriz de *pooling*, enquanto que o *pooling* de máximo repassa o valor máximo da vizinhança. A Figura 2.7 ilustra o funcionamento dos *pooling* de média e do *pooling* de máximo sobre mapas de recursos.

Após obter o último conjunto de mapas de recursos, a CNN vetoriza esse conjunto de dados e passa-o para uma camada densa completamente conectada, que por usa vez realiza a classificação.

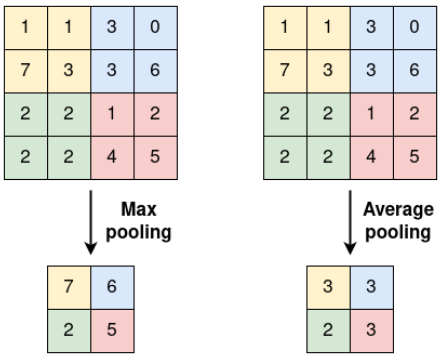


Figura 2.7: Operações de *pooling* para vizinhança retangular de dimensão (2×2). Cada vizinhança é representada por uma cor individual. As matrizes $(4 \times 4$ representam os mapas, enquanto que as matrizes (2×2) representam a matriz de *pooling*.

Dada a capacidade de aprender mapeamentos complexos para dados não-

linearmente separáveis após aplicar a *backpropagation* na rede, a CNN pode executar suas próprias extrações de características nas primeiras camadas [LeCun et al., 2015], tornando desnecessária a utilização de modelos auxiliares para esta tarefa. Nesse contexto, os *kernels* são as matrizes de pesos a serem ajustadas via gradiente descendente, enquanto que os mapas de recursos são as saídas da camada de convolução que deverão passar por funções de ativação.

A Figura 2.8 ilustra uma CNN simples recebendo como dados de entrada uma imagem (matriz de *pixels*), sua arquitetura consiste em uma pilha de camadas de convolução e *pooling*, como podemos observar.

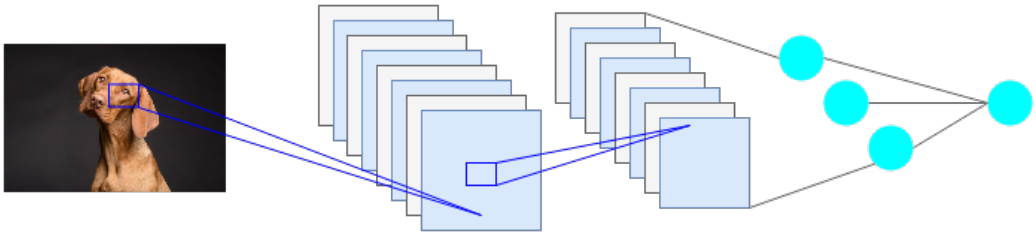


Figura 2.8: Arquitetura básica de CNN. Na imagem o *kernel* associado ao primeiro mapa está sobre a região do olho esquerdo do cão. A primeira e única camada de convolução dessa rede possui 8 filtros, ou seja, utiliza 8 *kernels* diferentes. Em seguida, cada mapa é reduzido a uma matriz de *pooling* e repassados a camadas densas para que, por fim, sejam classificados.

2.6 Random Forest

As *Random Forest* (RF) (Figura 2.9) são modelos *ensembles* baseados em árvores de decisão. Um modelo *ensemble* consiste em vários modelos independentes utilizados sobre o mesmo conjunto de dados que realizam uma tarefa em comum que são mais complexos, mais poderosos e mais capazes de realizar divisões não triviais de dados em espaços não lineares quando comparados com um único modelo mais simples [Dietterich, 2000, Rokach et al., 2014]. Outra característica fundamental da RF é sua baixa complexidade quando comparada com classificadores baseados em ANN, característica que a torna uma ferramenta poderosa na modelagem de pequenos conjuntos de dados. Além disso, por serem menos complexas, são mais fáceis de interpretar.

A estrutura básica das RF são árvores de decisão que, por sua vez, possuem como unidade básica árvores de decisão binárias (nós) que empregam particionamento recursivo nos dados. As árvores estabelecem seus nós calculando a informação pra cada variável do conjunto de treino, onde aquela com maior capacidade de divisão dos dados

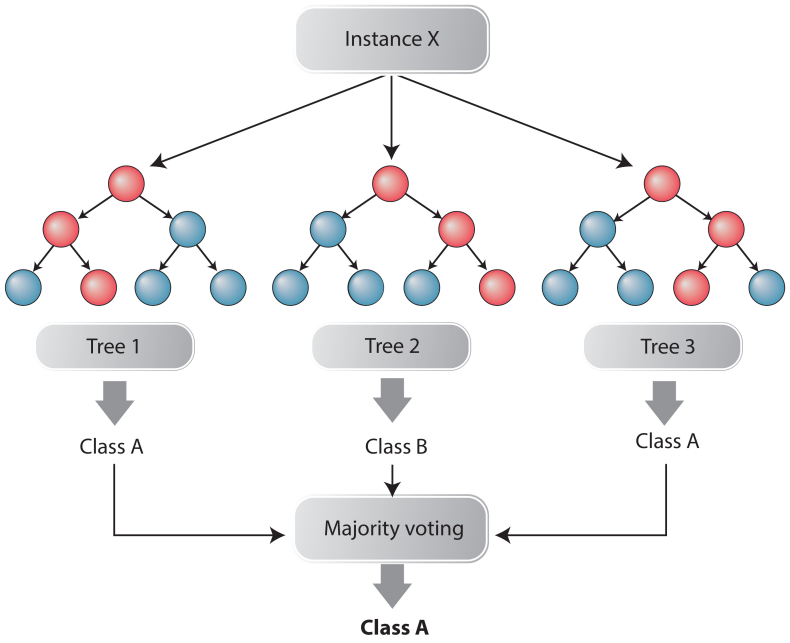


Figura 2.9: Classificador RF com três árvores de decisão. Neste exemplo a classe mais frequentes nas árvores foi a classe A, sendo essa a classe escolhida pela RF.

(maior informação) terá condições testadas no nó inicial da árvore (nó raiz). No fim da árvore encontram-se os nós folhas responsáveis pela classificação. Para definir a variável que será testada em cada nó o modelo quantifica a informação relativa entre variáveis e rótulos. Geralmente, esse valor é obtido pelo cálculo de entropia ou informação de Gini.

Após estabelecer a estrutura da árvore utilizando os dados de treino, as amostras tem suas variáveis testadas em cada nó. O resultado de um testes define o fluxo da amostra na árvore, até que essa chegue no último nó e seja classificada [Loh, 2011, Chen and Ishwaran, 2012].

O funcionamento básico de um RF é, dado um conjunto X de dados de treino, para cada árvore de decisão selecionar Θ variáveis aleatórias de X, portanto, podem ser matematicamente definidas como $\Sigma = \{h(X, \Theta_k), k = 1, ..., j\}$, onde h é a árvore de decisão, Θ_k representa as amostras aleatoriamente selecionadas para h e j é a quantidade de árvores na floresta aleatória. As árvores que constituem a RF devem ser independentes e identicamente distribuídas. A seleção de Θ variáveis aleatórias para cada árvore garante que essas sejam independentes, pois essas aprenderão regras diferentes de acordo com as variáveis atribuídas a cada uma delas. Essa característica também reduz relativamente o erro de *overfitting*, pois força o modelo a aprender padrões em

diferentes conjuntos de variáveis.

Além de selecionar Θ variáveis para cada árvore, como todo modelo *ensemble*, as RF fazem uso do método *Bootstrap Aggregation - Bagging* que consiste em: i) no treino, amostras de *bootstrap* são geradas para cada árvore, isso garante que o conjunto de treino em de cada árvore não será completamente igual e; ii) tanto na classificação quanto no treino a classe mais popular predita em cada árvore será a escolhida para representar a amostra, esse processo é chamado de votação ou *aggregation* [Breiman, 2001].

2.7 Interpretação de modelos

Muitos modelos de aprendizado de máquina profundos são caixas pretas funcionalmente, visto que a complexidade desses modelos torna praticamente impossível a compreensão do funcionamento interno. No entanto, em determinadas aplicações a interpretação de modelos é fundamental para que exista um domínio humano sobre o problema e evento de interesse. Tendo isso em vista, diversos métodos de interpretação de modelos foram propostos para explicar decisões tomadas por modelos por meio da avaliação da influência das variáveis de entrada nos resultados de predição [Molnar, 2019]. Tais métodos agrupam-se em duas classes: métodos globais e métodos locais. Os métodos globais interpretam os resultados do modelo para todas entradas de dados, enquanto que métodos locais interpretam uma entrada individual. Nesta seção, apresentaremos o método SHAP, utilizado na interpretação local de modelos.

Lundberg and Lee [2017] apresentam o método *Shapley Additive Explanations* (SHAP) para interpretação de modelos através da atribuição de importância para variáveis em uma predição particular. O método faz interpretações locais na saída do algoritmos unificando três métodos de Teoria dos Jogos conhecidos como *Shapley Values* [Lipovetsky and Conklin, 2001, Štrumbelj and Kononenko, 2014, Datta et al., 2016]. Tendo em vista que o melhor interpretador de modelos simples são eles próprios, pois são fáceis de entender, SHAP utiliza modelos simples para tarefa de interpretação de modelos complexos, como redes neurais profundas. Um modelo simples, também denominado *modelo de interpretação* é qualquer aproximação interpretável do modelo original. Seja g um modelo de interpretação, f o modelo preditivo original e x o conjunto de variáveis a serem explicadas pelo interpretador, o SHAP tenta garantir que o modelo g possua resultados aproximados aos do modelo f respeitando a condição $g(x') - f(h_x(x')) \approx 0$, tal que, x' são simplificações de x que podem ser mapeadas para entrada original a partir de uma função h_x, tal que, $x = h_x(x')$.

Para gerar interpretações, SHAP utiliza métodos de atribuição de variáveis aditivas, que consiste na remoção e adição de variáveis simplificadas ao decorrer das interpretações feitas por g. Primeiramente, g faz interpretações sem utilizar variáveis x', em seguida, cada variável x'_i é adicionada ao interpretador g, como apresentado na equação:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \ . \tag{2.2}$$

Por definição, os métodos de atribuição de variáveis aditivas possuem modelos de interpretação lineares que utilizam variáveis binárias, portanto, $x' \in \{0,1\}^M$, onde M é o número de entradas simplificadas, $\phi_i \in \mathbb{R}$ e g é um modelo linear. As soluções obtidas através de métodos de atribuição de variáveis aditivas possuem três propriedades fundamentais para interpretação de modelos:

1. **Precisão Local**: a precisão local exige que, ao obter o resultado do modelo original f para uma entrada específica x, o modelo de interpretação g corresponda à saída simplificada x'. Em outras palavras, $g(x')$ se aproxima de $f(x)$ quando $x = h_x(x')$, onde $\phi_0 = f(h_x(0))$ representa a saída do modelo com todas entradas simplificadas desativadas. A equação que representa a precisão pode ser visualizada a seguir,
$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i. \tag{2.3}$$

2. **Falta**: a propriedade de falta diz que, se as entradas simplificadas representam a presença de variáveis, a falta requer que as variáveis ausentes na entrada original não tenham impacto, conforme a lógica proposicional a seguir,
$$x'_i = 0 \implies \phi_i = 0, \forall i > 0, \tag{2.4}$$

3. **Consistência**: se um modelo for alterado para que a contribuição de alguma entrada simplificada permaneça a mesma, independentemente das outras entradas, a atribuição dessa entrada não deverá diminuir. A forma matemática da consistência pode ser vista na Equação 2.5. Sejam $f_x(z') = f_x(h_x(z'))$ e $z' \setminus i = \{x \in z'; x \notin i\}$ (o conjunto $z'_i = 0$ que representa falta de recursos). Para dois modelos f' e f, se,
$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i), \tag{2.5}$$

e todas entradas $z' \in \{0,1\}^M$, então, $\phi_i(f', x) \geq \phi_i(f, x)$.

Dada essas propriedades e a definição de modelos de interpretação (Equação 2.2), é possível obter o seguinte interpretador:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!}[f'_x(z') - f'_x(z' \setminus i)], \tag{2.6}$$

onde $|z'|$ é a quantidade de entradas não nulas em z' e $z' \subseteq x'$ representa todos os subconjuntos no conjunto de vetores não nulos x'.

O SHAP unifica a importância de variáveis através de uma função condicional de valor esperado do modelo original. Essa função é capaz de resolver a Equação 2.6, tal que, $f_x(z') = f(h_x(z')) = E(f(z)|z_S)$, onde S é o subconjunto não nulo de z'. Por fim, a equação geral do modelo de explicação do método SHAP assume a forma:

$$f(h_x(z')) = E(f(z)|z_S). \tag{2.7}$$

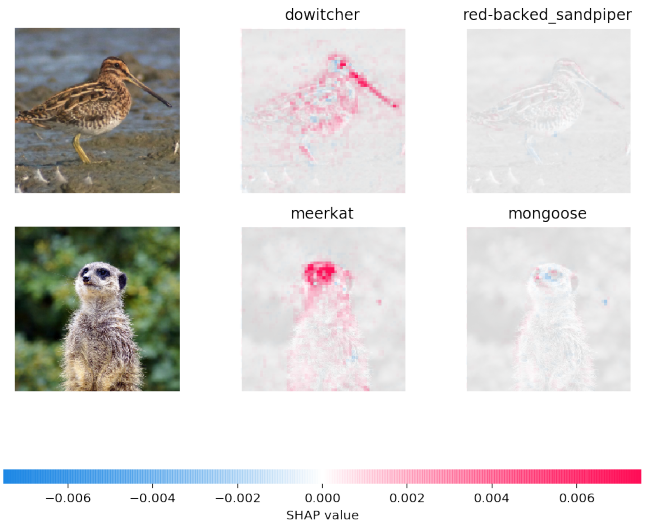


Figura 2.10: Interpretação de duas imagens utilizando SHAP Values para um modelo treinado a partir de imagens animais de diversas classes. Fonte: Repositório do SHAP no Github.

A Figura 2.10 apresenta a interpretação do SHAP para uma CNN-2D utilizada na classificação de imagens. Os pontos vermelhos são valores SHAP positivos e representam os *pixels* que aumentam a probabilidade da classe, enquanto que os pontos azuis são valores SHAP negativos que diminuem a probabilidade da classe. Na Figura 2.10, pode-se observar a imagem de uma ave *dowitcher* (esquerda), a interpretação do modelo para classe positiva (centro), e a interpretação do modelo para uma classe negativa

(direita). As áreas próximas ao bico aumentam a probabilidade para a classificação correta da ave, além disso, todos os *pixels* da imagem da classe verdadeira parecem não exercer nenhuma influência na classificação da ave *red backed sandpiper* (classe negativa), ou seja, poucos *pixels* da imagem desta ave *dowitcher* conseguem descrever uma ave *red backed sandpiper*.

2.8 Métricas de avaliação

As métricas avaliam o ajuste de modelos após o seu treinamento. Nesta seção serão apresentadas métricas aplicáveis em problemas de classificação e utilizadas na análise de resultados desta proposta. Uma das métricas mais simples na avaliação de classificadores é a matriz de confusão que mostra a frequência de acertos de classes. Na Figura 2.11 é possível visualizar a estrutura de uma matriz de confusão de duas classes.

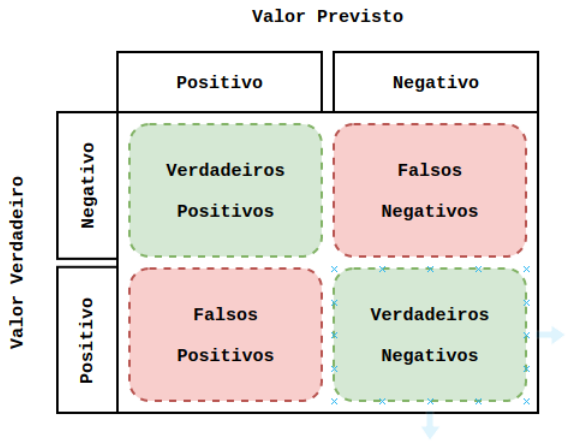


Figura 2.11: Matriz de confusão

Uma matriz de confusão binária, como a apresentada na Figura 2.11, é formada pelos seguintes itens:

- Verdadeiros Positivos (VP): é a quantidade de amostras pertencentes à primeira classe corretamente classificadas.
- Verdadeiros Negativos (VN): é a quantidade de amostras pertencentes à segunda classe corretamente classificadas.

- Falsos Positivos (*FP*): é a quantidade de amostras classificadas como pertencentes à primeira classe, quando na verdade pertencem à segunda.

- Falsos Negativos (*FN*): é a quantidade de amostras classificadas como pertencentes à segunda classe, quando na verdade pertencem à primeira.

As matrizes de confusão também são ótimas ferramentas para avaliação de classificadores em bases de dados desbalanceadas, pois estas possibilitam a visualização de acerto do modelo entre classes. Espera-se que em uma matriz de confusão os valores de sua diagonal principal sejam maximizados e que todos os outros sejam minimizados. Isto significa que a taxa de acerto do classificador entre as classes é confiável, pois mostra que este não está simplesmente atribuindo instâncias a classes mais frequentes, ou seja, não está "chutando" a classificação da amostra na classe mais frequente para aumentar a sua taxa de acerto.

Acurácia (acc) é uma métrica que representa a fração entre soma de instâncias classificadas corretamente e o total de instâncias. A acurácia possui a propriedade de ser sensível à distribuição de classes, isso ocorre pois a acurácia não atribui importância a uma classe específica, tornando-a facilmente influenciável pelo desbalanceamento de classes [OECD Statical Terms, 2020]. A acurácia pode ser calculada como:

$$\text{acc} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.8}$$

As métricas de *precisão* (prc) e *revocação* (rec) podem contornar este problema pois avaliam o comportamento do classificador sobre uma classe específica. A precisão é obtida pela razão do total de amostras classificadas corretamente pelo total de amostras de uma classe, conforme a Equação a seguir,

$$\text{prc} = \frac{VP}{VP + FP}, \tag{2.9}$$

já a revocação quantifica a razão do total de amostras classificadas corretamente pelo total de amostras que o classificador diz pertencer à classe, conforme a Equação s seguir,

$$\text{rec} = \frac{VP}{VP + FN}. \tag{2.10}$$

Outra métrica sensível ao desbalanceamento de classes é a pontuação F1, pois sua medida considera a precisão e a revocação para calcular sua pontuação. A pontuação F1 é a média harmônica da precisão e revocação, sua fórmula é:

$$\text{f1} = 2 \times \frac{\text{prc} \times \text{rec}}{\text{prc} + \text{rec}}. \tag{2.11}$$

As métricas revocação e pontuação F1 não quantificam somente o total de amostras classificadas corretamente para classe positiva, mas também dão peso à quantidade de amostras de classes negativas atribuídas a classe positiva pelo classificador.

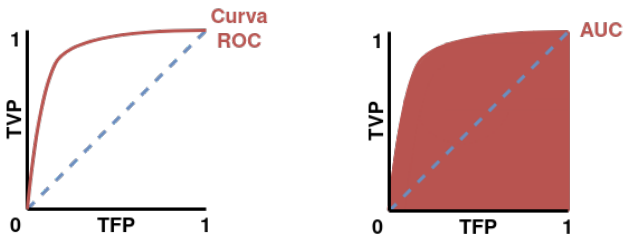
Precisão, revocação e pontuação F1 podem ser empregadas para problemas binários ou multiclasse, nesse último, a média dos valores obtidos para cada classe é calculada.

Outra métrica bastante utilizadas para avaliação de classificadores binários são a curva ROC (*Receiver Operating Characteristic*) e sua área AUC (*Area Under ROC Curve*). A curva ROC representa a razão entre revocação e Taxa de Falsos Positivos (TFP) ou $TVP \times TFP$, onde,

$$\text{TFP} = \frac{FP}{FP + VN}. \tag{2.12}$$

Em um mundo ideal, espera-se que a medida que a revocação cresce em seu eixo, o valor de TFP se mantenha baixo, podendo crescer lentamente. Essa característica indica a boa capacidade de classificação do modelo avaliado. Por outro lado, se revocação e TFP crescerem igualmente, significa que o classificador está tendo problemas para distinguir amostras de cada classe.

A curva ROC descreve a precisão que um classificador possui para distinguir um binômio [Ray et al., 2006]. Além disso, a curva representa o *tradeoff* entre revocação e TFP, de tal forma que um bom classificador apresenta baixa TFP e alta revocação. A Figura 2.12a apresenta o exemplo da curva ROC de um classificador. Para simplificar a curva ROC, a métrica AUC utiliza a área abaixo da curva ROC (Figura 2.12b) como valor para avaliação de classificadores e pode ser obtida através da integral da curva ROC.



(a) Exemplo de curva ROC (b) Exemplo de AUC

Figura 2.12: As métricas ROC e AUC derivam dos medidas TVP e TFP. Pode-se concluir que a métrica AUC é uma forma de representar a curva ROC.

As métricas apresentadas possuem valores ótimos quando alcançam 1 e valores péssimos quando chegam a 0. O AUC possui uma interpretação ligeiramente diferente das demais, neste, quando seu valor está próximo de 0.5 significa que o classificador avaliado não é melhor do que um classificador que se baseia nas probabilidades de cara e coroa ao jogar uma moeda não viciada para escolher as classes. Isso significa que o modelo avaliado não é superior a um modelo que define 50% de chance de uma amostra pertencer a cada classe.

2.9 Coleta de dados e ferramentas Bioinformáticas

Após o desenvolvimento de métodos de alto rendimento para o sequenciamento de RNA e DNA, a quantidade de sequências disponíveis para estudo cresceu significativamente [Schuster, 2008]. Tais métodos possibilitam que as sequências sejam armazenadas em arquivos no formato de texto, tornando possível sua análise por métodos computacionais [NCBI, 2020]. O portal *National Center for Biotechnology Information - NCBI* contém, atualmente, a maior coleção de dados públicos de sequências biológicas, incluindo RNA, DNA e proteínas de diversos organismos conhecidos que variam de plantas até vírus e bactérias. Além disso, o portal executa revisões esporádicas de dados. As sequências são disponibilizadas em diversos formatos de arquivo, entretanto, os mais utilizados são *FASTA* e *GenBank*.

Os arquivos *FASTA* são estruturas de dados simples usadas para representar sequências, contendo somente as sequências propriamente ditas, o identificador destas e eventualmente meta-dados adicionais disponibilizadas pelos autores dos sequenciamentos. Por outro lado, os arquivos *GenBank* são ricos em metadados, como por exemplo, localização e data de coleta, características genômicas, posições de inicio e fim de síntese proteica, taxonomia e tipo de organismo hospedeiro (em casos de sequências virais). Estas informações frequentemente transformam-se em rótulos para as sequências. Outra vantagem dos arquivos *GenBank* é que, em caso de sequências de RNA ou DNA, apresentam também sua tradução proteica.

Nossa base de dados foi estruturada a partir de diversos arquivos *GenBank*. A ferramenta de análise sequencial *UGENE* versão 36.0 [Okonechnikov et al., 2012] foi empregada no processo de alinhamento das sequências *GenBank* e sua transcriação para arquivos de dados no formato CSV (*Comma-separated values*). Adicionalmente, as amostras passaram por um processo de análise filogenética pela ferramenta *online Genome Detective*, obtendo-se os metadados de genótipos das amostras e novos metadados de sorotipo que puderam ser comparados com os sorotipos originalmente

declarados para cada amostra.

2.10 Considerações finais

Neste capítulo foram apresentadas descrições sobre a estrutura de sequências biológicas e outros conceitos necessários para compreensão do restante deste trabalho. Com todas as informações repassadas nesta capítulo, observa-se que o processamento de sequências biológicas não é uma tarefa trivial. Nesta seção, foi explicada a estrutura de matriz de co-ocorrências que será utilizada como um representante e atenuador de padrões de subestruturas sequenciais. Também foram descritos conceitos fundamentais de *Random Forest* e *CNN* que serão os classificadores utilizados nesse trabalho. Os conceitos e funcionamento do método *SHAP* para interpretação de modelos foram apresentados, visto que estes são de fundamental importância para investigar a importância dos padrões encontrados. Sendo assim, no próximo capítulo será exibido um conjunto de trabalhos relacionados que empregam diferentes técnicas de representação de sequências para diferentes fins.

Capítulo 3

Trabalhos relacionados

Este capítulo apresenta os trabalhos recentes que utilizam diferentes técnicas de representação de sequências biológicas através da sua codificação em vetores de *embeddings* por meio de métodos de PLN ou na codificação direta de sequências em estruturas matriciais. Os trabalhos apresentados neste capítulo resolvem problemas diferentes que, no entanto, estão relacionados com as etapas de representação e classificação de sequências biológicas. Por fim, será apresentada uma discussão sobre os trabalhos apresentados.

3.1 Representação de sequências por vetores de *embeddings*

Na literatura é possível encontrar diversos trabalhos que empregam modelos de representação de palavras oriundos do campo de PLN para representação de sequências biológicas. Essa associação entre sequências biológicas e linguagens naturais pode ser compreendida com mais clareza na seção 2.3. De fato, a abordagem PLN para pré-processamento e classificação de sequências biológicas vem crescendo no cenário de pesquisa atual, como vamos ver a seguir.

O método *BioVec* proposto por Asgari [2015] é um dos trabalhos mais evidentes entre os que empregam modelos de PLN para representação de sequências. O objetivo deste trabalho é apresentar a capacidade de *embeddings* codificarem relações estruturais aptas a caracterizarem sequências biológicas. Para isso, os autores desenvolvem uma metodologia para reconhecimento de 324 mil sequências de proteína divididas em 7 mil classes de famílias proteicas.

Primeiramente as subestruturas de cada sequência são obtidas pelo método *k-Mer*.

O algoritmo *k-mer* tornou-se muito difundido no campo de análise computacional genética por considerar sobreposições gênicas, pois este gera subsequências considerando sobreposições de nucleotídeos ou aminoácidos e, dessa forma, gerando sequências ocultas [Itzkovitz et al., 2010]. A Imagem 3.1 mostra códons obtidos através da aplicação do algoritmo *k-mer*, para $k = 3$ em uma sequência com 12 bp, além de k sequências ocultas. Observe que as primeiras palavras das sequências ocultas 2 e 3 possuem pelo menos um nucleotídeo contido na primeira subsequência da sequência oculta 1, ou seja, as sequências 2 e 3 possuem genes que se sobrepõem à sequência 1.

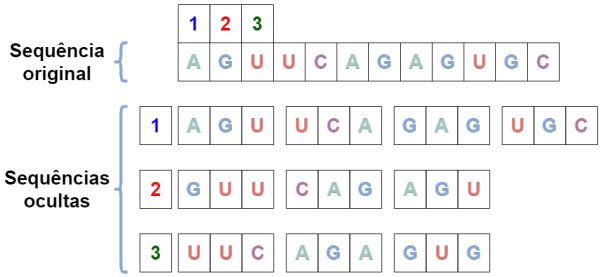


Figura 3.1: Três conjuntos de códons sobrepostos gerados pelo algoritmo *k-mer*, para $k = 3$. O valor k representa tando o tamanho da subestrutura quanto a quantidade de sequências desejadas.

O algoritmo *k-mer* (2) recebe como entrada as sequências biológicas originais X e um valor k que representa o tamanho da subsequência e, também, a quantidade de sequências ocultas desejadas. Para que um dicionário geral de subsequências seja gerado, o algoritmo itera sobre cada sequência oculta P contida em X. Por exemplo, a primeira sequência oculta tem seu inicio no primeiro caractere da sequência original, a segunda se inicia no segundo caractere e a k-ésima tem seu inicio no k-ésimo caractere da sequência original. Após obter a posição de inicio de uma subsequência, o algoritmo começa a etapa geração de sequências e formação do dicionário. A complexidade do algoritmo *k-mer* é $O(k|X||P|)$, onde $|P|$ é o tamanho das sequências originais e $|X|$ é a quantidades de sequências que o algoritmo recebe.

Os autores utilizam $k = 3$ para o algoritmo *k-Mer*, assinalando que a estrutura utilizada para geração de resultados será o códon (Seção 2.2). Dessa forma, cada sequência de entrada é representada por três conjuntos de códons (uma sequência original e duas subsequências ocultas). Em seguida, os conjuntos de códons são passados para o modelo *word2vec Skip-Gram* [Mikolov, 2012] para geração de *embeddings* de códons. O modelo *word2vec* consiste em uma rede neural rasa e aprende um vetor de relações entre uma palavra e as palavras no seu contexto ao realizar predições das

Algoritmo 2 *k-mer*

Entrada 1: $X \leftarrow$ Sequências biológicas
Entrada 2: $k \leftarrow$ Tamanho da subsequência
Saída 1: Dicionário de subsequências
Saída 2: k conjuntos de subsequências para cada sequência
$z \leftarrow 0$
$D \leftarrow \{\}$
$S \leftarrow \{\}$
para cada sequência P em X **faça**
    para $i \leftarrow 1$ até k **faça**
        $M \leftarrow \{\}$
        $l \leftarrow 1$
        para $j \leftarrow 1$ até P.tamanho **faça**
            se $P[l : (l + k - 1)] \notin D$ **então**
                $D.insere(P[(l + z) : (l + k - 1)])$
            fim se
            $M.insere(P[(l + z) : (l + k - 1)])$
            $l \leftarrow l + k$
        fim para
        $S.insere(M)$
        $z \leftarrow z + 1$
    fim para
fim para
Retorna D (Dicionário de subsequências) e S (k conjuntos de subsequências para cada sequência)

palavras de contexto dado uma palavra principal (Figura 3.2). No método *BioVec* o modelo aprende relações entre códons próximos.

Portanto, os códons são representados por um vetor de *embedding* individual. Cada uma das sequências é representada por três conjuntos contendo códons, logo, cada sequência é representada por três conjuntos de *embeddings*. Diante disso, os autores propõem a soma dos vetores de *embeddings* de cada conjunto e a concatenação desses vetores resultantes. A equação abaixo resume essa lógica,

$$C_p = \bigcup_{i}^{k} \sum_{j}^{n} B_i j, \tag{3.1}$$

onde, k é o valor de *k-Mer*, n é a quantidade de *embeddings* no conjunto i, $B_i j$ é um vetor de *embedding* e C_p é o vetor de representação da sequência p. Por fim, os vetores C são passados para o classificador *Support Vector Machine* (SVM) que obteve 93% de acurácia.

Como podemos observar, o método *BioVec* utiliza uma estatística resumida para

representação de sequências, tornando o método ideal para pré-processamento de grandes quantidades de dados e sequências muito extensas, pois a operação apresentada na Equação 3.1 diminui significativamente a complexidade de espaço do problema ao reduzir a matriz representação $M_{d\times n\times k}$ para um vetor $V_{1\times h}$, onde d é o tamanho do vetor de *embedding* e $h = d \times k$ e complexidade de tempo de classificação.

Essa compreensão das matrizes de representação torna praticamente inviável a identificação de padrões nas sequências brutas associadas classe, pois, ao somar os vetores de *embedding* perdemos o controle de observação da significância de subestruturas específicas na etapa de classificação. Além disso, a operação de soma de *embeddings* pode gerar vetores extremamente distintos para mesma classe, casos esta apresente sequências incompletas. *BioVec* será empregado como *baseline* de nossa pesquisa, onde desejamos observar a capacidade de representação entre o método proposto e *BioVec* para sequências proteicas da dengue.

O *ProtVecX*, proposto por Asgari et al. [2019] é empregado na descoberta de *motifs*[3] proteicos e classificação de proteínas de ligação. Diferente do *BioVec* que utiliza *k-Mer* como uma das etapas de pré-processamento de sequências, *ProtVecX* aplica o método PPE *(Peptide-pair Enconding)*, que consiste em uma modificação do algoritmo de compactação de texto BPE *(Byte-pair Enconding)* através da adição de uma *estrutura de amostragem.*[2] PPE é utilizado para segmentar sequências de diversas maneiras, obtendo codificações de pares de peptídeos presentes em uma sequência, que serão utilizados nas tarefas de descoberta de *motifs* proteicos e vetores de *embeddings*. Testes estatísticos são utilizados sobre as sequências PPE para identificar *motifs* discriminativos, enquanto que, as os vetores de *embedding* são obtidos através do modelo *word2vec - SkipGram* que recebe as sequências PPE como entrada de dados.

A Figura 3.2 ilustra a arquitetura do modelo *word2vec* e a geração dos vetores de *embeddings* W. Observe que o modelo recebe como entrada a subestrutura de sequência *svsr* e faz previsões das subestruturas que estão ao seu redor (contexto). *ProtVecX* possui a mesma arquitetura de *BioVec* [Asgari, 2015], mudando apenas a forma de tratamento inicial das sequências. Os resultados obtidos com as representações geradas por *ProtVecX* mostram que, na tarefa de descoberta de *motifs*, o modelo não consegue manter bom desempenho para diferentes bases de dados, apresentando precisões que variam entre 0% e 100%. Por outro lado, o algoritmo SVM, utilizado para classificação das proteínas de ligação, alcançou precisões entre 73% e 100% na classificação de três

motifs[3]: Um *motif* de sequência é um padrão de nucleotídeos ou aminoácidos que é difundido e tem, ou é conjecturado como tendo, um significado biológico.

estrutura de amostragem.[2]: Conjunto de características e atributos dos elementos de um universo que deverão estar presentes nos elementos que compõem uma amostra.

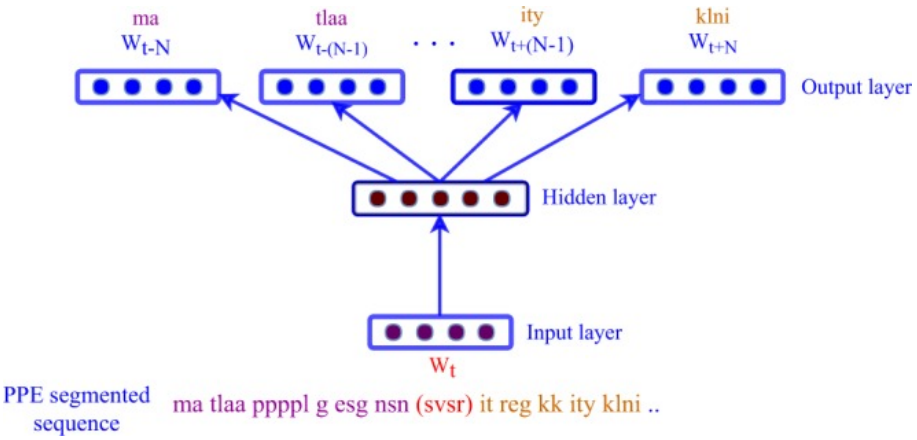


Figura 3.2: Extração de *embeddings* utilizando o modelo *word2vec*. Fonte: Asgari et al. [2019].

bases de proteínas de ligação.

O método *iDeepV* para detecção de *proteínas de ligação ao RNA*[4] é proposto por Pan and Shen [2018a]. O algoritmo *k-Mer* é empregado na etapa de pré-processamento das sequências, obtendo-se conjuntos se subsequências. Em seguida, os autores utilizam o modelo *word2vec - SkipGram* para representação das sequências e uma CNN-1D para classificação dos vetores de *embeddings*. Neste ponto da metodologia, a compreensão de matriz de *embeddings* é idêntica a realizada em *BioVec* e pode ser explicada pela equação 3.1. A arquitetura do método *iDeepV* é apresentada na Figura 3.3.

Por fim, os vetores de *embeddings* que representam as sequências alimentam uma CNN-1D, classificando cada sequência como *proteína de ligação* ou não. Os autores avaliaram o método em 24 bases de dados, apresentando uma melhoria em 10 delas ao se comparar com 3 outros *baselines* e alcançando um valor médio para AUC igual a 0.913. Uma das vantagens deste método está na arquitetura do classificador, pois, as CNN-1D são capazes de aprender longas dependências em estruturas sequenciais.

Hamid and Friedberg [2019] apresentam um trabalho interessante, pois seu método dispõe de um classificador específico para dados sequenciais. O método proposto utiliza *embeddings* de subestruturas gerados pelo modelo *word2vec - SkipGram* e uma *Recurrent Neural Network* (RNN) [LeCun et al., 2015] para classificar peptídeos antimicrobianos. Primeiramente, as sequências são pré-processadas pelo algoritmo *k-Mer*. No entanto, os autores convertem os k conjuntos de subestruturas geradas pelo *k-Mer* para uma sequência em um único conjunto ordenado. Em seguida, o modelo *word2vec - SkipGram*, extrai um conjunto $X = \{X_{(t-1)}, X_{(t),X_{(t+1)},\ldots,X_{(t+n)}}\}$ de vetores de *embed-*

proteínas de ligação ao RNA[4]: Proteínas de ligação ao RNA são proteínas que se ligam ao RNA de fita dupla ou única nas células e participam na formação de complexos de ribonucleoproteínas.

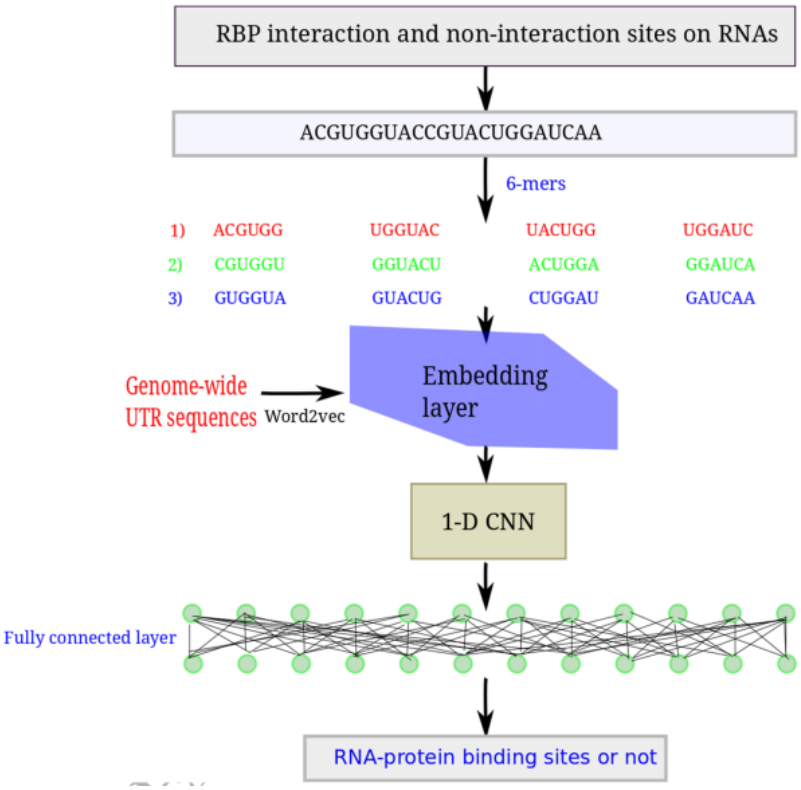


Figura 3.3: Arquitetura do método *iDeepV*. Fonte: Pan and Shen [2018a]

dings. Por fim, esses vetores alimentam a RNN para classificação das sequências. A Figura 3.4 ilustra parte da RNN proposta recebendo os *embeddings* para classificação. Na Figura, também fica claro as subestruturas são lidas na ordem da sequência original, no entanto, apresentando sobreposições de dois aminoácidos em cada subestrutura. Esse modo de leitura das subestruturas geradas pelo *k-Mer* é diferente da leitura tradicional e não considera a ordenação das subestruturas nas sequências ocultas, apenas considera suas sobreposições.

Neste trabalho, os autores tratam as sequências biológicas como séries temporais, onde, cada subestrutura representa um instante da série. Na Figura 3.4, U, V e W são pesos da rede, X_k representa o *embedding* associado à subestrutura no tempo k. Além disso, o neurônio h_k é um estado oculto associado ao instante k que contém informações das entradas anteriores e da atual. Esse mecanismo funciona como a memória da rede, possibilitando que a RNN preserve informações a longo prazo sobre *embeddings* processados em instantes anteriores. Por fim, a rede produz a probabilidade da sequência ser um peptídeo antimicrobiano ou não.

A metodologia proposta atingiu 95.8%, 94.6% e 94.7% para acurácia, revocação e pontuação F1, respectivamente. Também foi avaliado o desempenho de outros 9

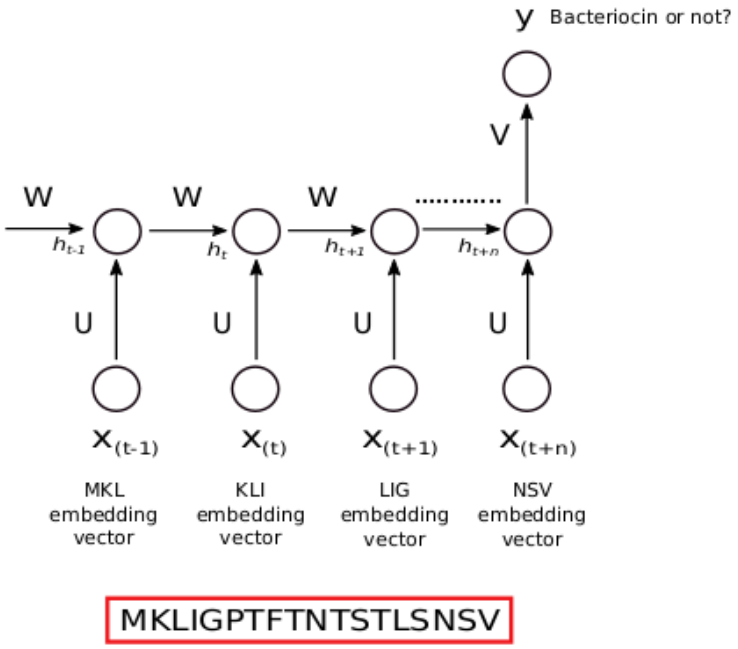


Figura 3.4: Arquitetura de uma RNN para classificação de peptídeos antimicrobianos. Fonte: Hamid and Friedberg [2019]

métodos na mesma base de dados. O método proposto obteve o melhor resultado quando as métricas de avaliação foram revocação e pontuação F1, por outro lado foi o terceiro melhor modelo em relação a acurácia, perdendo para os modelos BLAST e HMMER, que são baseados em cadeias ocultas de Markov (HMM) [Altschul et al., 1997, Eddy et al., 1995]. O desbalanceamento da base pode justificar a acurácia obtida pelo modelo proposto que, apesar ser terceira melhor, se mantém próxima aos dois melhores (BLAST: 97.2% e HMMER: 98.9%).

Zhang and Kabuka [2020] também utilizam o modelo *word2vec - SkipGram* para representação de sequências via vetores de *embeddings* de subestruturas geradas pelo algoritmo *k-Mer* e uma CNN-1D para classificação de famílias de proteínas. A arquitetura proposta pelos autores consiste em dois módulos, o primeiro módulo recebe a sequência e extraí os *embeddings* que, posteriormente, são passados para o segundo módulo que contém o classificador. Os autores avaliaram o desempenho da arquitetura em uma base de dados com 55,077 sequências proteicas distribuídas em 60 famílias, tal que as sequências podem apresentar tamanhos variando entre 50 e 1200 nucleotídeos. O modelo CNN obteve resultados superiores comparado com os modelos SVM, LSTM, LSTM Bidirecional e GRU *(Gated Recurrent Unit)*, alcançando 97.77% de pontuação F1.

Le et al. [2019a] propõem um método para identificação de *acentuassomos*[5], para tal, utilizam informações ocultas de sequências de DNA através da regra de 5 etapas de Chou e *embeddings* de subestruturas. As 5 etapas de Chou [Chou, 2011] foram úteis para refletir características essenciais que estão profundamente escondidas em sequências proteicas de difícil compreensão. O modelo *fastText* [Bojanowski et al., 2016] foi utilizado para representação das subestruturas geradas via *k-Mer*. A metodologia obteve 82.3% de acurácia na classificação de *acentuassomos* em sequências proteicas, onde, SVM foi utilizado como classificador. Diferente do *word2vec* o modelo *fastText* consegue fazer aproximações de *embeddings* para palavras não existentes no vocabulário, ou seja, é capaz de gerar vetores de *embeddings* para subestruturas que nunca foram treinadas.

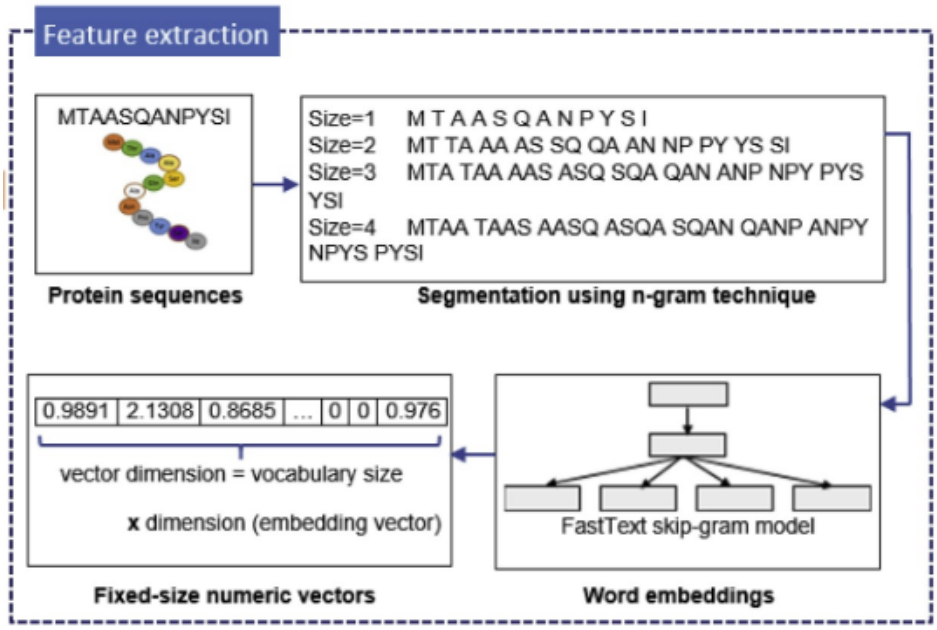


Figura 3.5: Representação de subestruturas utilizando o modelo *fastText*. Fonte: Ho et al. [2019].

Ho et al. [2019] também utilizam o modelo *fastText* para obter vetores de *embeddings*. Os *embedding* são utilizados para classificação de substratos de *proteínas transportadoras*[6]. A Figura 3.5 ilustra os processos básicos da metodologia. Ressalta-se que neste trabalho os *embeddings* tem dimensão diferente de outros, visto que as

acentuassomos[5]: Um acentuassomo é uma sequência de nucleotídeos de DNA específica ao qual se podem ligar proteínas que aumentam os níveis de transcrição.

proteínas transportadoras[6]: Os transportadores são uma das principais classes de proteínas de membrana que facilitam o movimento de substratos hidrofílicos através das membranas hidrofóbicas dentro e entre as células.

dimensões mais utilizadas para *embeddings* são valores entre 100 e 500. Neste, os autores utilizam *embeddings* de uma dimensão, ou seja, cada vetor contém apenas único valor numérico. Ho et al. [2019] eliminam sequências proteicas que apresentassem mais de 20% de similaridade com outras sequências, resultando na utilização de apenas 1050 proteínas de 3853. Os autores também utilizaram vetores de frequência de palavras biológicas em conjunto com os vetores de *embeddings* com o objetivo de disponibilizar informações extras para o modelo de classificação. Os vetores de frequência possibilitam que padrões específicos ocorrência de palavras biológicas nas sequências de interesse sejam revelados e aprendidos pelo classificador.

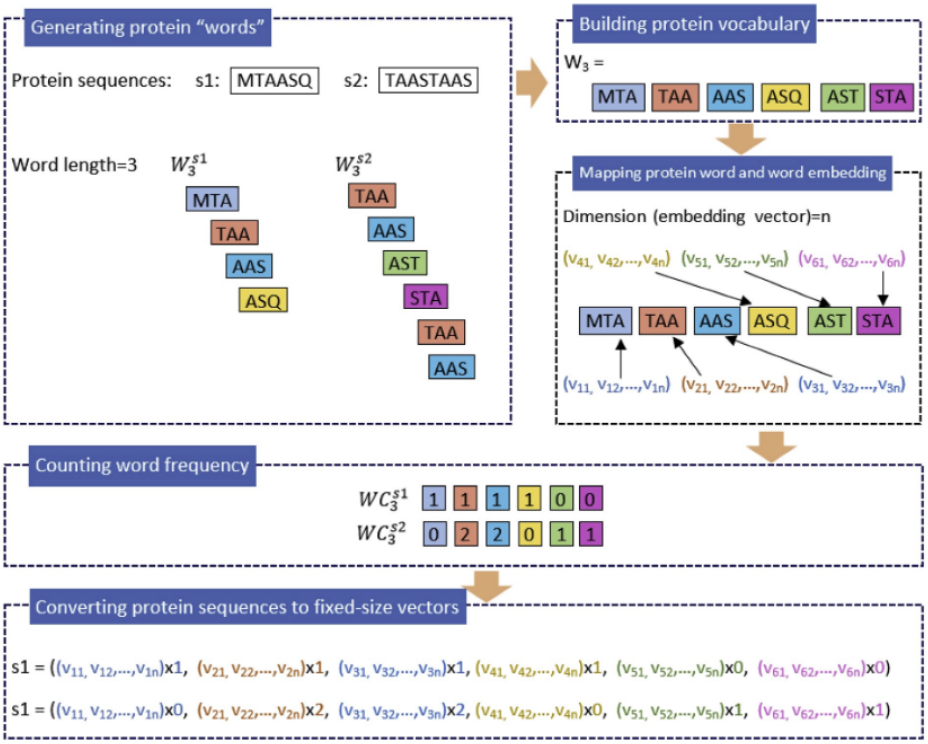


Figura 3.6: Diagrama da metodologia proposta por Ho et al. [2019]. Fonte: Ho et al. [2019]

A Figura 3.6 apresenta o diagrama da metodologia proposta por Ho et al. [2019]. Primeiramente as subestruturas são obtidas pelo algoritmo *k-Mer*, em seguida, um vocabulário de subestruturas é gerado. Após estas etapas, o modelo *fastText* codifica as subestruturasdo vocabulário em vetores de *embeddings* individuais. Os vetores de frequência de palavras para cada sequência são obtidos na quarta etapa do método. Com posse dos vetores de frequências de subestruturas e os vetores de *embeddings* das subestruturas o método segue para a última etapa de representação que é a multiplicação da frequência de cada subestrutura pelo seu *embedding*. Por fim, os vetores desses

produtos são passados para o classificador. Esta é uma ótima abordagem para resolver o problema de heterogeneidade do tamanho de sequências, no entanto, a estrutura genética original de ocorrência de palavras biológicas é perdida. Ho et al. [2019] dispõem de alguns classificadores binários para classificação de *transportadores* proteicos, dentre eles, os que apresentaram valores superiores à 99% para acurácia, sensibilidade e especificidade foram os modelos SVM e *Random Forest*.

Villmann et al. [2019] apresenta um método capaz de distinguir sequências reais e sintéticas de RNAs virais. Para tal, os autores optam por trabalhar com medidas de similaridade entre sequências obtidas por diferentes métodos. A primeira matriz de similaridades consiste na média das distâncias *NCD (Normalized Compression Distance)*. A segunda matriz de similaridades é obtida através do cálculo de distância Euclidiana ou divergências de *Kullback-Leibler (KL)* entre vetores de *embeddings* de cada sequência, que são obtidos através do método *BoW (bag-of-words)* [Harris, 1954], tal que, o *embedding* de uma sequência é a frequência das palavras biológicas desta. Por fim, utiliza-se o classificador *GLVQ* (*Generalized Learning Vector Quantization*) [Sato and Yamada, 1996] que classifica amostras baseando-se em dissimilaridades. Portanto, as matrizes de similaridades obtidas pelos métodos *NCD* e *BoW* (distância euclidiana ou KL) são passadas como dados de treino para o classificador *GLVQ*. Dentre as as abordagens avaliadas, observou-se que a matriz de distâncias Euclidiana entre os *embeddings* gerados pelo método *BoW* conseguiu representar de maneira mais adequada as sequências de RNA e, consequentemente, o modelo *GLVQ* alcançou seu melhor resultado, apresentando 76.1% de acurácia.

Nesta seção apresentamos alguns trabalhos que abordam diferentes problemas de bioinformática relacionados à sequências biológicas e que empregam modelos PLN para representação de sequências através de vetores de *embeddings*. Portanto, mostramos que diversos autores abordam o problema de representação de sequências biológicas com métodos de PLN. Apesar de recentes, estas abordagens tem alcançado bons resultados e em alguns casos superaram métodos clássicos de classificação de sequências. No entanto, estas abordagens possuem algumas limitações, tais como, a dificuldade de métodos como *word2vec* e *fastText* conservarem estatísticas globais de subestruturas de sequências distintas e a complexidade computacional de utilizar vetores esparsos como fonte de dados primária para modelos de representação. O longo comprimento das sequencias biológicas é outro desmotivador para utilizar abordagens de PLN. Tendo isso em vista, outros autores propõem a conversão de sequências biológicas em estruturas matriciais capazes de coletar e atenuar padrões de subestruturas ocultos ou pouco evidentes, como será visto na próxima seção.

3.2 Transformação de sequências em matrizes

Os trabalhos apresentados na seção anterior seguem uma abordagem oriunda de PLN, onde cada sequência e suas subestruturas podem ser tratadas, respectivamente, como uma frases e palavras de linguagem natural para, por fim, encontrar representações vetoriais que consigam descrever a estrutura com suas funções internas e externas. Nesta seção, serão apresentados trabalhos que utilizam técnicas para transformar dados sequenciais em matrizes, com o objetivo de capturar padrões funcionais de difícil acesso capazes de caracterizar sequências em tarefas de classificação.

Zeng et al. [2016] apresentam uma exploração sistemática de CNN em 690 experimentos de classificação de *proteínas de ligação ao DNA*[6] em várias bases com diferentes aspectos de informação. A exploração é realizada através da análise de desempenho de CNN-2D quando a largura e profundidade de cada uma é alterada, além de observação dos resultados após inserção e remoção de *kernels* e utilização diferentes parâmetros nas camadas de convolução. As sequências de DNA são representadas por matrizes de codificação binária (*one-hot*) com dimensões $L \times 4$, onde L é o tamanho da sequência e 4 representa os quatro tipos de nucleotídeos existentes. Por exemplo, a sequência {AGTTGC} pode ser representada pela matriz presente na Tabela 3.1.

Sequência	Nucleotídeos			
	A	C	G	T
A	1	0	0	0
G	0	0	1	0
T	0	0	0	1
T	0	0	0	1
G	0	0	1	0
C	0	1	0	0

Tabela 3.1: Matriz *one-hot* da sequência biológica {AGTTGC}. Cada coluna desta matriz representa um dos nucleotídeos existes.

Os autores concluem que os efeitos da arquitetura das CNN é específico para cada tarefa. Zeng et al. [2016] observam que a adição de *kernels* na arquitetura de modelos que treinam dados baseados em *motifs* ajuda na classificação de *proteínas de ligação ao DNA*. Também demostram a capacidade das CNN aprenderem recursos avançados de uma sequência, como contexto local em sequências. Zeng et al. [2016] evidenciam, em seu trabalho, que CNN são capazes de aprender informações sofisticadas presentes em sequências biológicas, e que, a complexidade das arquiteturas de CNN depende da

proteínas de ligação ao DNA[6]: Estas proteínas possuem domínios de ligação ao DNA e, portanto, apresentam afinidade com o DNA.

tarefa a ser realizada e da quantidade de dados de treino disponível, comprovando que redes mais complexas não geram melhores resultados para todos problemas.

No trabalho de Pan and Shen [2018b] é apresentado o método *iDeepE*, capaz de identificar proteínas de ligação em sequências de RNA através de uma CNN. Os autores utilizam duas redes, denominadas CNN Global e CNN Local, que são de canal único e multi-canal, respectivamente. As CNN exigem que suas entradas tenham tamanho fixo, no entanto, sequências biológicas raramente atendem esse requisito. Para contornar este problema Pan and Shen [2018b] fazem um pré-processamento nas sequências que serão classificadas. Os passos de pré-processamento incluem: 1) para a CNN Global, todas as sequências são preenchidas até alcançarem o tamanho da maior sequência do conjunto de treino; e 2) para a CNN Local, aplicando-se o algoritmo *k-Mer* com duas sobreposições de nucleotídeos, cada sequência é dividida em diversas subestruturas de tamanho 4, tal que, cada subestrutura é considerada um canal para CNN. Após a etapa de pré-processamento, cada sequência é codificada em uma matriz *one-hot*, seguindo critérios especificados pelos autores.

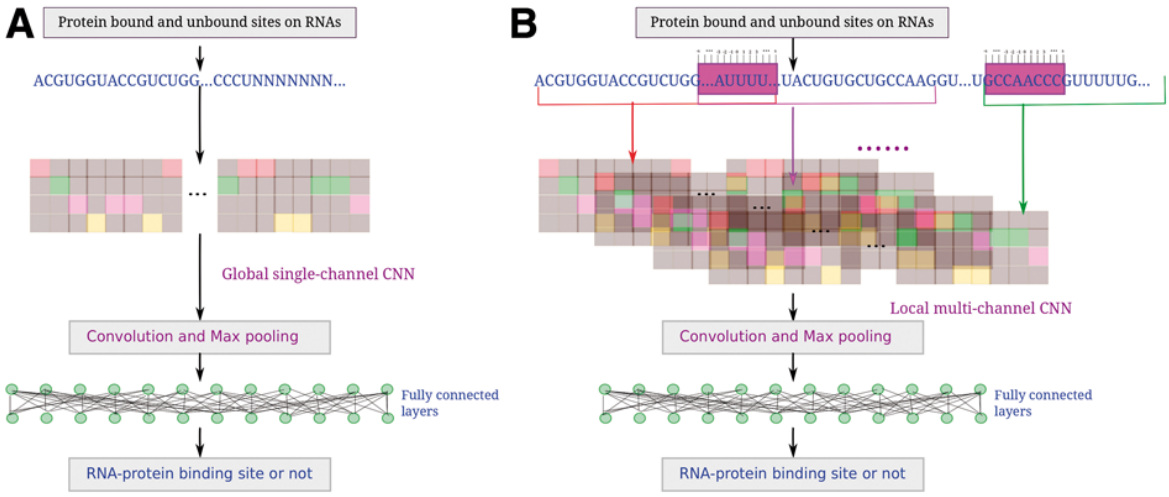


Figura 3.7: Fluxograma do método *iDeepE*. Os retângulos representam as matrizes *one-hot*, tal que, cada quadrado cinza representa o valor 0 e os coloridos 1. Fonte: Pan and Shen [2018b].

A Figura 3.7 apresenta a metodologia proposta por Pan and Shen [2018b]. Vale ressaltar que as probabilidades de classificação obtidas pelas CNN Global e CNN Local são combinadas para produzir as probabilidades finais de classificação, essa combinação é feita através do calculo da média das probabilidades de saída do modelo local e global. A metodologia foi testada em 24 bases de dados referentes à proteínas de ligação e, quando comparado com outros 5 métodos, *iDeepE* obteve melhor resultado em 17

bases, alcançando, em média, 0.937 de AUC.

O método *DeepInsight* [Sharma et al., 2019] transforma sequências de expressão gênica em matrizes através de uma transformação $T(g) : \mathbb{R}^{d\times 1} \rightarrow \mathbb{R}^{p\times q}$, como ilustrado na Figura 3.8. Sequências de expressão gênica são similares as sequências genéticas (RNA e DNA), no entanto, estas não são formadas por caracteres, mas por valores reais. Apesar desta grande diferença entre sequências de expressão gênica e sequências genéticas, é possível aplicar a metodologia *DeepInsight* em sequências genéticas após o pré-processamento destas. As localizações dos valores de expressões gênicas g_j dentro da matriz M dependem de medidas de similaridade. Por exemplo, seja a sequência de expressão $x = \{g_1, g_2, g_3, g_4, g_5, g_6, ..., g_d\}$, caso sejam calculadas similaridades significantes entre os valores g_1, g_3, g_6e g_d de x, estes serão agrupados em coordenadas próximas dentro da matriz M, conforme mostra a Figura 3.8.

Seja a matriz de amostras $X = \{x_1, x_2, ..., x_n\}$ e $x_i = \{g_1, g_2, ..., g_d\}$ uma sequência de expressão gênica, após determinar a localização dos valores g_d de x_i na matriz M_i, os valores g_d características serão mapeados para M_i. O método irá gerar matrizes exclusivas para cada amostra de X. Com isso, serão geradas n matrizes $M_{p\times q}$, tal que, $d = p \times q$, onde p e q são quaisquer valores inteiros que tenham produto igual a d. Esse conjunto de n matrizes poderá, posteriormente, ser utilizado para treino de CNN.

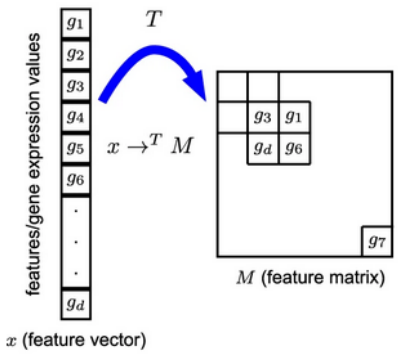


Figura 3.8: Codificação de uma sequência de expressão gênica em matriz através de uma transformação T. Fonte: Sharma et al. [2019].

Para testar a metodologia proposta, os autores utilizaram a base de dados RNA-seq que contém 6216 sequências de expressões gênicas divididas em 10 classes de cânceres, tal que, cada sequência apresenta 60483 valores. A Figura 3.9 ilustra o classificador proposto por Sharma et al. [2019], o modelo consiste de duas redes convolutivas em paralelo que apresentam a mesma profundidade e quantidade de filtros, no entanto, possuem tamanhos distintos de *kernels*. A classificação das matrizes geradas pelo método através da arquitetura da Figura 3.9 apresentou 99% de acurácia, o resultado alcan-

çado é superior a outros classificadores que utilizaram sequencias sem a transformação matricial.

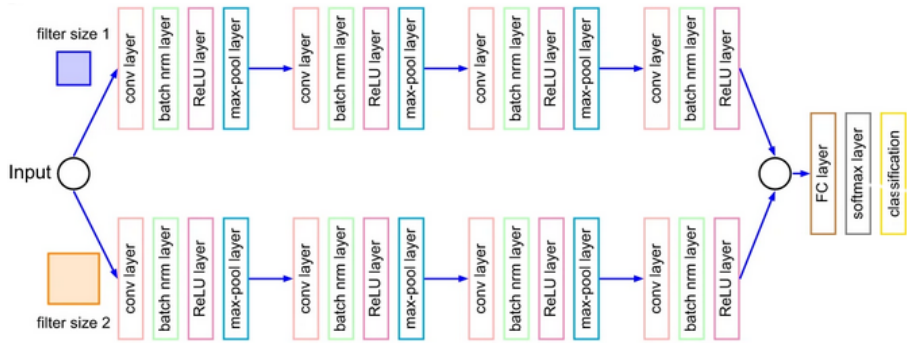


Figura 3.9: Arquitetura paralela de CNN usada para classificar as imagens geradas pela metodologia proposta. Fonte: Sharma et al. [2019].

Le et al. [2019b] propõem o método *iMotor* baseado em CNN e na regra de 5 etapas de Chou [Chou, 2011] para identificar funções moleculares de proteínas motoras do citoesqueleto. A perda funcional de uma molécula específica de uma proteína motora pode estar relacionada à diversas doenças humanas, tais como síndrome de Kartagener e distrofia muscular de Dechenne. Portanto, um modelo que consiga classificar precisamente essas proteínas é de grande valia para área médica e biológica. Os autores utilizam o algoritmo PSI-BLAST [Altschul et al., 1997] para obter a Matriz de Pontuação Específica de Posição (*Position-Specific Scoring Matrix* – PSSM) [Stormo et al., 1982] das sequências. As PSSM geralmente são derivadas de um conjunto de sequências alinhadas funcionalmente relacionadas, e acabam tornando-se parte de muitas ferramentas de bioinformática para a descoberta de *motifs*. Para obtenção da matriz PSSM, primeiramente, deve-se descobrir a matriz de *score* de posição (*Position Frequency Matrix* – PFM) de cada nucleotídeo nas sequências. Sejam os nucleotídeos conhecidos {A, C, G, T} e as sequências $S_1 = \{\text{AGTGTCGAC}\}$ e $S_2 = \{\text{AGTCGGAAA}\}$, ambas com comprimentos iguais à 9, a matriz PFM de S_1 e S_2 pode ser visualizada na Tabela 3.2

Nucleotídeos conhecidos	Posição								
	1	2	3	4	5	6	7	8	9
A	2	0	0	0	0	1	1	2	1
C	0	0	0	1	0	1	0	0	1
G	0	2	0	1	1	0	1	0	0
T	0	0	2	0	1	0	0	0	0

Tabela 3.2: Matriz PFM de S_1 e S_2.

Para gerar a matriz PSSM, é necessário converter a matriz PFM em uma matriz de probabilidades de posição dos nucleotídeos. Esta matriz pode ser adquirida dividindo os valores da matriz PFM pela quantidade de sequências, obtendo-se então a matriz de probabilidades PFM, presente na Tabela 3.3.

Nucleotídeos	Posição								
conhecidos	1	2	3	4	5	6	7	8	9
A	1	0	0	0	0	0.5	0.5	1	0.5
C	0	0	0	0.5	0	0.5	0	0	0.5
G	0	1	0	0.5	0.5	0	0.5	0	0
T	0	0	1	0	0.5	0	0	0	0

Tabela 3.3: Matriz de probabilidades PFM de S_1 e S_2.

Por fim, calcula-se os *scores* da matriz PSSM utilizando a equação abaixo,

$$\text{PSSM}_{jk} = 2\log_2\left(\sum_{i=1}^{4} P_{ik}\frac{Z_{ij}}{q_i q_j}\right), \tag{3.2}$$

onde, P_{jk} é a probabilidade do nucleotídeo j na posição k, Z_{ij} é a probabilidade de substituição do nucleotídeo i com o nucleotídeo j, q_i e q_j são as probabilidades esperadas dos nucleotídeos i e j. Para nucleotídeos assume-se que todos tem as mesmas probabilidades de ocorrência em uma posição, ou seja, $q_i = 0.25$. A matriz de substituição utilizada para nucleotídeos pode ser visualizada na Tabela 3.4.

	A	C	G	T
A	0.13	0.03	0.03	0.06
C	0.03	0.13	0.06	0.03
G	0.03	0.06	0.13	0.03
T	0.06	0.03	0.03	0.13

Tabela 3.4: Matriz de substituição de nucleotídeos. Apresenta as probabilidades de trocas entre nucleotídeos em uma posição da sequência.

Portanto, o *score* PSSM do nucleotídeo A na posição 1 é,

$$\begin{aligned}
&\text{PSSM}_{A1} = \log_2\left(P_A \times \tfrac{P_{AA}}{q_A q_A} + P_C \times \tfrac{P_{CA}}{q_C q_C} + P_G \times \tfrac{P_{GA}}{q_G q_G} + P_T \times \tfrac{P_{TA}}{q_T q_A}\right) \\
&= \log_2\left(1 \times \tfrac{0.13}{0.25^2} + 0 \times \tfrac{0.03}{0.25^2} + 0 \times \tfrac{0.03}{0.25^2} + 0 \times \tfrac{0.06}{0.25^2}\right) \\
&= \log_2(2.08) = 1.056
\end{aligned} \tag{3.3}$$

Por fim, os valores obtidos após a aplicação da Equação 3.3 na matriz de probabilidades PFM são arredondados para o valor inteiro mais próximo. A Tabela 3.5

apresenta a matriz PSSM após todos os procedimentos descritos anteriormente.

Nucleotídeos	Posições								
Conhecidos	1	2	3	4	5	6	7	8	9
A	1	-1	0	-1	0	0	0	1	0
C	-1	0	-1	1	0	0	0	-1	0
G	-1	1	-1	1	0	0	0	-1	0
T	0	-1	1	-1	0	0	0	0	0

Tabela 3.5: Matriz PSSM das sequências S_1 e S_2.

Os vetores para cada nucleotídeo presentes na matriz PSSM são utilizados como *embeddings*, por exemplo, a sequência S_1 pode ser representada pela matriz de *embeddings* da Tabela 3.6. Essas matrizes de *embeddings* foram utilizadas como dados de treino para CNN. A classificação de proteínas motoras utilizando *iMotor* atingiu acurácias entre 92% e 96.4% para diferente bases de dados. O modelo também mostrou ser melhor que métodos clássicos de aprendizagem de máquina, como SVM, kNN e *Random Forest*.

	Sequência S_1								
	A	G	T	G	T	C	G	A	C
Embedding	1	-1	0	-1	0	-1	-1	1	-1
	-1	1	-1	1	-1	0	1	-1	0
	0	-1	1	-1	1	-1	-1	0	-1
	-1	1	-1	1	-1	1	1	-1	1
	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
	1	-1	0	-1	0	-1	-1	1	-1
	0	0	0	0	0	0	0	0	0

Tabela 3.6: Matriz de *embeddings* PSSM da sequência S_1.

Wang et al. [2019] utilizam matrizes PSSM, CNN e o classificador FSRF *(Feature-selective Rotation Forest)* para identificar regiões de ocorrência de *interações proteína-proteína*[7] (PPIs). Para codificação de sequências proteicas em dados numéricos, os autores utilizaram o método PSSM, descrito anteriormente na apresentação do trabalho de Le et al. [2019b]. Posteriormente essas matrizes servem de entrada para uma rede convolutiva que irá extrair mapa de características, para que no fim, esses vetores alimentem um modelo FSRF que irá classificar a sequência como região de *interação proteína-proteína* ou não.

interações proteína-proteína[7]: Locais de sequências biológicas onde ocorre contato físico e intencional entre duas ou mais proteínas.

Diferente de outros trabalhos mencionados, este aprende características das sequências em duas etapas: primeiro utilizam o método PSSM e posteriormente uma CNN. O método foi avaliado em uma única base de dados e, considerando a baixa quantidade de amostras disponíveis, cinco validações cruzadas foram realizadas sobre a base de treino. O classificador FSRF alcançou 97.75% de acurácia média e apresentou um desvio padrão de 0.54% entre as acurácias obtidas para cada subconjunto de treino.

Nesta seção apresentamos alguns trabalhos que solucionam diferentes problemas de genética através da transformação de sequências biológicas em estruturas matriciais que revelam alguma forma de padrão. Essas abordagens são computacionalmente econômicas e geram estruturas mais fáceis de interpretar, quando comparadas com vetores de *embeddings* obtidos via PLN. Na próxima seção serão apresentados alguns trabalhos que fogem do escopo de transformação matricial e PLN para representação de sequências, mas que sugerem abordagens bastante eficientes para pequenas bases de dados.

3.3 Outras abordagens

Apesar de atualmente existir uma grande concentração de trabalhos que abordam representação de sequências biológicas utilizando métodos de PLN e codificações matriciais, outros trabalhos apresentam metodologias diferentes. Conque et al. [2016] propõem uma metodologia baseada em redes complexas para classificação de sequências biológicas. Os autores baseiam-se na possibilidade de aplicação de redes complexas, principalmente em genética, para representar sequências. As métricas possíveis para redes complexas, tais como coeficiente de agrupamento, centralidade e número de comunidades, são utilizadas como dados para classificação, pois descrevem características internas das sequências, como associações entre seus nucleotídeos. Além disto, o cálculo da entropia entre redes revela informações globais sobre o comportamento estrutural destas sequências.

O método pré-processa as sequências em duas etapas. Primeiramente converte as sequências em subestruturas através do algoritmo *k-Mer*, obtendo um conjunto de subsequências ocultas. Por exemplo, seja a sequência $S = \{\text{ATGGAGTCCGAA}\}$, S pode ser representada pelo conjunto P de subestruturas obtidas para $k = 2$, de modo que, $P = \{\text{AT, TG, GG, GA, AG, GT, TC, CC, CG, GA, AA}\}$ ou pelos conjuntos W_1 e W_2 de sequências ocultas de S, tal que, $W_1 = \{\text{AT, GG, AG, TC, CG, AA}\}$ e $W_2 = \{\text{TG, GA, GT, CC, GA}\}$. Na segunda etapa, os autores definem as arestas do grafo

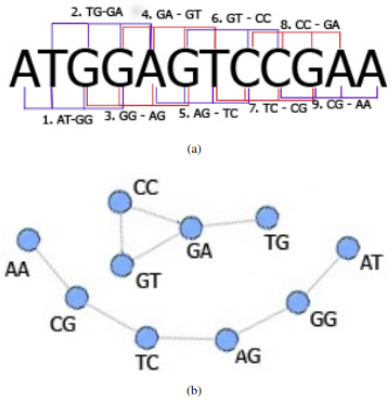


Figura 3.10: Etapas de pré-processamento de uma sequência pelo método proposto por Conque et al. [2016]. Em a) aplica-se o algoritmo *k-Mer* na sequência para obtenção de subestruturas. Um grafo das subestruturas obtidas em a) é gerado em b). Fonte: Conque et al. [2016].

a partir da relação das subestruturas dentro da sequência, levando em consideração uma janela de contexto de tamanho ST. Por exemplo, seja a palavra GG presente no conjunto W_1, se $ST = 1$, então GG se conecta com AT e AG, por outro lado, caso $ST = 2$, então GG se conecta com AT, AG e TC. A Figura 3.10 ilustra estas duas etapas.

Uma das vantagens deste método é sua capacidade de representar sequências através de uma estrutura não linear (grafo) semelhante à estrutura biológica tridimensional de sequências. No entanto, não é capaz de obter representações completas de sequências, pois não gera conexões individuais para cada item da sequência. Os autores avaliaram a metodologia utilizando cinco classificadores: *Random Forest*, SVM, kNN, Árvore de Decisão, *Naive Bayes*. *Random Forest* apresentou o melhor resultado na classificação de três classes de sequências genômicas, alcançando 0.912 de acurácia.

Chen et al. [2019] apresentam o método *CKN-seq*, que consiste em uma abordagem híbrida entre redes convolutivas e métodos de *kernel* para modelagem de sequências biológicas. Os métodos de *kernel* atuam mapeando uma amostra x_i para um espaço vetorial. Na etapa de pré-processamento do método são obtidas as subestruturas para cada sequência, o algoritmo *k-Mer* foi utilizado nesta tarefa. O *kernel* proposto pelos autores compara duas sequências x e x', de tamanhos m e m', através do produto escalar entre representações *one-hot* das subestruturas presentes em cada sequência, da seguinte maneira:

$$K(x, x') = \frac{1}{mm'} \sum_{i=1}^{m} \sum_{j=1}^{m'} K_0(P_i(x) P_j(x')), \tag{3.4}$$

tal que, a função $P_i(x)$ codifica o *k-Mer* da sequência x_i em um vetor *one-hot* e K_0 é um *kernel* com valores positivos utilizado para comparação dos *k-Mers*. A Equação que descreve K_0 representa o *kernel* homogêneo de produto escalar entre dois vetores z e z':

$$K_0(z, z') = \|z\|\|z'\|\kappa\left(\left\langle \frac{z}{\|z\|}, \frac{z'}{\|z'\|} \right\rangle\right), \tag{3.5}$$

onde, $\kappa : u \longrightarrow e^{\frac{1}{\sigma^2}(u-1)}$. A Equação 3.4, então, permite obter vetores de representação para cada sequência biológica. Posteriormente, tais vetores serão utilizados como dados de treino pela CNN. O modelo *CKN-seq* foi aplicado na tarefa de classificação de *motifs* e apresentou um valor AUC igual à 0.986, mostrando-se superior a outros métodos utilizados para mesma tarefa.

3.4 Síntese dos trabalhos

A revisão bibliográfica deste trabalho teve como objetivo identificar os avanços referentes ao processo representação e classificação de sequências biológicas. Conforme apresentado, os trabalhos da Seção 3.1 possuem uma etapa de pré-processamento onde subestruturas das sequências são modeladas por métodos de PLN, representando-as por vetores de *embedding* capazes de armazenar informações de relação e funcionamento interno dessas subestruturas. Grande parte dos trabalhos relacionados apresentados na Seção 3.1 empregam o modelo *word2vec - SkipGram* no processo de modelagem das subestruturas. Foi observado que os classificadores mais utilizados estão no grupo de algoritmos de aprendizagem de máquina clássica, no entanto, alguns autores propõem classificadores mais complexos capazes de identificar e empregar relações temporais no seu aprendizado, tais como, LSTM, RNN e CNN-1D.

Uma boa parcela dos trabalhos revisados na Seção 3.2 representam sequências por matrizes binárias (*one-hot*), por outro lado, outros convertem sequências biológicas em matrizes mais complexas utilizando metodologias adicionais. Por fim, estas matrizes são empregadas como dados de treino para redes de convolução. A abordagem de representação por matrizes tende a gerar representações mais simples, exigindo modelos menos complexos para seu tratamento. Além disso, dependendo da necessidade da pesquisa, a interpretação de resultados baseando-se em vetores de *embeddings* modelados por métodos de PLN pode ser inviável. Esse problema não ocorre com matrizes de representação, pois são estruturas mais simples e que evidenciam padrões de difícil acesso. No entanto, a complexidade dos modelos empregados para classificação

pode adicionar dificuldade de interpretação nos métodos baseados em representação matricial.

Os trabalhos apresentados na Seção 3.3 apresentam metodologias interessantes para o representação e classificação de sequências biológicas. Conque et al. [2016] transforma as sequências biológicas em grafos e utiliza métricas de avaliação de grafos e entropia para extrair informações que possam ser utilizadas para classificação de sequências. Já Chen et al. [2019] apresenta uma metodologia híbrida entre CNN e métodos de *kernel* para modelagem de sequências, o que permite que o método trabalhe com sequências de tamanhos heterogêneos. As métricas mais utilizadas nos trabalhos apresentados nas Seções 3.1, 3.2 e 3.3 foram acurácia, revocação e pontuação F1. Um resumo dos trabalhos apresentados pode ser visualizado na tabela 3.7.

3.5 Considerações finais

Diante dos trabalhos expostos nas Seções 3.1, 3.2 e 3.3, é possível concluir que, apesar dos avanços em tarefas de classificação de sequências biológicas, nenhum dos trabalhos mencionados aqui preocupam-se, de fato, com a interpretação dos classificadores e seu conhecimento biológico de funções capazes de distinguir classes de sequências. Outro ponto que merece destaque é que poucos trabalhos exploram a possibilidade de abordar a relação temporal entre palavras biológicas de uma sequência, a utilização de modelos temporais pode encontrar padrões que não sejam possíveis de obter com modelos estáticos.

Os trabalhos da Seção 3.1 utilizam representações binárias para subestruturas de sequências como entrada de dados para os modelos de representação *word2vec* e *fastText*. Estes modelos não capturam estatísticas globais de um conjunto de sequências e, além disso, geram estruturas de dados com interpretações complexas, fazendo com que essas representações se tornem um empecilho, dependendo da necessidade da pesquisa. Os trabalhos apresentados na Seção3.2 comumente utilizam matrizes binárias para representar as sequências e as utilizam como dados para classificadores baseados em CNN. Apesar disso, um ponto positivo é que essas estruturas tendem a aumentar a visibilidade de padrões ocultos, agregando a capacidade de interpretação de resultados de maneira geral.

Nesse cenário, surgem duas oportunidades de reprodução de metodologias para classificação de sequências da dengue, que futuramente podem ser aplicadas a sequências em geral. A primeira consiste na utilização de estruturas matriciais que descrevem padrões extremamente importantes em genética, como por exemplo, matrizes de co-

ocorrências. Ainda na etapa de representação, podemos abordar técnicas para redução do espaço de representação, sem perder as características das sequências brutas. Isso permite que modelos menos complexos sejam usados para classificação, facilitando ainda mais a tarefa de interpretação.

A Tabela 3.7 contém informações resumidas sobre as metodologias e problemas abordados pelos trabalhos apresentados neste capítulo.

Autor	Problema		Metodologia				Resultados	
	Tarefa	Sequência	Pré-processamento	PLN	Matriz	Classificador	Métricas	%
Asgari [2015]	Classificação de famílias proteicas	Proteína	k-mer	word2vec	-	SVM	Acurácia	93.00
Pan and Shen [2018a]	Classificação de proteínas de ligação	Proteína	k-mer	word2vec	-	CNN 1D	AUC	91.30
Asgari et al. [2019]	Descoberta de motifs e classificação de proteínas de ligação	Proteína	PPE	word2vec	-	SVM	Precisão	[73,100]
Hamid and Friedberg [2019]	Classificação de peptídeos antimicrobianos	Proteína	k-mer	word2vec	-	RNN	Acurácia Revocação F1	95.80 94.60 94.70
Zhang and Kabuka [2020]	Classificação de famílias proteicas	Proteína	k-mer	word2vec	-	CNN 1D	F1	97.77
Le et al. [2019a]	Classificação de acentuossomos	DNA	k-mer	fastText	-	SVM	Acurácia	82.30
Ho et al. [2019]	Classificação de proteínas transportadoras	Proteína	k-mer	fastText	-	Random Forest / SVM	Revocação Especificidade	>99.00 >99.00
Villmann et al. [2019]	Classificação de sequências sintéticas e reais	RNA	NCD	BoW	-	GLVQ	Acurácia	76.10
Zeng et al. [2016]	Exploração de CNN 2D na classificação de proteínas de ligação	Proteína	k-mer	-	Matriz binária	CNN 2D	AUC	Diversos
Pan and Shen [2018b]	Classificação de proteínas de ligação	RNA	k-mer	-	Matriz binária	CNN 2D	AUC	93.70
Sharma et al. [2019]	Classificação de classes de canceres	Expressão Gênica	-	-	Matriz de números reais	CNN 2D	Acurácia	99.00
Le et al. [2019b]	Classificação de funções moleculares de proteínas motoras	Proteínas	PSSM	-	Matriz de embeddings PSSM	CNN 1D	Acurácia	96.40
Wang et al. [2019]	Classificação de interações proteína-proteína	Proteína	PSSM + CNN 1D	-	Matriz de embeddings CNN 1D	FSRF	Acurácia	97.75
Conque et al. [2016]	Classificação de três classes de sequências genéticas	RNA	Redes complexas	-	-	Random Forest	Acurácia	91.20
Chen et al. [2019]	Classificação de motifs	RNA	kernel de string	-	Matriz binária	CNN 1D	AUC	98.60

Tabela 3.7: Resumo dos trabalhos apresentados. Trabalhos em **negrito** são candidatos a *baseline* desse trabalho

Capítulo 4

Abordagem proposta

Neste capítulo apresentamos um método para representação, classificação e interpretação sequências biológicas aplicados à amostras de proteínas da dengue. As principais contribuições metodológicas incluem: a introdução de uma estrutura capaz de capturar co-ocorrências de códons para representação de sequências; a classificação de severidade da dengue por proteína; a identificação de padrões específicos relacionados a dengue severa.

Nosso método é dividido em 5 etapas, sendo elas: i) alinhamento do RNA viral e segmentação por proteína para que estas sejam exploradas de forma independente; ii) normalização e tokenização de sequências como etapas para padronização e obtenção de códons das proteínas; iii) geração de matrizes de co-ocorrências de códons que servirão como dados de treino para o classificador; iv) predição do grau de infecção através dos algoritmos RF/CNN e; (v) interpretação local do modelo de classificação para as amostras de treinamento de forma a extrair conjuntos de co-ocorrências de códons significantes para predição de dengue severa.

4.1 Método

Propomos a representação de proteínas da dengue através de matrizes de co-ocorrência de códons. Para tal, os dados brutos de RNA da dengue, ou seja, amostras incompletas, não alinhadas e com possíveis erros de segmentação devem passar pelo processo de alinhamento e normalização, onde os nucleotídeos de cada amostra serão alinhados e padronizados conforme o código de aminoácidos (Seção 2.2). O alinhamento permite que cada amostra seja segmentada por proteínas. Após estas etapas, inicia-se o processo de tokenização onde cada amostra de segmento de proteína é convertido em um conjunto de subestruturas de três nucleotídeos (códons). Por fim, matrizes

de co-ocorrência são geradas a partir dos conjuntos de códons. Tais matrizes passam por processos de redução de dimensão e vetorização para redução de complexidade do problema. Por fim, as matrizes resultantes alimentam o classificador que por sua vez permite que o método *SHAP Values* gere interpretações. A Figura 4.1 apresenta o diagrama da metodologia proposta. A seguir, as etapas da metodologia serão explicadas com mais detalhes.

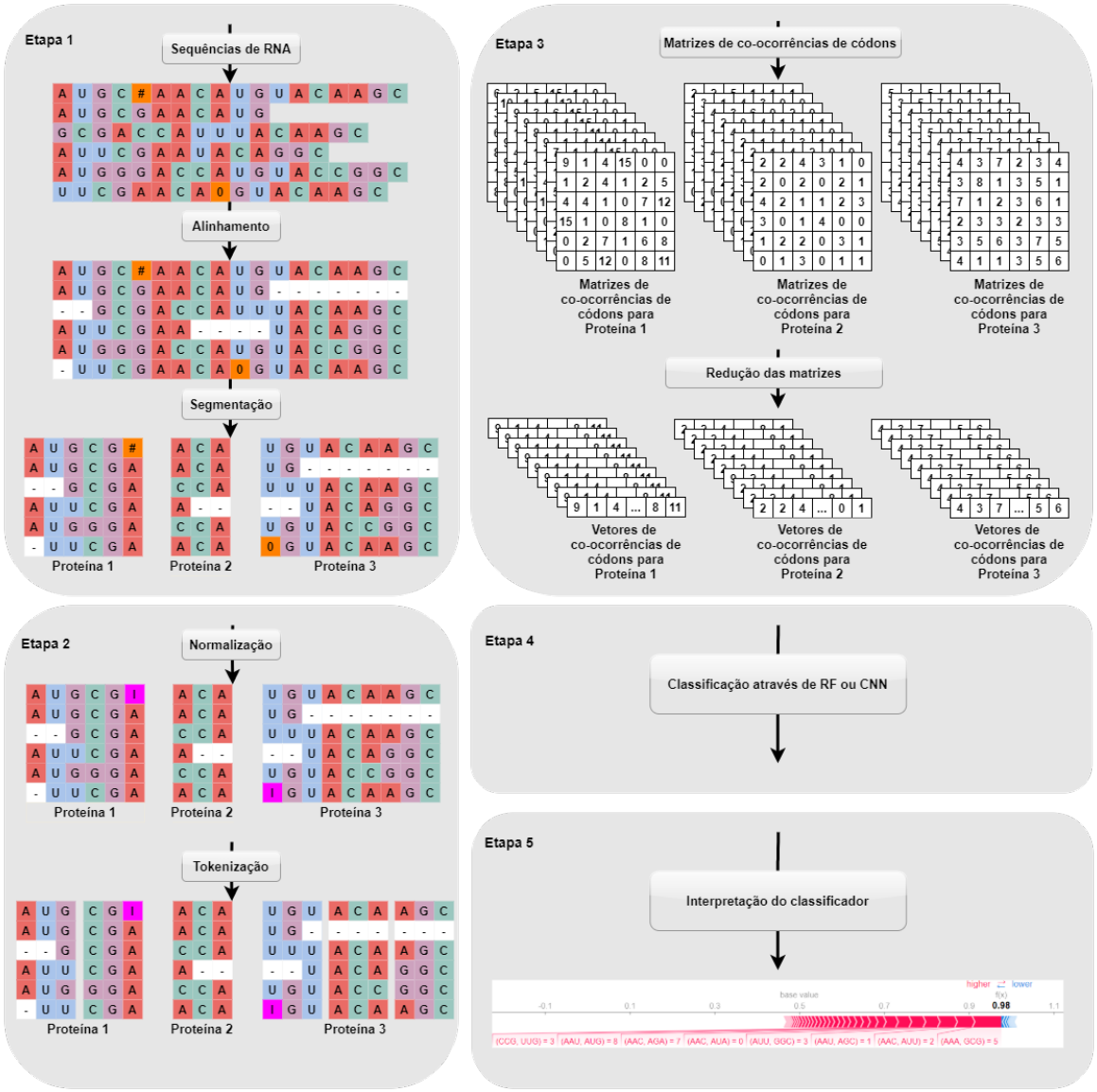


Figura 4.1: As 5 etapas da metodologia proposta. As etapas 1 e 2 realizam o pré-processamento das sequências de RNA. As representações das sequências são obtidas na terceira etapa. Na quarta etapa, cada representação é classificada de acordo com a severidade associada ao seu RNA. Por fim, padrões significantes para caracterização de dengue severa são extraídos do classificador na etapa 5.

4.1.1 Alinhamento e segmentação de sequências

As sequências foram alinhadas através do algoritmo MUSCLE disponibilizado no software *UGENE* versão 36.0 [Okonechnikov et al., 2012]. MUSCLE é um algoritmo de alinhamento de múltiplas sequências que acontece em três estágios: 1) o estágio de rascunho progressivo utiliza a contagem de *k-Mer* para calcular a similaridade entre cada par de sequências, resultando em um alinhamento global dos pares e, a partir disso, gera-se uma matriz de distância triangular, que permite a criação de uma árvore filogenética; 2) no estágio progressivo aprimorado, a árvore criada no estágio anterior é melhorada e um novo alinhamento progressivo é feito. Realiza-se o cálculo de similaridade entre as sequências e uma nova árvore é construída, a árvore anterior e a nova são comparadas para o aprimoramento de um novo alinhamento progressivo com base na ordem de ramificação de cada subárvore que não foi alterada; 3) no estágio de refinamento, uma aresta é apagada da árvore, dividindo, assim, as sequências em subgrupos. O perfil de alinhamento de cada subgrupo é obtido a partir do alinhamento múltiplo atual. Os dois perfis obtidos são realinhados e caso o novo alinhamento seja melhor, ele é mantido [Edgar, 2004]. A segmentação das sequências em proteínas foi realizada com base nas sequências de referência disponíveis no GenBank NCBI para cada sorotipo do vírus da dengue.

4.1.2 Normalização

A etapa de normalização consiste na análise de nucleotídeos das sequências, padronizando nucleotídeos sem significado biológico definido ou prováveis erros de sequenciamento. Logo, na normalização, os nucleotídeos que não estejam definidos no código de nucleotídeos IUPAC são substituídos pelo caractere "I" que representa indeterminação.

4.1.3 Tokenização

As matrizes de co-ocorrências dependem de subestruturas de proteínas pré-definidas para que sejam geradas. Além disso, é de grande importância que essas subestruturas tenham significado biológico. Para nossos experimentos, códons serão essas subestruturas de sequência. Códons consistem em tripletes de nucleotídeos que podem ser transcritos para aminoácidos [Yanofsky, 2007]. Então, na etapa de tokenização são obtidos os códons de cada sequência proteica. Portanto, cada sequência é representada por um conjunto de códons ordenados de acordo com sua posição na sequência de RNA, conforme a Figura 4.1 (etapa 2).

4.1.4 Matriz de co-ocorrências de códons

Matrizes de co-ocorrências tem sido utilizadas para coleta de estatísticas de dados variados, especialmente dados de imagem e texto [Carr and De Miranda, 1998, Zhang et al., 2017, Brochier et al., 2019]. Na análise de imagens médicas, as matrizes de co-ocorrência são empregadas para mensurar texturas de imagens [Abdel-Nasser et al., 2019]. No campo de PLN, as co-ocorrências podem fornecer indícios de relações semânticas entre palavras em um corpo de texto [Pennington et al., 2014]. Suas aplicações também se expandem para o campo da bioinformática, por exemplo, evidências de relações funcionais importantes em sequências proteicas para os processos de proteínas podem ser encontradas quando padrões idênticos de co-ocorrências de aminoácidos estão presentes em diferentes regiões [Lee et al., 2013, 2014].

Uma co-ocorrência de códons é a ocorrência de dois códons em uma segmento de proteína. Seja P uma sequência de códons e S um segmento de P, a matriz de co-ocorrências de códons X pode ser obtida pela fórmula: $X_{ij} = \sum_S K_{ij}$, onde,

$$K_{ij} = \begin{cases} 1, & \text{if } i, j \in S \\ 0, & \text{caso contrário} \end{cases} \tag{4.1}$$

e X_{ij} denota a quantidade de vezes que o códon j estava no mesmo segmento que o códon i. Desta forma, $X_{i,j}$ é proporcional à probabilidade conjunta $P(i, j)$, que representa a probabilidade de ocorrência dos códons i e j em um mesmo segmento.

O segmento, ou janela de contexto, reflete no alcance das co-ocorrências dos códons, por exemplo, segmentos extensos refletem na cobertura de grandes áreas do genoma, gerando co-ocorrências entre códons distantes e refletindo na capacidade das matrizes de capturar correlações de longa distância. De maneira análoga, segmentos pequenos definem uma análise local e conseguem capturar regiões com padrões extremamente conservados.

4.1.4.1 matriz de co-ocorrências molde

Para que as matrizes de co-ocorrências de cada amostra de mesma proteína tivessem dimensões idênticas, fez-se necessário a criação de um dicionário global contendo todos os códons presentes nas amostras. Com posse do dicionário global foi possível gerar uma matriz de co-ocorrências molde que integra todas co-ocorrências de códons presentes na proteína. Por exemplo, sejam as amostras $A_1 = \{CAU,\ ICG,\ GGC\}$, $A_2 = \{CAU,\ GCG\ UGU\}$ e $A_3 = \{-AU,\ GCG,\ AIC\}$ é possível obter o dicionário global de códons $d = \{CAU,\ ICG,\ GGC,\ GCG,\ UGU,\ -AU,\ AIC\}$ que nos permite ge-

rar a matriz de co-ocorrências molde presente na Figura 4.2. O fato das co-ocorrências serem intercambiáveis gera uma matriz de co-ocorrências simétrica.

	CAU	ICG	GGC	GCG	UGU	-AU	AIC
CAU	CAU CAU	CAU ICG	CAU GGC	CAU GCG	CAU UGU	CAU -AU	CAU AIC
ICG	CAU ICG	ICG ICG	ICG GGC	ICG GCG	ICG UGU	ICG -AU	ICG AIC
GGC	CAU GGC	ICG GGC	GGC GGC	GGC GCG	GGC UGU	GGC -AU	GGC AIC
GCG	CAU GCG	ICG GCG	GGC GCG	GCG GCG	GCG UGU	GCG -AU	GCG AIC
UGU	CAU UGU	ICG UGU	GGC UGU	GCG UGU	UGU UGU	UGU -AU	UGU AIC
-AU	CAU -AU	ICG -AU	GGC -AU	GCG -AU	UGU -AU	-AU -AU	-AU AIC
AIC	CAU AIC	ICG AIC	GGC AIC	GCG AIC	UGU AIC	-AU AIC	AIC AIC

Figura 4.2: Matriz de co-ocorrência de códons para as amostras A_1, A_2 e A_3

Algoritmos de aprendizagem de máquina dependem de entradas numéricas para realizar tarefas de predição. Embasados nesta necessidade, propomos a codificação de sequências proteicas em matrizes de co-ocorrência de códons. A capacidade das matrizes de co-ocorrência de concentrar co-ocorrências de códons em um único valor numérico a torna ideal para revelar padrões que ocorrem de forma imperceptível. Além disso, a possibilidade de escolher o tamanho dos segmentos permite controlar a coleta de relações entre códons, onde segmentos mais extensos tendem a capturar relações de longo alcance, enquanto que segmentos pequenos capturam relações locais.

O fato da matriz de co-ocorrências molde ser constituída por todas combinações de pares de códons do dicionário pode introduzir esparsidade de dados, levando em consideração que as dimensões da matriz molde podem crescer devido à combinações extremamente raras. Diante disso, o método *Quick Hull* (Seção 2.4) é empregado para calcular a região mais informativa em todas as matrizes de uma proteína. A Figura 4.3 ilustra o resultado esperado do algoritmo *Quick Hull*.

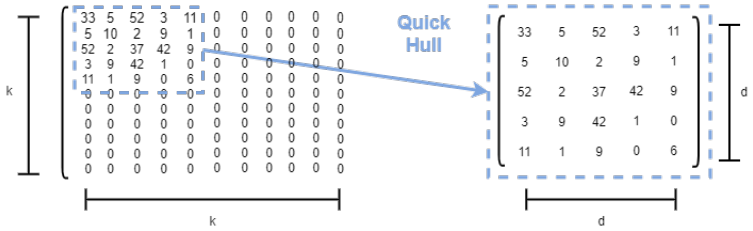


Figura 4.3: Após calcular a região mais informativa das matrizes, representada pelo retângulo tracejado, os pontos que pertencem a essa região são extraídos e estruturados em uma matriz menos esparsa.

Essa região densa de pontos pode ser considerada a matriz de co-ocorrência com exclusão de co-ocorrências raras. Diante disso, a região informativa gerada pelo algo-

ritmo *Quick Hull* é, na verdade, a região da matriz de co-ocorrência que armazena as co-ocorrências mais frequentes.

4.1.4.2 Redimensionamento das matrizes de co-ocorrências

As matrizes obtidas após aplicação do método *Quick Hull* são, de fato, sub-matrizes de co-ocorrência e possuem as mesmas características das matrizes originais, ou seja, são simétricas e intercambiáveis. Baseando-se nisso, a primeira etapa de redimensionamento é extrair apenas os elementos da matriz triangular superior. Conforme o exemplo a seguir, para matriz de co-ocorrências C serão extraídos apenas os índices em vermelho.

$$C = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & \dots & c_{1d} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & \dots & c_{2d} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} & \dots & c_{3d} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{45} & \dots & c_{4d} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & \dots & c_{5d} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ c_{d1} & c_{d2} & c_{d3} & c_{d4} & c_{d5} & \dots & c_{dd} \end{bmatrix}$$

As matrizes de co-ocorrências geradas possuem dimensões $\mathbb{R}^{d \times d}$, onde d é o tamanho de linhas ou colunas da matriz molde após a aplicação do método *Quick Hull*. O fato das matrizes serem simétricas e intercambiáveis permite o redimensionamento da matriz triangular superior em um vetor de dimensão $\mathbb{R}^{1 \times d(d+1)/2}$.

Para o classificador RF estruturamos uma base tabular com os vetores resultantes, onde cada coluna da base representa uma co-ocorrência entre pares de códons. Por outro lado, diante das características de dados exigidas nas operações de convolução, quando o classificador CNN foi utilizado a matriz triangular superior foi redimensionada para uma nova matriz com dimensão $\mathbb{R}^{d \times (d+1)/2}$.

Portanto, o classificador RF recebe dados com o formato $\mathbb{R}^{1 \times d(d+1)/2}$, enquanto que o classificador CNN recebe matrizes no formato $\mathbb{R}^{d \times (d+1)/2}$.

4.1.5 Seleção de variáveis

Com o objetivo de alcançar o desempenho máximo do classificador RF através da redução de complexidade do problema, eliminamos co-ocorrências que carregam pouca ou nenhuma informação. Para isso, utilizamos informação mútua (*Mutual Information* – MI) que mede a dependência entre duas variáveis através do cálculo de entropia de Shannon utilizando os k vizinhos mais próximos. Sejam as variáveis x e y, MI quantifica

a redução ded incerteza sobre x ao conhecer y [Cover, 1999]. Neste contexto, duas variáveis podem ser consideradas independentes se, e somente se, o coeficiente MI entre elas é zero. Em contrapartida, quanto maior a dependência entre duas variáveis, maior será seu valor de informação mútua [Kozachenko and Leonenko, 1987, Kraskov et al., 2004]. Portanto, foram calculados os valores de informação mútua entre co-ocorrências e rótulos (severidade) para cada base de proteínas. Diante disso, foram selecionadas para cada base as 50 co-ocorrências que apresentaram as maiores informações mútuas relacionadas aos rótulos.

A seleção de variáveis foi empregada apenas nos vetores $\mathbb{R}^{1\times d(d+1)/2}$ passados para o classificador RF. Essa etapa de pré-processamento não se fez necessária no classificador CNN diante da sua capacidade de extração de características em suas primeiras camadas [LeCun et al., 2015].

4.1.6 Classificadores

A quantidade amostras disponíveis para esta pesquisa insere dificuldade na tarefa de classificação dos dados por modelos profundos pois, estes não são eficientes no aprendizado de padrões em pequenas quantidades de amostras. Diante disso, optamos pela utilização de classificadores não paramétricos por reduzirem significativamente a quantidade de amostras exigidas para realizarem mapeamentos. O classificador escolhido foi *Random Forest*.

Levando em consideração a capacidade de modelagem de dados tabulares das CNN, avaliamos seu desempenho nas representações geradas. Além disso, suas operações de convolução e *max pooling* reduzem significativamente a quantidade de parâmetros dessas redes, quando comparadas com outras redes profundas [LeCun et al., 2015]. Diante disso, os dois classificadores empregados em nossa proposta foram *Random Forest* e CNN.

O classificador RF (Seção 2.6) é composto por 20 árvores de decisão, onde, o critério para definição de nós das árvores é baseado na entropia de Shannon das variáveis. A profundidade máxima de cada árvore é 20, tal que, o número mínimo de amostras para estar em um nó folha e o número mínimo de amostras para dividir um nó interno são 4 e 2, respectivamente. Diante da quantidade de amostras e da necessidade de controle de variância no modelo, não empregou-se *bootstrap* no treino. Para comparação da capacidade de caracterização de severidade de cada proteína, a mesma arquitetura de RF foi utilizada em todas as bases.

Diferente da RF, o classificador CNN (Seção 2.5) é ligeiramente mais complexo. É formado por dois blocos de convolução que consistem em: uma camada de convo-

lução, uma camada de normalização, uma camada de ativação com função tangente hiperbólica, uma camada de *dropout* e uma camada de *pooling* de valor máximo. No primeiro bloco, apenas dois filtros são utilizados, enquanto que no segundo esse valor é duplicado. As matrizes de *kernel* e *pooling* possuem as mesmas dimensões nos dois blocos, sendo 3×3 para o *kernel* e 2×2 para *pooling*. A taxa de *dropout* para todas camadas foi de 10%. Nas camadas de convolução empregou-se o regularizador L_2 com taxa de penalização igual a 1×10^{-4}. Portanto, esse classificador consiste em uma arquitetura rasa de CNN que é pouco complexa e possui, em média, 1.500 parâmetros para cada proteína.

4.2 Interpretação

Modelos de aprendizagem de máquina realizam internamente múltiplas operações matemáticas para obtenção de resultados. Por exemplo, pra realizar predições, classificadores geram valores reais que por sua vez serão associados aos rótulos. Como descrito anteriormente, SHAP Values realiza interpretações de variáveis a partir da função de esperança condicional apresentada na Equação 2.7 da Seção 2.7. A partir disso, o método atribui impactos positivos e negativos para as variáveis da instância de entrada de modo que o valor esperado do interpretador $E(f(z)|z_S)$ seja igual ao valor de saída do modelo original f. Desse modo, a grandeza do impacto reflete na influência da variável na classificação da amostra, tal que, impactos positivos aumentam a probabilidade de classificação correta da amostra, enquanto que impactos negativos tem o efeito oposto, sugerindo que variáveis com impactos positivos tem maior capacidade de caracterizar a classe da amostra [Lundberg et al., 2018a]. Essas informações podem ser facilmente visualizadas nos gráficos de força gerados pelo método, apresentado na Figura 4.4.

Figura 4.4: Exemplo de gráfico de força para uma amostra de dengue severa. As co-ocorrências em vermelho elevam $f(x)$ (probabilidade de dengue severa), enquanto que as co-ocorrências em azul reduzem $f(x)$

Nossas interpretações são baseadas em gráficos de forças obtidos para cada instância de treino e teste. Os impactos das co-ocorrências podem ser obtidos através da extração de dados dos gráficos de força. Para cada amostra de dengue severa serão selecionadas as z co-ocorrências com maiores impactos positivos para classificação.

Posteriormente essas z co-ocorrências serão agrupadas em uma matriz $M \in \mathbb{R}^{n \times z}$ de impactos, portanto, $M_i = \text{RANK}(f(x_i), z)$, tal que, $f(x_i)$ representa as interpretações da amostra x_i e $i = 1, 2, ..., n$, onde n é o número de amostras. Em seguida, o impacto médio de cada co-ocorrência em M será calculado por uma função g, selecionando-se as z co-ocorrências com maior impacto médio, logo, $h = \text{RANK}(g(M), z)$, onde h representa o conjunto ranqueado de co-ocorrências que serão utilizadas no restante das interpretações. Sendo assim, nossas interpretações são agrupamentos de dados dos gráficos de força.

É possível comparar a distribuição das co-ocorrências significantes para dengue severa com sua distribuição em amostras de dengue clássica, com o intuito de observar dissimilaridades entre cada distribuição. Diante disso, utilizamos gráficos de violino para apresentar graficamente as interpretações finais. Os gráficos de violino são ótimos para visualizar distribuição de dados e suas probabilidades, sendo estes uma combinação de *box plots* e gráficos de densidade. A forma dos gráficos de violino permitem observar o comportamento modal da distribuição de dados e valores com maior probabilidade de ocorrência. Por fim, nosso método exibirá as comparações entre as distribuições de valores de co-ocorrência presentes na matriz M (dengue severa) com a distribuição das mesmas co-ocorrências em dengue clássica. Um exemplo de gráfico de interpretação pode ser observado na Figura 4.5.

4.3 Considerações finais

Neste capítulo foi apresentada a proposta de solução capaz de atingir os objetivos deste trabalho. A metodologia baseia-se na representação de sequências de proteína da dengue em matrizes de co-ocorrências de códons. Essas matrizes consideram relações internas de códons dentro de cada sequência e são capazes de extrair e apresentar padrões de difícil acesso que possam estar associados com dengue severa. Também foram apresentadas as arquiteturas dos classificadores RF e CNN e, diante da quantidade de amostras disponíveis para treino e teste, a complexidade de ambos devem ser baixas, permitindo que mapeamentos entre matrizes e rótulos sejam realizados com mais facilidade. Por fim, exibimos o processo de obtenção e estruturação gráfica dos dados de interpretação. Os resultados parciais obtidos serão apresentados no próximo capítulo.

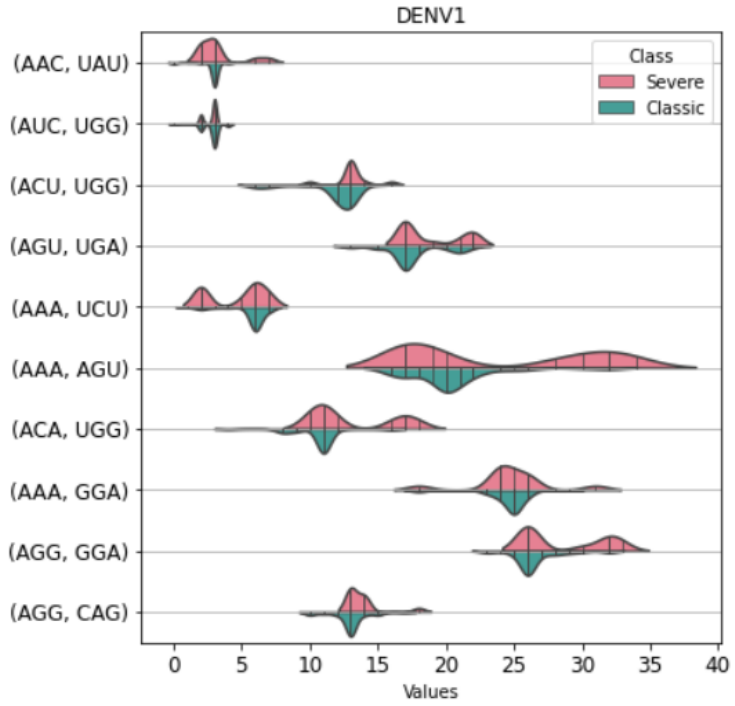


Figura 4.5: Exemplo de gráfico de violino para amostras de dengue sorotipo 1 (DENV1). Em rosa, distribuição de co-ocorrências significantes para dengue severa. Em verde, distribuição das mesmas co-ocorrências em amostras de dengue clássica.

Capítulo 5

Resultados

Os algoritmos da abordagem propostas neste trabalho foram implementados em *Python* e utilizam as bibliotecas *Tensorflow*, *Pandas*, *NumPy*, *SciPy* e *SkLearn*. Os classificadores foram treinados com bases de proteínas segmentadas a partir de 562 amostras de RNA da dengue, onde, algumas dessas podem apresentar somente algumas proteínas (RNA parcial) enquanto outras podem conter todas proteínas (RNA completo). Neste capítulo serão apresentados todos os procedimentos experimentais realizados e os resultados parciais obtidos. Esse capítulo está organizado da seguinte maneira: i) informações sobre a base de dados; ii) resultados de classificação iii) teste estatístico para avaliar o desempenho das proteínas na caracterização de dengue severa; iv) interpretações do melhor classificador treinado com dados da proteína com maior capacidade de caracterização de dengue severa; v) observação do comportamento de co-ocorrências significantes para dengue severa em regiões funcionais da proteína com maior capacidade de caracterização.

5.1 Bases de dados

Apesar da grande quantidade de genomas da dengue disponíveis publicamente em repositórios de sequências genéticas de universidades e institutos de pesquisa, constatamos grande escassez de amostras rotuladas com quadro clínico do paciente infectado. Diante disso, realizamos buscas manuais nos repositórios do *NCBI* para coleta de dados rotulados, obtendo 45 amostras rotuladas. Adicionalmente, 517 amostras foram coletadas através do repositório *NCBI Virus Variation*, totalizando 562 amostras de genoma da dengue rotuladas como quadro clínico do hospedeiro. Desse total, apenas 61 amostras apresentam genoma completo da dengue com suas 10 proteínas.

Os rótulos encontrados foram: febre da dengue (DF), febre hemorrágica da den-

gue (DHF) e síndrome de choque da dengue (DSS). Diante da baixa quantidade de amostras de DHF e DSS e por se tratarem de casos graves de dengue, realizamos a rotulação binária da base, onde, DF tornou-se dengue clássica e DHF e DSS, dengue severa. Todas as amostras, com exceção de duas amostras coletadas no baço, foram coletadas através do material sanguíneo isolado de humanos infectados entre os anos de 1985 e 2017. Os dados são provenientes de 20 países: Brasil, Camboja, Chile, China, Colômbia, Cuba, Espanha, Filipinas, Gana, Índia, Indonésia, Japão, Malásia, México, Paraguai, Polinésia Francesa, Sri Lanka, Tailândia, Taiwan e Vietnã.

Como dito na Seção 4.1.1, as sequências foram segmentadas em proteínas através de alinhamento. O alinhamento de sequências permite a padronização das amostras dos dados brutos, preenchendo sequências incompletas com *gaps* ou lacunas, representadas pelo simbolo "–" para que estas se alinhem aos genomas de referência do seu sorotipo (Seção 4.1.1), permitindo a criação de uma base de dados para cada proteína (Figura 4.1). O processo de alinhamento de sequências baseia-se no cálculo de similaridade de regiões conservadas entre sequências. Diante disso, é natural que o alinhamento adicione *gaps* em sequências parcialmente incompletas para que as regiões conservadas de cada sequência fiquem alinhadas, elevando a similaridade entre elas [Altschul et al., 1990, Thompson et al., 1994, Edgar, 2004]. Este procedimento pode resultar em regiões extensas de *gaps* para sequências muito incompletas, fazendo com que proteínas inteiras sejam representadas unicamente por *gaps*. Para contornar esse problema, antes de qualquer processamento para geração de códons e matrizes de co-ocorrências, optamos por remover amostras nas bases de proteínas formadas por mais de 15% de *gaps*. Os alinhamentos permitiram a geração de 10 bases de dados, uma para cada proteína. Após a eliminações de amostras, a distribuição final das bases pode ser observada na Tabela 5.1.

	Amostras		
Proteína	Dengue clássica	Dengue severa	Total
C	206	92	298
M	199	89	288
E	279	115	394
NS1	194	81	275
NS2A	194	76	270
NS2B	194	76	270
NS3	194	76	270
NS4A	194	76	270
NS4B	194	76	270
NS5	194	76	270

Tabela 5.1: Distribuição das bases de dados.

Os resultados de interpretação apresentados neste capítulo foram gerados por 115 amostras de dengue severa, sendo 11 amostras de dengue sorotipo 1 (DENV1), 73 amostras de dengue sorotipo 2 (DENV2), 28 amostras de dengue sorotipo 3 (DENV3) e 3 amostras de dengue sorotipo 4 (DENV4). A partir das amostras com informação de local disponíveis, constatou-se que as amostras de DENV1 foram coletadas, em grande parte, no Brasil e Polinésia Francesa, enquanto que amostras de DENV2 e DENV3 derivam, majoritariamente, do Brasil, México e Paraguai. Todas amostras de DENV4 foram coletadas no Camboja.

5.2 Resultados de classificação

Diante do evidente desbalanceamento das bases de dados apresentadas na Tabela 5.1, assim como a pequena quantidade de amostras em cada uma delas. Portanto, fez-se necessária a observação da capacidade de generalização dos classificadores para diferentes conjuntos de treino e teste, reduzindo a probabilidade de resultados gerados sob conjuntos de treino e teste ótimos. Os resultados apresentados nesta seção foram obtidos através de 5 experimentos compostos por 5 validações cruzadas cada. O algoritmo utilizado para validação cruzada foi *k-Fold* estratificado.

Considerando ainda o desbalanceamento das bases, optamos por empregar métricas de avaliação sensíveis a bases desbalanceadas, sendo elas: AUC (*Area Under The Curve*), precisão, revocação e pontuação F1 balanceadas. As métricas balanceadas compensam o desbalanceamento de classes através do cálculo de média ponderada entre instâncias corretamente classificadas. A média das métricas, assim como seus intervalos de confiança para todas proteínas podem ser observadas na Tabela 5.2. Os resultados foram obtidos por três abordagens: 1) Representações de proteínas por matrizes de co-ocorrência de códons e classificador RF; 2) Representações de proteínas por matrizes de co-ocorrência de códons e classificador CNN e; 3) Representações de proteínas por *embeddings* gerados pelo método *BioVec* e classificador RF. Os valores de cada célula são obtidos pela equação $\bar{x} \pm \epsilon$, onde $\bar{x}$ é a média de resultados e ϵ é erro da média obtido pela distribuição *t de Student* com 95% de nível de confiança.

Como podemos observar na Tabela 5.2 o classificador RF consegue obter melhores resultados de classificação a partir de matrizes de co-ocorrências de códons quando comparado com o classificador CNN e com a abordagem *BioVec* mais RF (*baseline*). Embora o modelo CNN sejam mais apropriado em tarefas onde a entrada de dados é multidimensional, modelos baseados em árvore podem ser mais precisos do que CNN onde os dados de treinamento possuem variáveis individualmente significativas e que

carecem de forte multi-escala temporal ou estruturas espaciais. Portanto, o restante dos resultados apresentados neste capítulo são baseados na abordagem matrizes de co-ocorrências mais RF.

Adicionalmente, realizamos análises exploratórias sobre os resultados obtidos nos 5 experimentos para observar graficamente o desempenho do RF em cada base de dados de matrizes de co-ocorrência. Para realizar a comparação visual entre os resultados obtidos para cada base empregamos *box-plots* (Figura 5.1) para verificar a distribuição empírica das métricas.

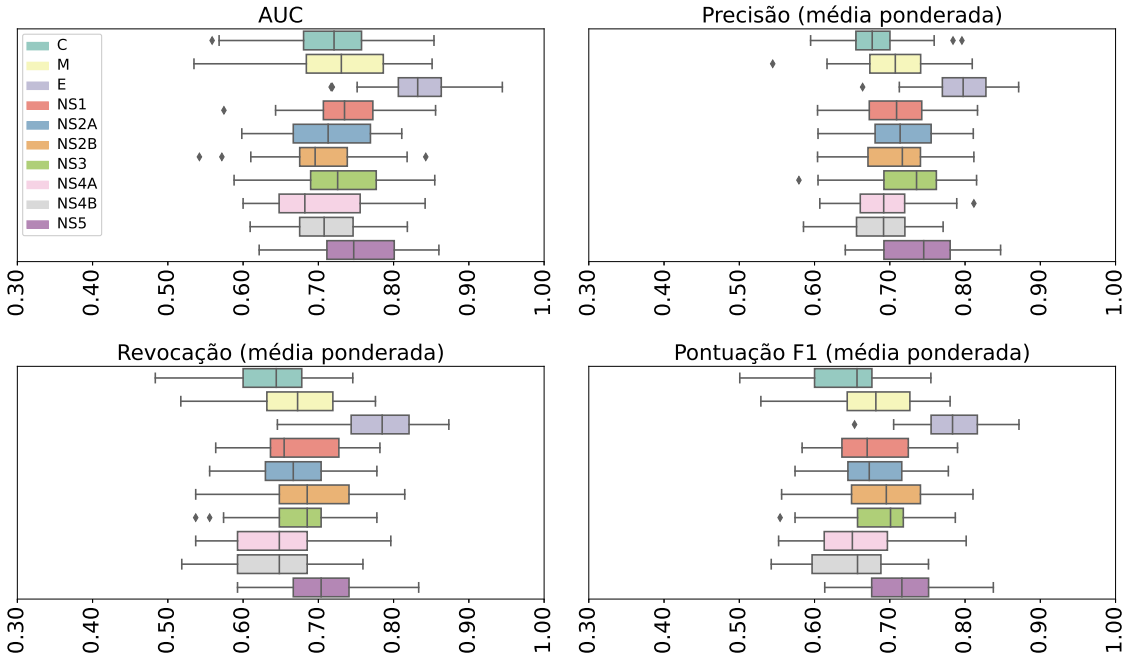


Figura 5.1: A mediana das métricas da proteína E são superiores as de outras proteínas, dando evidências de que seus resultados são superiores. Quando comparado com outras proteínas, os *box-plots* de proteína E indicam baixa dispersão dos resultados e simetria, sugerindo baixa variabilidade e que o classificador manteve um desempenho constante para cada conjunto de teste.

5.3 Testes estatísticos

Os box-plots da Figura 5.1 evidenciam uma possível diferença entre os resultados obtidos em cada proteína para abordagem de matrizes de co-ocorrências de códons com classificação via RF. Diante disso, para testar estatisticamente a hipótese de que as médias dos resultados são distintas para cada proteína, utilizamos o modelo de análise de variância unilateral (ANOVA) que compara médias amostrais através da distribuição F de Fisher-Snedecor [St et al., 1989, Girden, 1992]. Os dados empregados no

Proteína	Matrizes de co-ocorrências + RF				Matrizes de co-ocorrências + CNN				BioVec + RF			
	AUC	PRC	REV	F1	AUC	PRC	REV	F1	AUC	PRC	REV	F1
C	0.72 ± 0.03	0.68 ± 0.02	0.64 ± 0.03	0.64 ± 0.02	0.71 ± 0.03	0.67 ± 0.02	0.69 ± 0.02	0.66 ± 0.02	0.67 ± 0.04	0.66 ± 0.03	0.63 ± 0.03	0.64 ± 0.03
M	0.73 ± 0.03	0.70 ± 0.03	0.67 ± 0.03	0.68 ± 0.03	0.73 ± 0.03	0.69 ± 0.03	0.70 ± 0.02	0.67 ± 0.02	0.72 ± 0.03	0.68 ± 0.03	0.65 ± 0.03	0.66 ± 0.03
E	0.83 ± 0.02	0.79 ± 0.02	0.78 ± 0.02	0.78 ± 0.02	0.78 ± 0.02	0.72 ± 0.02	0.73 ± 0.02	0.71 ± 0.02	0.80 ± 0.02	0.74 ± 0.02	0.73 ± 0.02	0.73 ± 0.02
NS1	0.74 ± 0.03	0.71 ± 0.02	0.67 ± 0.02	0.68 ± 0.02	0.73 ± 0.03	0.69 ± 0.04	0.71 ± 0.02	0.66 ± 0.02	0.71 ± 0.03	0.70 ± 0.02	0.69 ± 0.03	0.69 ± 0.02
NS2A	0.72 ± 0.03	0.71 ± 0.02	0.67 ± 0.02	0.68 ± 0.02	0.70 ± 0.03	0.63 ± 0.04	0.70 ± 0.02	0.63 ± 0.02	0.69 ± 0.03	0.69 ± 0.02	0.67 ± 0.03	0.67 ± 0.02
NS2B	0.70 ± 0.03	0.71 ± 0.02	0.69 ± 0.03	0.70 ± 0.03	0.71 ± 0.04	0.66 ± 0.04	0.70 ± 0.02	0.64 ± 0.02	0.67 ± 0.03	0.67 ± 0.02	0.63 ± 0.03	0.64 ± 0.02
NS3	0.73 ± 0.03	0.72 ± 0.03	0.68 ± 0.03	0.69 ± 0.03	0.69 ± 0.03	0.59 ± 0.03	0.69 ± 0.02	0.62 ± 0.02	0.67 ± 0.03	0.68 ± 0.02	0.67 ± 0.02	0.67 ± 0.02
NS4A	0.70 ± 0.03	0.69 ± 0.02	0.64 ± 0.03	0.65 ± 0.02	0.71 ± 0.02	0.63 ± 0.03	0.69 ± 0.01	0.65 ± 0.02	0.67 ± 0.03	0.66 ± 0.02	0.65 ± 0.03	0.65 ± 0.02
NS4B	0.70 ± 0.02	0.69 ± 0.02	0.64 ± 0.03	0.65 ± 0.02	0.69 ± 0.03	0.62 ± 0.03	0.69 ± 0.01	0.63 ± 0.01	0.64 ± 0.03	0.66 ± 0.02	0.63 ± 0.03	0.64 ± 0.03
NS5	0.75 ± 0.03	0.74 ± 0.02	0.71 ± 0.02	0.72 ± 0.02	0.70 ± 0.03	0.62 ± 0.03	0.69 ± 0.02	0.64 ± 0.02	0.65 ± 0.03	0.65 ± 0.02	0.66 ± 0.02	0.65 ± 0.02

Tabela 5.2: Média dos resultados de classificação em 5 experimentos de 5 *folds* cada. Precisão (média ponderada), revocação (média ponderada) e pontuação F1 (média ponderada) são representadas, respectivamente, pelas siglas *PRC*, *REV* e *F1*.

teste ANOVA devem atender ao pressuposto de homogeneidade de variâncias, assim como os resíduos do modelo devem ser normalmente distribuídos. Para estas tarefas empregamos os testes Levene [Levene, 1961] e Shapiro-Wilk [Shapiro and Wilk, 1965], respectivamente. As hípoteses nulas (H_0) e alternativas (H_1) para os testes Levene, Shapiro-Wilk e ANOVA estão listadas a seguir,

Teste Levene

- H_0: as variâncias dos grupos são homogêneas;
- H_1: as variâncias dos grupos não são homogêneas.

Teste Shapiro-Wilk

- H_0: os dados são normalmente distribuídos;
- H_1: os dados não são normalmente distribuídos.

Teste ANOVA

- H_0: as médias da amostra são iguais;
- H_1: pelo menos uma das médias é diferente das outras.

onde, as hipóteses nulas são aceitas se, e somente se, o p-valor do teste for maior que um nível descritivo ϵ. Ressalta-se que o teste de normalidade de Shapiro-Wilk é aplicado nos resíduos do modelo ANOVA para verificação do pressuposto de resíduos normais. A Tabela 5.3 apresenta os resultados dos testes ANOVA para cada métrica, assim como os testes de seus pressupostos. A Tabela 5.3 apresenta os resultados dos testes ANOVA para cada métrica, assim como os testes de seus pressupostos.

Após obter os resultados do teste ANOVA, realizamos então o teste de Tukey para verificar a diferença entre as médias das métricas para cada proteína. A hipótese nula para o teste de Tukey assume que não existe diferença estatística significativa entre as médias de duas amostras, enquanto que a hipótese alternativa assume o oposto. Os pares de proteínas com médias de métricas estatisticamente distintas podem ser observados na Figura 5.2. Como podemos observar, para todas as métricas a proteína E apresenta resultados médios estatisticamente distintos quando comparados com outras proteínas, indicando sua capacidade superior de caracterizar dengue clássica e severa em nossos experimentos. Baseados nessas evidências estatísticas, as interpretações do classificador RF foram realizadas apenas para matrizes de co-ocorrência da proteína E.

	Levene p-valor	ANOVA p-valor	Shapiro-Wilk p-valor
AUC	0.869	9.309×10^{-11}	0.190
PRC	0.961	6.388×10^{-11}	0.390
REV	0.978	7.354×10^{-14}	0.194
F1	0.983	1.229×10^{-14}	0.393

Tabela 5.3: Para um nível descritivo $\epsilon = 0.05$ as hipótese nulas para os teste de Levene e Shapiro-Wilk são aceitas, portanto, apresentando indícios de que as métricas possuem variâncias homogêneas e que os resíduos do modelo ANOVA são normalmente distribuídos. Por fim, a hipótese nula do teste ANOVA é rejeitada, indicando que pelo menos uma das médias das métricas é diferente das demais. Precisão (média ponderada), revocação (média ponderada) e pontuação F1 (média ponderada) são representadas, respectivamente, pelas siglas *PRC*, *REV* e *F1*.

5.4 Interpretações de resultados para proteína E

Após treinados, os classificadores foram interpretados pelo do método *SHAP Values* através do algoritmo *TreeExplainer*. O *TreeExplainer* é um método específico para interpretações locais de modelos baseados em árvores, fornecendo interpretações locais rápidas e precisas através do calculo dos valores *SHAP* para cada folha de uma árvore. O método é formado por três algoritmos que estimam $f(h_x(z')) = E(f(z)|z_S)$ (Seção 2.7) seguindo recursivamente o caminho de decisão para uma instância de entrada x. A metodologia completa, assim como os algoritmos que definem o *TreeExplainer*, podem ser encontrados em Lundberg et al. [2020].

Como dito na Subseção 2.7 o método *SHAP Values* gera interpretações individuais para cada amostra de dados. Essas interpretações podem ser visualizadas através de diversos gráficos apresentados em Lundberg et al. [2018a]. Diante disso, empregamos gráficos de força para compreender os valores *SHAP* gerados pelo algoritmo *TreeExplainer*. Os gráficos de força apresentam o impacto das variáveis na predição, tal que, os impactos descrevem a capacidade de determinada variável elevar a probabilidade da classe da amostra. Para obter intuições globais sobre o modelo, os resultados apresentados nesta Seção são obtidos através de extrações sucessivas de dados de interpretação dos gráficos de força e compactados em gráficos de violino. Informações sobre o processo de extração e compactação desses dados, assim como a geração de gráficos de interpretação por esses dados, podem ser observadas na Seção 4.2.

Optamos por estratificar as interpretações por sorotipo, dessa forma, revelando características únicas para cada família sorológica de vírus da dengue. Em seguida, os dados das 10 co-ocorrências mais significantes serão apresentados em gráficos de interpretação. Por fim, os valores de co-ocorrências serão comparados com amostras

Figura 5.2: A comparação entre pares de proteínas indica que a média dos resultados da proteína E é estatisticamente diferente das médias das outras proteínas.

de mesmo sorotipo de dengue clássica, atenuando as diferenças estruturais entre amostras de dengue clássica e severa. As interpretações são constituídas por dois gráficos: 1) gráfico com interpretações que alavancam a probabilidade de dengue severa e; 2) gráfico com interpretações que reduzem a probabilidade de dengue severa. Os gráficos de interpretação com co-ocorrências que elevam a probabilidade de dengue severa nas amostras e os gráficos de interpretação com co-ocorrências que reduzem a probabilidade de dengue severa nas amostras podem ser encontrados nas Figuras 5.3 e 5.4, respectivamente.

As interpretações foram realizadas sobre um modelo RF (Seção 4.1.6) treinado com todas amostras, que incluem dengue clássica e severa da proteína E. Em seguida, apenas amostras de dengue severa foram interpretadas. As amostras de dengue severa interpretadas fazem parte do treinamento da árvore. A abordagem de interpretar amostras de treino é justificável pois de fato estamos interessados em obter os padrões de particionamento de dados que são gerados no momento da construção das árvores do modelo RF, como demostram Lundberg et al. [2018b,a, 2020], Molnar [2019]. A matriz

de confusão (Tabela 5.4) apresenta o comportamento de classificação do modelo RF sobre todo conjunto de dados da proteína E.

		Real	
		Dengue Severa	Dengue Clássica
Previsto	Dengue Severa	110	5
	Dengue Clássica	33	246

Tabela 5.4: Matriz de confusão do modelo RF para matrizes de co-ocorrência de códons da proteína E.

As interpretações da proteína E revelam características distintas entre as co-ocorrências de códons significantes para dengue severa quando comparadas com dengue clássica. Em geral, como podemos visualizar nas Figuras 5.3 e 5.4, as distribuições de co-ocorrências são na maioria das vezes distintas para dengue clássica e severa. Muitas dessas distribuições são multimodais e apresentam modas diferentes em amostras de dengue clássica e severa, como fica evidente na Figura 5.3 ao analisar as co-ocorrências (AGG, GGA) e (AAG, CAU) em DENV1, (CAU, UAC) e (AGG, GGA) em DENV2, (GGA, GGC) e (AAU, AGG) em DENV3 e (CAC, CUG) e (AAA, UCU) em DENV4. Esse comportamento distintos de distribuições também se mantém para os gráficos da Figura 5.4, onde as 10 co-ocorrências com maior impacto negativo na classificação de dengue severa são apresentadas.

É evidente que existam combinações diferentes de co-ocorrências que caracterizam dengue severa para cada sorotipo, porém, examinando os gráficos da Figura 5.3 podemos ver que em todos os sorotipos de dengue as co-ocorrências (GGA - Glicina, GGC - Glicina) e (CAC - Histidina, CUG - Leucina) aparecem como as duas primeiras colocadas no *ranking* de importância para classificação de dengue severa. A co-ocorrência (GGA, GGC) em DENV3 tem distribuição ampla, apresentando maior densidade em valores elevados de co-ocorrência, o que não ocorre nos outros sorotipos.

Nas Figuras 5.3 e 5.4 os gráficos de violino em algumas co-ocorrências de DENV4 foram comprimidos, isso se da ao fato deles não possuírem variância, o que pode ser justificado pela quantidade de amostras disponíveis desse sorotipo. Nesse ponto, podemos verificar que nos sorotipos DENV2 e DEN3, que possuem o maior número de amostras, a distribuição de co-ocorrências apresenta comportamento mais variante entre as duas classes. Outro ponto é que, em alguns casos raros, as co-ocorrências podem ser significantes para dengue severa e clássica, como é o caso da co-ocorrência (AUU, CGG) que possui impacto positivo e negativo em DENV1, isso significa que o classificador atribui bastante importância a essas co-ocorrências pois, dependendo da grandeza dos seus valores de co-ocorrência, podem caracterizar dengue clássica ou severa.

O mapeamento das sequências da proteína E em matrizes de co-ocorrências de códons revelaram padrões de co-ocorrências ligeiramente distintos para dengue clássica e severa, dando indícios de que essas co-ocorrências ocorrem de formas independentes entre amostras de dengue clássica e severa. Adicionalmente, os resultados também fortificam a hipótese de diferenças estruturais entre amostras de dengue clássica e severa, no entanto, outras formas de mapeamento devem ser empregadas nas sequências de proteína E da dengue para que essa afirmação possa ser realizada.

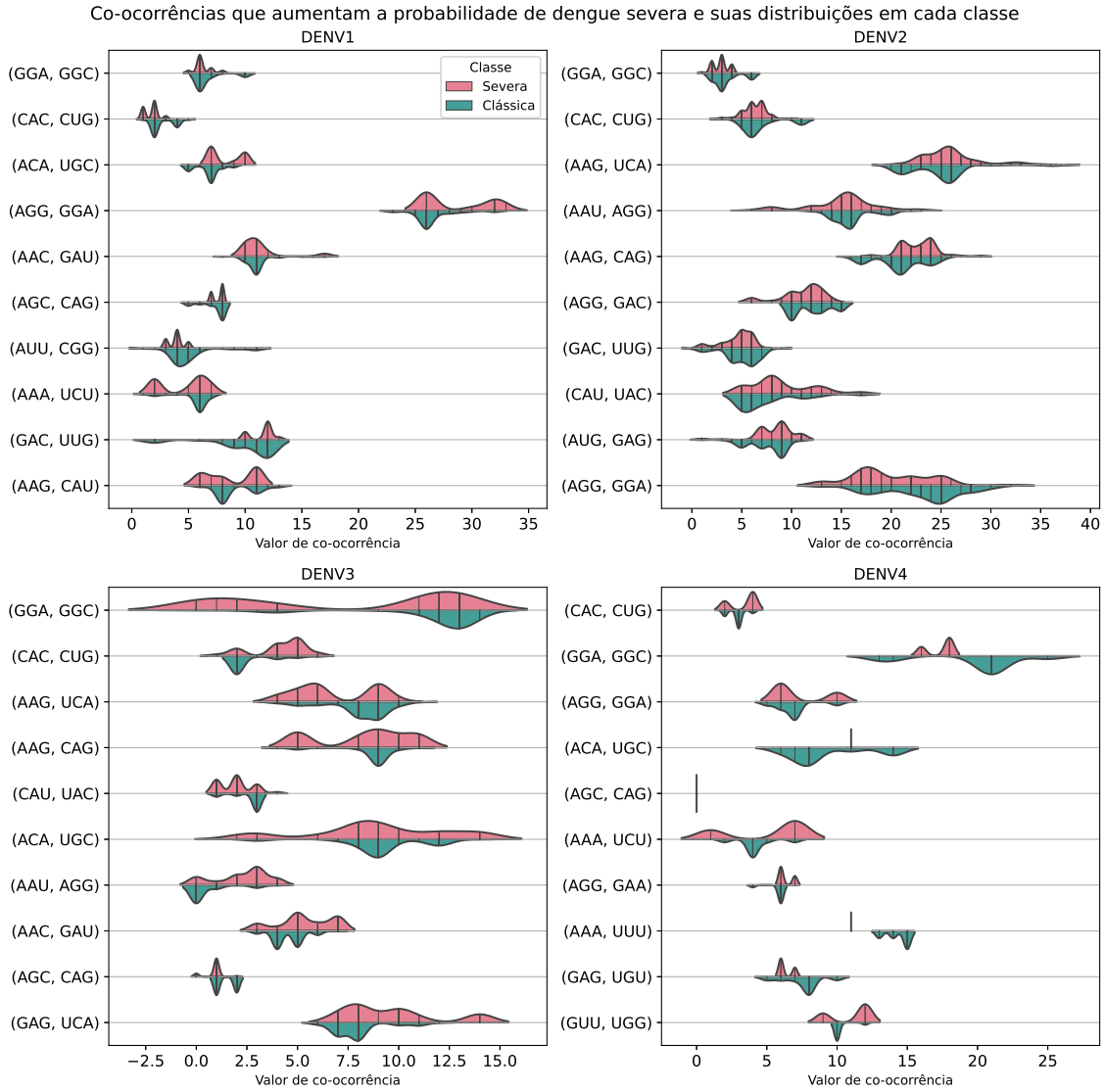


Figura 5.3: Gráficos de violino das 10 co-ocorrências com maior impacto positivo na classificação de dengue severa em cada sorotipo

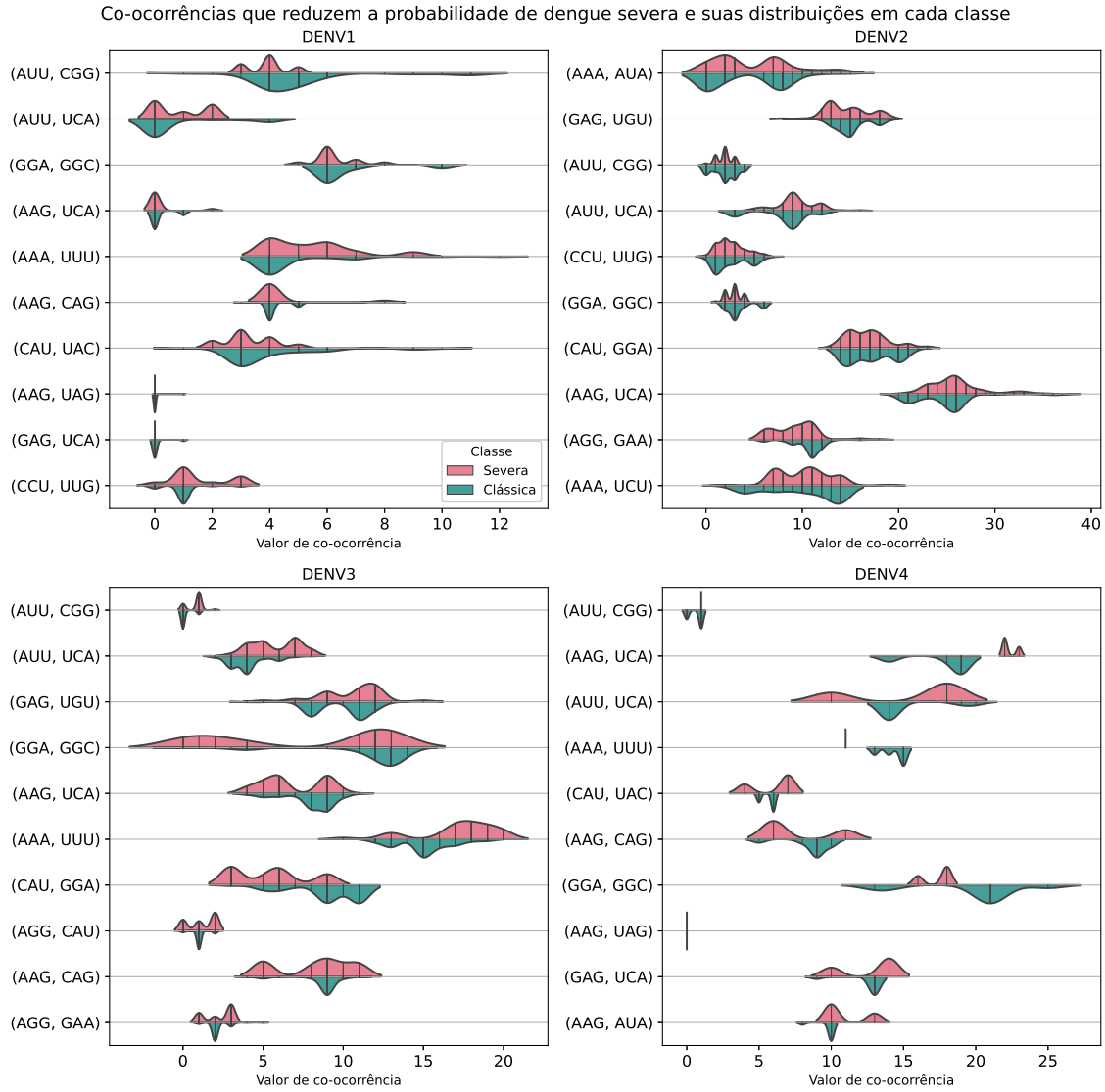


Figura 5.4: Gráficos de violino das 10 co-ocorrências com maior impacto negativo na classificação de dengue severa em cada sorotipo

5.5 Análise de proteína E por regiões

Nesta seção apresentaremos o comportamento das co-ocorrências de códons com impacto positivo na classificação de dengue severa dentro de quatro regiões fundamentais da proteína E, sendo elas: Domínio 1, Domínio 2, Transmembrana 1 e Transmembrana 2. A Tabela 5.5 apresenta o inicio e fim dessas regiões além do tamanho das sequências de proteína E para cada sorotipo. O intervalo das regiões da proteína E foram obtidos nos seguintes artigos: Laille and Roche [2004], Foster et al. [2004], Li et al. [2008], Ito et al. [2007], Midgley et al. [2012] e Patil et al. [2012].

	DENV1		DENV2		DENV3		DENV4	
	Início	Fim	Início	Fim	Início	Fim	Início	Fim
Domínio 1	1	296	1	269	2	294	4	296
Domínio 2	298	393	271	366	296	391	298	393
Transmembrana 1	444	468	412	436	442	466	443	467
Transmembrana 2	475	493	448	466	472	491	473	493
Tamanho da proteína E	495		492		492		495	

Tabela 5.5: Estrutura das regiões da proteína E para cada sorotipo da dengue.

As Tabelas 5.6, 5.7, 5.8 e 5.9 apresentam a média de co-ocorrências nas regiões da proteína E de todas amostras de mesmo sorotipo. Podemos observar que as co-ocorrências se concentram nas regiões Domínio 1 e 2, para todos sorotipos. Em especial, para todos sorotipos o Domínio 1 aparenta ser a região com maior co-ocorrências significantes para dengue severa.

	Domínio 1	Domínio 2	Transmembrana 1	Transmembrana 2
(GGA, GGC)	4.3	1.0	0.0	0.0
(CAC, CUG)	1.6	0.0	0.0	0.0
(ACA, UGC)	6.7	0.5	0.0	0.0
(AGG, GGA)	21.0	3.6	0.0	0.0
(AAC, GAU)	3.6	5.2	0.0	0.1
(AGC, CAG)	3.5	0.1	0.0	0.0
(AUU, CGG)	2.4	0.2	1.3	0.0
(AAA, UCU)	3.3	0.0	0.0	0.0
(GAC, UUG)	9.9	0.0	1.5	0.0
(AAG, CAU)	2.1	1.7	0.0	0.0

Tabela 5.6: Valores médios de co-ocorrências de códons em cada região da proteína E para amostras de dengue severa do sorotipo 1.

	Domínio 1	Domínio 2	Transmembrana 1	Transmembrana 2
(GGA, GGC)	2.8	0.1	0.0	0.0
(CAC, CUG)	2.8	0.0	0.0	0.2
(AAG, UCA)	9.5	8.0	0.1	0.0
(AAU, AGG)	11.0	0.8	0.0	0.0
(AAG, CAG)	6.2	8.6	0.2	0.0
(AGG, GAC)	10.8	0.0	0.0	0.0
(GAC, UUG)	1.8	1.2	0.0	0.0
(CAU, UAC)	5.9	2.9	0.0	0.0
(AUG, GAG)	2.2	2.3	0.2	0.0
(AGG, GGA)	18.9	0.0	0.0	0.0

Tabela 5.7: Valores médios de co-ocorrências de códons em cada região da proteína E para amostras de dengue severa do sorotipo 2.

A co-ocorrência (AGC, CAG) em DENV4 não apresenta co-ocorrências em nenhuma região, assim como não apresenta co-ocorrências em amostras de dengue clás-

	Domínio 1	Domínio 2	Transmembrana 1	Transmembrana 2
(GGA, GGC)	5.7	0.7	0.0	0.0
(CAC, CUG)	2.6	0.9	0.0	0.0
(AAG, UCA)	5.5	0.8	0.0	0.0
(AAG, CAG)	6.6	1.5	0.0	0.0
(CAU, UAC)	1.0	0.9	0.0	0.0
(ACA, UGC)	5.3	3.4	0.0	0.0
(AAU, AGG)	1.2	0.0	0.0	0.0
(AAC, GAU)	4.1	1.2	0.0	0.0
(AGC, CAG)	0.4	0.0	0.0	0.0
(GAG, UCA)	9.4	0.2	0.0	0.0

Tabela 5.8: Valores médios de co-ocorrências de códons em cada região da proteína E para amostras de dengue severa do sorotipo 3.

	Domínio 1	Domínio 2	Transmembrana 1	Transmembrana 2
(CAC, CUG)	1.7	0.0	0.0	0.0
(GGA, GGC)	11.7	0.7	0.7	0.7
(AGG, GGA)	6.0	1.0	0.0	0.0
(ACA, UGC)	9.0	1.0	0.0	1.0
(AGC, CAG)	0.0	0.0	0.0	0.0
(AAA, UCU)	4.3	0.0	0.0	0.0
(AGG, GAA)	6.3	0.0	0.0	0.0
(AAA, UUU)	7.7	1.0	0.0	0.0
(GAG, UGU)	3.7	2.0	0.0	0.0
(GUU, UGG)	6.3	0.7	0.0	0.0

Tabela 5.9: Valores médios de co-ocorrências de códons em cada região da proteína E para amostras de dengue severa do sorotipo 4.

sica, no entanto, ainda sim é significante para a classificação de dengue sorotipo 4. Isso pode ser justificado pelo padrão gerado pelo conjunto de valores de todas as co-ocorrências apresentadas na Tabela 5.9, que por sua vez, foi significante para o particionamento das árvores do classificador e na modelagem de dengue severa.

5.6 Considerações finais

Neste capítulo foram apresentados resultados de classificação severidade da dengue através de amostras de RNA viral e evidências de que o desenvolvimento de dengue severa pode estar, em partes, associado a estrutura genética do vírus. Constatou-se que, para nosso problema, o presente método apresentou representações de sequências superiores as obtidas pelo método *BioVec*, ficando evidente com os resultados de classificação apresentados. Com o objetivo de verificar a capacidade de caracterização de severidade da infecção por parte das proteínas da dengue, realizamos testes estatísticos sobre os resultados de classificação em cinco experimentos com cinco validações cruzadas para cada proteína.

Para centralizar nossas explorações em uma região específica do RNA da dengue, optamos por interpretar os resultados da proteína com maior capacidade de caracterização de dengue severa. Apresentamos os padrões de co-ocorrências de códons significantes para dengue severa e comparamos suas distribuições com distribuições dos mesmos padrões em amostras de dengue clássica. Além disso, também verificamos o comportamento dos padrões de códons associados a dengue severa em sub-regiões estruturais da proteína selecionada, levantando a hipótese que uma sub-região específica pode apresentar mutações associadas a severidade.

Nosso experimentos foram capazes de gerar os resultados esperados. Apresentamos resultados que indicam que o método proposto pode gerar representações adequadas de sequências biológicas e que podem ser empregadas em diversos classificadores. Por fim, relatamos três evidências de que a severidade da dengue pode estar associada à particularidades moleculares do vírus.

Capítulo 6

Conclusões

Apresentamos uma proposta para classificação de severidade da dengue baseado em amostras de genomas do vírus. Além disso, apresentamos resultados que sugerem maior capacidade de caracterização de dengue severa pela proteína E que, por sua vez, desempenha o papel de reconhecimento e entrada na célula a ser infectada pelo vírus da dengue [Kuhn et al., 2002]. A glicoproteína viral E também é o principal alvo antigênico da resposta de anticorpos humanos, portanto, após a infecção, a resposta dos anticorpos é direcionada a esta proteína [Flipse and Smit, 2015]. Também apresentamos resultados de interpretação que dão indícios de possíveis particularidades estruturais, mais especificamente padrões de ocorrência conjunta de códons em regiões próximas na proteína E que são capazes de levar o hospedeiro infectado a desenvolver dengue severa.

Adicionalmente, também observamos que os padrões significantes para dengue severa encontrados em nosso trabalho se concentram em grande parte na região de Domínio 1 da proteína E, revelando possíveis associações dessa região com dengue severa. Os domínios da da proteína E são regiões estruturais que compartilham graus variáveis de homologia entre os diferentes vírus da dengue [Slon Campos et al., 2017]. Além disso, os domínios são as regiões de maior densidade da proteína e que apresenta mais dobras em sua estrutura tridimensional [Wodak and Janin, 1981].

Apresentamos evidências de que a severidade da infecção por dengue pode estar associada ao material genético e que mutações na estrutura de códons da proteína E podem liderar o desenvolvimento de dengue severa, independentemente do sorotipo. Além disso, nossa metodologia pode contribuir para identificação prévia de infecções severas baseando-se apenas na amostra genética do vírus.

De modo geral, os resultados obtidos neste trabalho podem aumentar a compreensão do problema de desenvolvimento de dengue severa em alguns hospedeiros e

associar diretamente esse evento a características genômicas do vírus. Os resultados de interpretação obtidos fortificam os resultados apresentados por Pandey et al. [2000] e Pandey and Igarashi [2000] que, ao realizar experimentos em amostras de DENV2, sugere que a severidade da infecção pode estar associada à variações moleculares que ocorrem nos genomas.

Para nosso conhecimento, nosso trabalho é o primeiro a estruturar bases de dados de proteínas da dengue dos quatro sorotipos rotuladas com grau de severidade da infecção, classificar e analisar essas amostras para apresentar padrões de co-ocorrências de códons associados a dengue severa em cada sorotipo. Nossos resultados também demonstram a alta capacidade de matrizes de co-ocorrências representarem sequências biológicas, fazendo com que nosso método possa ser empregado em outros problemas relacionados a sequencias. A versatilidade das matrizes permite que elas sejam empregadas em diversos modelos que vão de aprendizagem de máquina clássica até aprendizagem profunda, tornando nosso método flexível quanto ao uso de outros classificadores. Portanto, conseguimos alcançar os objetivos propostos que eram, de modo geral: 1) desenvolver um método de representação de sequências e; 2) encontrar padrões capazes de caracterizar dengue severa.

6.1 Limitações

Uma das limitações deste trabalho é a coleta da base de dados. De fato, não foi possível obter quantidades massivas de amostras de RNA da dengue com os rótulos de interesse, apesar do esformço e tempo depositados a essa tarefa.

Também podemos mencionar como limitação a dimensão dos dados. Nossas amostras possuem dimensões elevadas para pequena quantidade de amostras, dificultando a tarefa de aprendizado de padrões por parte dos classificadores. Além disso, nossas estruturas de representação não codificam estrutura de correlação entre códons das proteínas, o que é possível com *embeddings* obtidos por métodos de NLP e mapas de características gerados por convolução.

Por fim, as interpretações podem variar de acordo com o classificador, visto que classificadores estruturalmente diferentes podem atribuir importância à variáveis distintas.

6.2 Trabalhos futuros

Fica como principal sugestão para futuros trabalhos, a utilização de métodos adicionais de representação e que sejam capazes de gerar resultados interpretáveis, tais como, matrizes PSSM e matrizes de recorrência, unidos a matrizes de co-ocorrência. Isso poderia gerar novas hipóteses baseadas em outros padrões estruturais. Ao aumentar a complexidade das representações, outros classificadores clássicos, porém mais robustos que as RF, poderiam ser empregados, tais como *XGBoost*, *Gradient Boosting Tree* e *(Feature-selective Rotation Forest).*

Diante da estrutura tridimensional que as sequências biológicas assumem, uma importante sugestão de trabalhos futuros seria quantificar a proximidade de subestruturas na estrutura tridimensional. Dessa forma, correlações entre subestruturas distantes na sequência linear mas próximas na estrutura tridimensional devem ser úteis para que o classificador aprenda novos padrões nas sequências, aumentando ainda mais a confiabilidade das interpretações.

Por fim, nossos trabalho é focado na obtenção de interpretações do classificador. Podemos elevar o nível da exploração dos genomas da dengue ao realizar, além das interpretações, a explicação dos classificadores e do genoma em si.

Referências Bibliográficas

Ehsaneddin Asgari, Alice C McHardy, and Mohammad RK Mofrad. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (dimotif) and sequence embedding (protvecx). *Scientific reports*, 9(1):1–16, 2019.

Xiaoyong Pan and Hong-Bin Shen. Learning distributed representations of rna sequences and its application for predicting rna-protein binding sites with a convolutional neural network. *Neurocomputing*, 305:51–58, 2018a.

Md-Nafiz Hamid and Iddo Friedberg. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics*, 35(12):2009–2016, 2019.

Quang-Thai Ho, Dinh-Van Phan, Yu-Yen Ou, et al. Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. *Analytical biochemistry*, 577:73–81, 2019.

Xiaoyong Pan and Hong-Bin Shen. Predicting rna–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, 34(20):3427–3436, 2018b.

Alok Sharma, Edwin Vans, Daichi Shigemizu, Keith A Boroevich, and Tatsuhiko Tsunoda. Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific reports*, 9(1):1–7, 2019.

Bruno MM Conque, Andre Yoshiaki Kashiwabara, and Fabricio Martins Lopes. A feature extraction approach based on complex networks for genomic sequences recognition. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1803–1807. IEEE, 2016.

Robert E Shope and James M Meegan. Arboviruses. In *Viral Infections of Humans*, pages 151–183. Springer, 1997.

WHO. *Dengue: guidelines for diagnosis, treatment, prevention and control*. World Health Organization, 2009.

WHO. Comprehensive guideline for prevention and control of dengue and dengue haemorrhagic fever. 2011.

Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, et al. The global distribution and burden of dengue. *Nature*, 496(7446): 504–507, 2013. doi: https://doi.org/10.1038/nature12060.

CNN Brasil. Brasil tem quase 1 milhão de casos de dengue em 2020, diz ministério da saúde. `https://www.cnnbrasil.com.br/saude/2020/11/24/brasil-tem-quase-1-milhao-de-casos-de-dengue-em-2020-diz-ministerio-da-saude`, 2020. Acessado: 2021-02-20.

Richard J Kuhn, Wei Zhang, Michael G Rossmann, Sergei V Pletnev, Jeroen Corver, Edith Lenches, Christopher T Jones, Suchetana Mukhopadhyay, Paul R Chipman, Ellen G Strauss, et al. Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. *Cell*, 108(5):717–725, 2002.

Jason M Mackenzie, Alexander A Khromykh, Malcolm K Jones, and Edwin G Westaway. Subcellular localization and some biochemical properties of the flavivirus kunjin nonstructural proteins ns2a and ns4a. *Virology*, 245(2):203–215, 1998.

Panisadee Avirutnan, Nuntaya Punyadee, Sansanee Noisakran, Chulaluk Komoltri, Somchai Thiemmeca, Kusuma Auethavornanan, Aroonroong Jairungsri, Rattiyaporn Kanlaya, Nattaya Tangthawornchaikul, Chunya Puttikhunt, et al. Vascular leakage in severe dengue virus infections: a potential role for the nonstructural viral protein ns1 and complement. *The Journal of infectious diseases*, 193(8):1078–1088, 2006.

Thomas J Chambers, David W McCourt, and Charles M Rice. Yellow fever virus proteins ns2a, ns213, and ns4b: Identification and partial n-terminal amino acid sequence analysis. *Virology*, 169(1):100–109, 1989.

Stephen Clum, Kurt E Ebner, and R Padmanabhan. Cotranslational membrane insertion of the serine proteinase precursor ns2b-ns3 (pro) of dengue virus type 2 is required for efficient in vitro processing and is mediated through the hydrophobic regions of ns2b. *Journal of Biological Chemistry*, 272(49):30715–30723, 1997.

Xuping Xie, Shovanlal Gayen, CongBao Kang, Zhiming Yuan, and Pei-Yong Shi. Membrane topology and function of dengue virus ns2a protein. *Journal of virology*, 87 (8):4609–4622, 2013.

Sven Miller, Stefan Kastner, Jacomine Krijnse-Locker, Sandra Bühler, and Ralf Bartenschlager. The non-structural protein 4a of dengue virus is an integral membrane protein inducing membrane alterations in a 2k-regulated manner. *Journal of Biological Chemistry*, 282(12):8873–8882, 2007.

Debashish Ray, Aaloki Shah, Mark Tilgner, Yi Guo, Yiwei Zhao, Hongping Dong, Tia S Deas, Yangsheng Zhou, Hongmin Li, and Pei-Yong Shi. West nile virus 5'-cap structure is formed by sequential guanine n-7 and ribose 2'-o methylations by nonstructural protein 5. *Journal of virology*, 80(17):8362–8370, 2006.

Maudry Laurent-Rolle, Elena F Boer, Kirk J Lubick, James B Wolfinbarger, Aaron B Carmody, Barry Rockx, Wenjun Liu, Joseph Ashour, W Lesley Shupert, Michael R Holbrook, et al. The ns5 protein of the virulent west nile virus ny99 strain is a potent antagonist of type i interferon-mediated jak-stat signaling. *Journal of virology*, 84 (7):3503–3515, 2010.

Scott B Halstead. Observations related to pathogensis of dengue hemorrhagic fever. vi. hypotheses and discussion. *The Yale journal of biology and medicine*, 42(5):350, 1970.

Leon Rosen. The emperor's new clothes revisited, or reflections on the pathogenesis of dengue hemorrhagic fever. *The American journal of tropical medicine and hygiene*, 26(3):337–343, 1977. doi: https://doi.org/10.4269/ajtmh.1977.26.337.

Hilda Palacios Serrano, María Elsa Vargas Caballero, and Tania María Aguirre Portuondo. Dengue hemorrágico en dengue primario. *Revista Cubana de Medicina Tropical*, 53(1):59–62, 2001.

WMP. Dengue. `https://www.worldmosquitoprogram.org/en/learn/mosquito-borne-diseases/dengue`, 2020. Acessado: 2020-02-14.

Fiocruz. Dengue: retorno do tipo 2 preocupa especialistas. `https://portal.fiocruz.br/noticia/dengue-retorno-do-tipo-2-preocupa-especialistas`, 2020. Acessado: 2020-02-14.

Folha UOL. Brasil registra em 2019 segundo maior número de mortes por dengue em 21 anos. `https://www1.folha.uol.com.br/cotidiano/2020/01/`

brasil-registra-em-2019-segundo-maior-numero-de-mortes-por-dengue-em-21-anos.shtml, 2020. Acessado: 2020-02-14.

Sunit Singhi, Niranjan Kissoon, and Arun Bansal. Dengue e dengue hemorrágico: aspectos do manejo na unidade de terapia intensiva. *Jornal de Pediatria*, 83(2): S22–S35, 2007.

Nicholas G Reich, Sourya Shrestha, Aaron A King, Pejman Rohani, Justin Lessler, Siripen Kalayanarooj, In-Kyu Yoon, Robert V Gibbons, Donald S Burke, and Derek AT Cummings. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *Journal of The Royal Society Interface*, 10(86):20130414, 2013.

David W Vaughn, Sharone Green, Siripen Kalayanarooj, Bruce L Innis, Suchitra Nimmannitya, Saroj Suntayakorn, Timothy P Endy, Boonyos Raengsakulrach, Alan L Rothman, Francis A Ennis, et al. Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. *The Journal of infectious diseases*, 181(1):2–9, 2000.

Eric S Halsey, Morgan A Marks, Eduardo Gotuzzo, Victor Fiestas, Luis Suarez, Jorge Vargas, Nicolas Aguayo, Cesar Madrid, Carlos Vimos, Tadeusz J Kochel, et al. Correlation of serotype-specific dengue virus infection with clinical manifestations. *PLoS Negl Trop Dis*, 6(5):e1638, 2012.

Mohammad RK Asgari, Ehsaneddin; Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11), 2015.

Ian Korf, Mark Yandell, and Joseph Bedell. *Blast.* "O'Reilly Media, Inc.", 2003.

Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, et al. Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3):272, 2018.

Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, 6(1):1–11, 2016.

Hong Kim, Seoung-Ae Lee, and Bum-Joon Kim. X region mutations of hepatitis b virus related to clinical severity. *World journal of gastroenterology*, 22(24):5467, 2016a.

Hong Kim, Seoung-Ae Lee, Seung Yeon Do, and Bum-Joon Kim. Precore/core region mutations of hepatitis b virus related to clinical severity. *World journal of gastroenterology*, 22(17):4287, 2016b.

Xiaoxia Yang, Qiongshu Wang, Beibei Liang, Fuli Wu, Hao Li, Hongbo Liu, Chunyu Sheng, Qiuxia Ma, Chaojie Yang, Jing Xie, et al. An outbreak of acute respiratory disease caused by a virus associated rna ii gene mutation strain of human adenovirus 7 in china, 2015. *PloS one*, 12(2), 2017.

IUPAC-IUB. Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents. *Biochemistry*, 9(20):4022–4027, 1970.

John S Mackenzie, Duane J Gubler, and Lyle R Petersen. Emerging flaviviruses: the spread and resurgence of japanese encephalitis, west nile and dengue viruses. *Nature medicine*, 10(12):S98–S109, 2004.

Rushika Perera and Richard J Kuhn. Structural proteomics of dengue virus. *Current opinion in microbiology*, 11(4):369–377, 2008.

EG Westaway and J Blok. Dengue and dengue hemorrhagic fever. *CAB International, New York*, pages 147–173, 1997.

BD Lindenbach, CL Murray, HJ Thiel, and CM Rice. Flaviviridae. fields virology. *Knipe DM, editor*, pages 712–746, 2013.

Nguyen Quoc Khanh Le, Edward Kien Yee Yapp, Quang-Thai Ho, N Nagasundaram, Yu-Yen Ou, and Hui-Yuan Yeh. ienhancer-5step: Identifying enhancers using hidden information of dna sequences via chou's 5-step rule and word embedding. *Analytical biochemistry*, 571:53–61, 2019a.

En-Shiun Annie Lee, Sanderz Fung, Ho-Yin Sze-To, and Andrew KC Wong. Confirming biological significance of co-occurrence clusters of aligned pattern clusters. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 422–427. IEEE, 2013.

En-Shiun Annie Lee, Sanderz Fung, Ho-Yin Sze-To, and Andrew KC Wong. Discovering co-occurring patterns and their biological significance in protein families. *BMC bioinformatics*, 15(S12):S2, 2014.

James R Carr and Fernando Pellon De Miranda. The semivariogram in comparison to the co-occurrence matrix for classification of image texture. *IEEE Transactions on geoscience and remote sensing*, 36(6):1945–1952, 1998.

Xin Zhang, Jintian Cui, Weisheng Wang, and Chao Lin. A study for texture feature extraction of high-resolution satellite images based on a direction measure and gray level co-occurrence matrix fusion algorithm. *Sensors*, 17(7):1474, 2017.

Robin Brochier, Adrien Guille, and Julien Velcin. Global vectors for node representations. In *The World Wide Web Conference*, pages 2587–2593, 2019.

Mohamed Abdel-Nasser, Antonio Moreno, and Domenec Puig. Breast cancer detection in thermal infrared images using representation learning and texture analysis methods. *Electronics*, 8(1):100, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Haoyang Zeng, Matthew D Edwards, Ge Liu, and David K Gifford. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32 (12):i121–i127, 2016.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. The structure and function of dna. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

Mark Welch, Sridhar Govindarajan, Jon E Ness, Alan Villalobos, Austin Gurney, Jeremy Minshull, and Claes Gustafsson. Design parameters to control synthetic gene expression in escherichia coli. *PloS one*, 4(9), 2009.

Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. Codon bias and heterologous protein expression. *Trends in biotechnology*, 22(7):346–353, 2004.

Scott C Perry and Robert G Beiko. Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. *Genome biology and evolution*, 2:117–131, 2010.

Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic acids research*, 34(2):564–574, 2006.

Samuel Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Current opinion in microbiology*, 1(5):598–610, 1998.

Ruth Hershberg and Dmitri A Petrov. Selection on codon bias. *Annual review of genetics*, 42:287–299, 2008.

Jean Lehmann and Albert Libchaber. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. *Rna*, 14(7):1264–1269, 2008.

Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, and Ilya Mazo. Extracting human protein interactions from medline using a full-sentence parser. *Bioinformatics*, 20(5):604–611, 2004.

Vasileios Hatzivassiloglou, Pablo A Duboue, and Andrey Rzhetsky. Disambiguating proteins, genes, and rna in text: a machine learning approach. *Bioinformatics*, 17 (suppl_1):S97–S106, 2001.

Rabie Saidi, Sabeur Aridhi, Engelbert Mephu Nguifo, and Mondher Maddouri. Feature extraction in protein sequences classification: a new stability measure. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 683–689, 2012.

Didier Debroas, Jean-François Humbert, François Enault, Gisèle Bronner, Michael Faubladier, and Emmanuel Cornillot. Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (lac du bourget–france). *Environmental microbiology*, 11(9):2412–2424, 2009.

Elizabeth D Liddy. Natural language processing. 2001.

Jyotshna Dongardive and Siby Abraham. Protein sequence classification based on n-gram and k-nearest neighbor algorithm. In *Computational Intelligence in Data Mining—Volume 2*, pages 163–171. Springer, 2016.

SM Ashiqul Islam, Benjamin J Heil, Christopher Michel Kearney, and Erich J Baker. Protein classification using modified n-grams and skip-grams. *Bioinformatics*, 34(9): 1481–1487, 2018.

C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Rafael C Gonzales and Richard E Woods. Digital image processing, 2002.

Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

Lior Rokach, Alon Schclar, and Ehud Itach. Ensemble methods for multi-label classification. *Expert Systems with Applications*, 41(16):7507–7523, 2014.

Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2019.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3): 647–665, 2014.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

OECD Statical Terms. Glossary of statical terms. `https://stats.oecd.org/glossary/detail.asp?ID=21`, 2020. Acessado: 2020-02-17.

Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–18, 2008.

NCBI. Biological sequence. `https://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/BIOSEQ.HTML`, 2020. Acessado: 2020-02-20.

K Okonechnikov, O Golosova, and M Fursov. team u.(2012). *Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics*, 28(8):1166–1167, 2012.

Shalev Itzkovitz, Eran Hodis, and Eran Segal. Overlapping codes within protein-coding sequences. *Genome research*, 20(11):1582–1589, 2010.

Tomáš Mikolov. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80, 2012.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17): 3389–3402, 1997.

Sean R Eddy et al. Multiple alignment using hidden markov models. In *Ismb*, volume 3, pages 114–120, 1995.

Da Zhang and Mansur Kabuka. Protein family classification from scratch: A cnn based deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

Kuo-Chen Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247, 2011.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Thomas Villmann, Marika Kaden, Szymon Wasik, Mateusz Kudla, Kaja Gutowska, Andrea Villmann, and Jacek Blazewicz. Searching for the origins of life–detecting rna life signatures using learning vector quantization. In *International Workshop on Self-Organizing Maps*, pages 324–333. Springer, 2019.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Atsushi Sato and Keiji Yamada. Generalized learning vector quantization. In *Advances in neural information processing systems*, pages 423–429, 1996.

Nguyen Quoc Khanh Le, Edward Kien Yee Yapp, Yu-Yen Ou, and Hui-Yuan Yeh. imotor-cnn: identifying molecular functions of cytoskeleton motor proteins using 2d convolutional neural network via chou's 5-step rule. *Analytical biochemistry*, 575: 17–26, 2019b.

Gary D Stormo, Thomas D Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, 10(9):2997–3011, 1982.

Lei Wang, Hai-Feng Wang, San-Rong Liu, Xin Yan, and Ke-Jian Song. Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Scientific reports*, 9(1):1–12, 2019.

Dexiong Chen, Laurent Jacob, and Julien Mairal. Biological sequence modeling with convolutional kernel networks. *Bioinformatics*, 35(18):3294–3302, 2019.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

Charles Yanofsky. Establishing the triplet nature of the genetic code. *Cell*, 128(5): 815–818, 2007.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018a.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Julie D Thompson, Desmond G Higgins, and Toby J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22 (22):4673–4680, 1994.

Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

Ellen R Girden. *ANOVA: Repeated measures*. Number 84. Sage, 1992.

Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pages 279–292, 1961.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018b.

Manola Laille and Claudine Roche. Comparison of dengue-1 virus envelope glycoprotein gene sequences from french polynesia. *The American journal of tropical medicine and hygiene*, 71(4):478–484, 2004.

Jerome E Foster, Shannon N Bennett, Christine VF Carrington, Helen Vaughan, and W Owen McMillan. Phylogeography and molecular evolution of dengue 2 in the caribbean basin, 1981–2000. *Virology*, 324(1):48–59, 2004.

Long Li, Shee-Mei Lok, I-Mei Yu, Ying Zhang, Richard J Kuhn, Jue Chen, and Michael G Rossmann. The flavivirus precursor membrane-envelope protein complex: structure and maturation. *Science*, 319(5871):1830–1834, 2008.

Mikako Ito, Ken-Ichiro Yamada, Tomohiko Takasaki, Basu Pandey, Reiko Nerome, Shigeru Tajima, Kouichi Morita, and Ichiro Kurane. Phylogenetic analysis of dengue viruses isolated from imported dengue patients: possible aid for determining the countries where infections occurred. *Journal of travel medicine*, 14(4):233–244, 2007.

Claire M Midgley, Aleksandra Flanagan, Hai Bac Tran, Wanwisa Dejnirattisai, Kriangkrai Chawansuntati, Amonrat Jumnainsong, Wiyada Wongwiwat, Thaneeya Duangchinda, Juthathip Mongkolsapaya, Jonathan M Grimes, et al. Structural analysis of a dengue cross-reactive antibody complexed with envelope domain iii reveals the molecular basis of cross-reactivity. *The Journal of Immunology*, 188(10):4971–4979, 2012.

JA Patil, S Cherian, AM Walimbe, A Bhagat, J Vallentyne, M Kakade, PS Shah, and D Cecilia. Influence of evolutionary events on the indian subcontinent on the phylogeography of dengue type 3 and 4 viruses. *Infection, Genetics and Evolution*, 12(8):1759–1769, 2012.

Jacky Flipse and Jolanda M Smit. The complexity of a dengue vaccine: a review of the human antibody response. *PLoS neglected tropical diseases*, 9(6):e0003749, 2015.

Jose L Slon Campos, Monica Poggianella, Sara Marchese, Monica Mossenta, Jyoti Rana, Francesca Arnoldi, Marco Bestagno, and Oscar R Burrone. Dna-immunisation with dengue virus e protein domains i/ii, but not domain iii, enhances zika, west nile and yellow fever virus infection. *PloS one*, 12(7):e0181734, 2017.

Shoshana J Wodak and Joel Janin. Location of structural domains in proteins. *Biochemistry*, 20(23):6544–6552, 1981.

Basu Dev Pandey, Kouichi Morita, Futoshi Hasebe, Maria del Carmen Parquet, and Akira Igarashi. Molecular evolution, distribution and genetic relationship among the dengue 2 viruses isolated from different clinical severity. *Southeast Asian journal of tropical medicine and public health*, 31(2):266–272, 2000.

Basu Dev Pandey and Akira Igarashi. Severity-related molecular differences among nineteen strains of dengue type 2 viruses. *Microbiology and immunology*, 44(3): 179–188, 2000.