Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Classificação de RNA telomerase usando extração de características e técnicas de aprendizado de máquina

Ana Luísa Salvador Alvarez

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientadora
Profa. Maria Emília Walter

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Classificação de RNA telomerase usando extração de características e técnicas de aprendizado de máquina

Ana Luísa Salvador Alvarez

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Profa. Maria Emília Walter (Orientadora)
CIC/UnB

Prof. Dr. Li Weigang
CIC/UnB

Prof. João Victor de Araújo Oliveira
Instituto Federal de Brasília

Prof. Dr. Marcelo Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 24 de fevereiro de 2023

Dedicatória

Dedico esse trabalho às pessoas que me inspiraram na busca do conhecimento ao longo da minha vida, sejam eles professores, cientistas, colegas ou amigos. Em especial dedico aos meus pais e ao meu marido, que sempre me apoiaram nessa busca.

Agradecimentos

Gostaria de agradecer a todos que generosamente cederam tempo e compartilharam seus conhecimentos para que esse trabalho fosse possível de ser executado e avaliado. Muito obrigada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

A detecção de RNA telomerase por métodos de Biologia Molecular e de Bioinformática vem se mostrando extremamente difícil, devido à variabilidade na sua sequência genômica e na sua estrutura nas diferentes espécies de organismos. Esse problema é importante pois a RNA telomerase é considerável para a saúde e longevidade humanas e sua predição é de grande interesse científico. Por outro lado, como o uso de técnicas de aprendizado de máquina vem mostrando bons resultados para a classificação de DNA e de RNA, neste trabalho, utilizamos modelos obtidos com essas técnicas para verificar se poderiam ser utilizados para a classificação de RNA telomerase. Neste contexto, este trabalho propõe métodos baseados em aprendizado de máquina supervisionado para classificar a RNA telomerase. Foram obtidos modelos de classificação de RNA telomerase usando quatro algoritmos de aprendizado de máquina supervisionados: *Random Forest* (RF), *Naive Bayes* Gaussiano (NBG), *Naive Bayes* de Bernoulli (NBB) e Máquina de Vetor de Suporte, ou *Suport Vector Machine* (SVM), com uma etapa anterior de uso de métodos de extração de características (em inglês, *features*): transformada de Fourier com representação real (FR), curva-z (FZ) e pseudopotencial de interação elétron-ion (EIIP); e Redes Complexas (RC). Os resultados atingidos a partir dos modelos obtidos foram avaliados usando as métricas tradicionais de aprendizado de máquina. Foram testados dezesseis (16) modelos para classificar RNA telomerase, considerando os métodos de aprendizado de máquina e de extração de características. Os quatro melhores modelos, segundo a métrica de *F1-score*, foram os modelos baseados em *Random Forest* com extração de características por transformada de Fourier com representação real (RF FR), por Redes Complexas (RF RC) e por pseudopotencial de interação elétron-íon (RF EIIP); e *Naive Bayes* Gaussiano com extração de características por transformada de Fourier com representação real (NBG FR). Entre eles, obtiveram resultados menos satisfatórios RF RC e NBG FR, considerando-se os verdadeiros positivos, que são de maior interesse para esse estudo.

Palavras-chave: RNA telomerase, classificação de RNA telomerase, extração de características, aprendizado de máquina

Abstract

Detecting telomerase RNA by Molecular Biology and Bioinformatics has been proving to be extremely difficult, due to the variability in its genomic sequence and structure in different species of organisms. This is a significant problem since telomerase RNA is important to human health and longevity and its prediction is of great scientific interest.

And since the use of machine learning techniques is presenting good results in classifying DNA and RNA, in this work we used models obtained with those machine learning techniques to verify if they could be used to telomerase RNA classification. In this context, this work suggests methods based on supervised Machine Learning algorithms to classify telomerase RNA. Telomerase RNA classification models were obtained by the use of four Machine Learing algorithms: Random Forest, Gaussian Naive Bayes Gaussiano, Bernoulli Naive Bayes de Bernoulli and Suport Vector Machine, with an anterior step of feature extraction: Fourier transform with real representation, z-curve and electron-ion interection pseudopotential; and Complex Network. The results achieved from the models obtained were evaluated by Machine Learning traditional metrics. Sixteen (16) models for telomerase RNA classification, considering Machine Learning methods and feature extraction. The best four models, according to F1-score metric, were the ones based on Random Forest, with Fourier transform and real representation feature extraction, with Complex Network feature extraction and with electron-ion interaction pseudopotential feature extraction; and Gaussian Naive Bayes with Fourier transform and real representation feature extraction. Among them, the least satisfactory results were obtained by Random Forest with Complex Network and Gaussian Naive Bayes with Fourier transform and real representation, when we considered true positive results,that are of most interest to this study.

Keywords: RNA telomerase, RNA telomerase classification, feature extraction, machine learning, Terc, Ter, extração de características de sequências biológicas

Sumário

1 Introdução **1**
1.1 Problema e Hipótese . . . 3
1.2 Justificativa . . . 3
1.3 Objetivos . . . 3
1.3.1 Principal . . . 3
1.3.2 Específicos: . . . 3
1.4 Descrição dos Capítulos . . . 4

2 Biologia Molecular e Bioinformática **5**
2.1 Conceitos básicos . . . 5
2.1.1 Ácidos nucleicos . . . 5
2.1.2 Proteínas . . . 8
2.1.3 Dogma Central da Biologia Molecular - síntese de proteínas . . . 8
2.2 Conceitos básicos de Bioinformática . . . 10
2.3 Telômero e telomerase . . . 12
2.3.1 Componentes da telomerase . . . 13
2.3.2 Métodos para predição de telomerase RNA . . . 14
2.4 Levantamento bibliográfico sobre RNA telomerase . . . 16

3 Aprendizado de Máquina **17**
3.1 Conceitos básicos . . . 17
3.1.1 Aprendizado de máquina supervisionado . . . 18
3.1.2 Aprendizado de máquina não-supervisionado . . . 18
3.1.3 Aprendizado de máquina semi-supervisionado . . . 19
3.1.4 Aprendizado por reforço . . . 19
3.1.5 Treinamento e teste . . . 20
3.1.6 Otimização de hiperparâmetros e validação cruzada . . . 20
3.2 Extração de características . . . 21
3.2.1 Extração de características para sequências biológicas . . . 22

3.3 Técnicas de Aprendizado de Máquina 24
3.3.1 *Random Forest* 24
3.3.2 *Naive Bayes* 24
3.3.3 Máquinas de Vetores de Suporte 26
3.3.4 Métricas 27

4 Projeto **31**
4.1 Método de geração de modelos 31
4.2 Dados de entrada 31
4.3 Extração de características 33
4.4 *Tuning* de hiperparâmetros, treinamento e teste 35
4.4.1 *Random Forest* 36
4.4.2 *Naive Bayes* Gaussiano 36
4.4.3 *Naive Bayes* de Bernoulli 36
4.4.4 Máquina de Vetores de Suporte 37
4.4.5 Treinamento e teste 37

5 Resultados e discussão **38**
5.1 *Random Forest* 38
5.2 *Naive Bayes* Gaussiano 39
5.3 *Naive Bayes* de Bernoulli 39
5.4 Máquina de Vetores de Suporte 40
5.5 Observações gerais 41

6 Conclusões e trabalhos futuros **43**

Referências **45**

Anexo **51**

I Resultados completos dos experimentos **52**

Lista de Figuras

2.1 Ácidos nucleicos: à esquerda RNA e à direita DNA [1] 6
2.2 Estrutura da ribose e da desoxirribose [2]. 6
2.3 Estrutura de dupla hélice do DNA [3]. 7
2.4 Estrutura química de um aminoácido, evidenciando um grupo amino, um grupo lateral e um grupo carboxila [3]. 8
2.5 Dogma da Biologia Molecular: Replicação do DNA (evidenciando o reparo e recombinação genética), síntese do RNA (ou transcrição), síntese de proteínas (ou tradução), evidenciando os aminoácidos componentes delas [4]. 9
2.6 Repetição do telômero e estrutura da alça-t, evidenciando as extremidades da molécula e a colaboração delas na formação estrutural [4]. 12
2.7 Estrutura do núcleo do telômero inclui uma transcriptase reversa (TERT) e proteínas associadas, um molde de RNA (TER) e um pequeno pedaço do DNA do telômero [5]. 14
2.8 Relação evolutiva de filos metazoários basais e classes com RNA telomerase identificados no trabalho de [6]. 15

3.1 Esquema de divisão dos dados utilizados para criar o modelo de aprendizado de máquina, para as fases de treinamento e teste. 20
3.2 Random Forest, evidenciando os dados de entrada, as árvores e a classificação [7] 25
3.3 *Naive Bayes* Gaussiano: curvas Gaussianas baseadas em médias e desvios padrão dos dados de características [8]. 26
3.4 Melhor hiperplano gerado por uma máquina de vetores de suporte, separando os objetos de classe diferentes, evidenciados como símbolos diferentes (círculo e triângulo) [9]. 27
3.5 Matriz de confusão: visualização dos resultados obtidos em um modelo de classificação, relacionando os valores verdadeiros com os previstos. 30

4.1 Fluxograma que representa o método. Inicialmente os conjuntos de dados positivo e negativo, obtidos dos bancos de dados, são passados por métodos de extração de características (*feature extraction*). Cada um dos arquivos obtidos por essa extração é utilizado para a criação de modelos de classificação de RNA telomerase, através de algoritmos de aprendizado de máquina. 32
4.2 Comparação entre estruturas de telomerases de vertebrados e equinodermos. Dois domínios estruturais, T-PK e box H/ACA são conservados em ambos, enquanto o domínio CR4/5 dos vertebrados contém um stem loop P6.1, nos equinodermos ele é substituído por um domínio funcionalmente equivalente [6]. 33

Lista de Tabelas

5.1 Médias e desvios padrão dos resultados obtidos nos modelos de *Random Forest*, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z. 39
5.2 Médias e desvios padrão dos resultados obtidos nos modelos de *Naive Bayes* Gaussiano, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z. 40
5.3 Médias e desvios padrão dos resultados obtidos nos modelos de *Naive Bayes* de Bernoulli, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z. 40
5.4 Médias e desvios padrão dos resultados obtidos nos modelos de Máquinas de Vetores de Suporte, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z. 41
5.5 Comparação das matrizes de confusão dos melhores modelos obtidos, onde RF FR refere-se ao modelo obtido com *Random Forest* e transformada de Fourier com representação real, RF RC refere-se ao modelo obtido com *Random Forest* e Redes Complexas, RF EIIP refere-se ao modelo obtido com *Random Forest* e método de extração de EIIP, NBG FR refere-se ao modelo obtido com *Naive Bayes* Gaussiano com transformada de Fourier e representação real. 42

I.1 Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 52
I.2 Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 52
I.3 Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 53
I.4 Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de Transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 53
I.5 Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de Transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 53
I.6 Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 54
I.7 Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 54
I.8 Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 54
I.9 Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 55
I.10 Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 55

I.11 Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 55
I.12 Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 56
I.13 Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 56
I.14 Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 56
I.15 Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 57
I.16 Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo. 57

Capítulo 1

Introdução

Por muitos anos, acreditou-se que o papel principal do RNA seria a de permitir, a partir das informações do DNA, produzir (ou sintetizar) proteínas, que inclui o RNA transportador (em inglês, *transfer RNA* - tRNA) e o RNA ribosomal (em inglês, *ribosomal RNA* - rRNAs). Cerca de 20 anos atrás, essa conjectura mudou drasticamente com a identificação do primeiro pequeno RNA não-codificador (em inglês, *micro non-coding RNA* - microRNA). Essa denominação foi criada em referência ao RNA que se refere àquelas moléculas que são capazes de codificar (ou sintetizar) proteínas, os RNAs mensageiros (em inglês, *messenger RNA* - mRNA). Essa descoberta permitiu o surgimento de uma área de estudo, em Biologia Molecular e Bioinformática, a de RNAs não-codificadores de proteínas (em inglês, *non-coding RNAs* - ncRNAs), que possuem pouca ou nenhuma capacidade de codificar (ou sintetizar) proteínas [10]. Os ncRNAs possuem importantes papeis nos mecanismos celulares, como controle de doenças, apesar de não codificarem proteínas [11].

As extremidades físicas dos cromossomos lineares eucarióticos são denominadas de telômeros, que formam estruturas especiais que encapsulam as extremidades dos cromossomos para evitar sua degradação [12]. Em células eucarióticas e também em organismos unicelulares, o DNA telomérico é replicado pelas ações da telomerase, uma enzima especializada chamada transcriptase [12]. A telomerase é um complexo ribonucleoproteico essencial para a estabilidade do genoma. Desempenha essa função por meio da adição de DNA repetitivo às extremidades dos cromossomos. A enzima telomerase é composta por uma transcriptase reversa especializada (em inglês *reverse transcriptase* - TERT), que utiliza um domínio molde em uma subunidade de RNA (TERC), a qual fornece uma sequência molde, para adicionar repetidamente DNA telomérico nas extremidades dos cromossomos [13, 14].

A telomerase desempenha um papel vital na senescência celular das células somáticas. Mutações nos genes da telomerase afetam a capacidade de proliferação das células e têm

sido associadas a três doenças humanas, disceratose congênita (DKC), anemia aplástica (AA) e fibrose pulmonar idiopática (FPI) [15, 16].

A ausência da atividade da telomerase nas células somáticas leva ao encurtamento dos cromossomos após cada divisão celular, causando a senescência celular e, em algum momento, a morte celular [17]. Por outro lado, a reativação da atividade da telomerase nas células somáticas é observada em 90% dos cânceres, permitindo o crescimento intenso e descontrolado da célula cancerosa [16].

Dado a relevância dos temas citados, na área de Bioinformática, um dos principais desafios é a extração de informações a partir de sequenciamento de material genético (DNA ou RNA), que são geradas de forma exponencial. Esse desafio requer o desenvolvimento de ferramentas e métodos capazes de transformar esses dados, muito heterogêneos pois são gerados por diferentes sequenciadores e em diferentes instituições, em conhecimento biológico [18].

A Bioinformática utiliza, para diferentes problemas advindos do sequenciamento de DNA em genomas (área de Genômica) e de transcritos (área de Transcritômica), técnicas de aprendizado de máquina e de Inteligência Artificial, que tenta extrair conhecimento biológico a partir dos dados primários (chamados de dados brutos) [18, 19]. O aprendizado de máquina é um campo multidisciplinar que usa uma variedade de algoritmos para simular o aprendizado humano, pela exploração de padrões nos dados, transformando-os em informações [20, 18]. Na tecnologia, a Inteligência Artificial, do inglês *Artificial Intelligence* (AI) é a inteligência demonstrada por máquinas ao executar tarefas complexas associadas a seres inteligentes, além de também ser um campo de estudo acadêmico, no qual o principal objetivo é de executar funções de modo autônomo [21].

Em particular, na Genômica, busca-se extrair informações como o local e a estrutura dos genes e suas funções [22]. Em Transcritômica, busca-se predizer funções dentro dos mecanismos celulares. Mais recentemente, a identificação de funções regulatórias dos ncRNAs tem sido feitas a partir da estrutura secundária de RNA, extraída da sua estrutura sequencial (ou a sequência de nucleotídeos em DNA ou RNA) [18].

Algoritmos de aprendizado de máquina tratam apenas de dados numéricos, o que significa que sequências biológicas devem ser traduzidas em sequências numéricas [23]. Para esses algoritmos, devem ser definidas características (em inglês, *features*), para as quais existem métodos de seleção e extração de atributos a partir de sequências biológicas, que usam meta heurísticas e modelos matemáticos [23].

1.1 Problema e Hipótese

Existe uma dificuldade grande em detectar RNA telomerase por métodos baseados em homologia, devido à rápida evolução das sequências em nível de nucleotídeos, que sofrem variação em tamanho e nas suas estruturas secundárias [14]. Tanto quanto saibamos, não há métodos baseados em técnicas computacionais para classificação de RNA telomerase.

A hipótese é que seja possível utilizar métodos baseados em aprendizado de máquina para a classificação de RNA telomerase de forma automatizada, isto é, sem a intervenção humana, e portanto, mais eficiente.

1.2 Justificativa

O uso de técnicas de aprendizado de máquina vem mostrando bons resultados para a classificação de DNA e de RNA [20].

Devido à sua importância para a saúde e longevidade humana, estudar a estrutura e função da telomerase é de enorme interesse. No entanto, as pesquisas tanto em Biologia Molecular quanto em Bioinformática sobre telomerase se mostrado extremamente difíceis devido à baixa quantidade natural da telomerase e uma incapacidade geral de superexpressar de forma solúvel as proteínas de comprimento total, bem como a falta de uma caracterização completa dos componentes da holoenzima da telomerase. Tanto o tamanho quanto a grande diversidade da sequência da telomerase em diferentes organismos tornam difícil identificá-la e caracterizá-la nas várias espécies [24].

1.3 Objetivos

1.3.1 Principal

O objetivo principal desse trabalho é avaliar o desempenho de diferentes técnicas de extração de características (*features*) e de aprendizado de máquina na classificação de telomerase.

1.3.2 Específicos:

- criar uma coleta de dados com sequências de RNA telomerase, a partir de informações coletadas no banco de dados *The Telomerase Database* [15];
- realizar a etapa de extração de características em sequências biológicas de TERCs a partir de técnicas de extração de características baseadas em modelos matemáticos;

- usar técnicas de tuning de hiperparâmetros, de forma a melhorar a performance dos algoritmos de aprendizado de máquina supervisionados;
- selecionar algoritmos de aprendizado de máquina supervisionado, a serem utilizados para a classificação de sequências de RNA telomerase.

1.4 Descrição dos Capítulos

No Capítulo 2, são apresentados de forma breve conceitos básicos de Biologia Molecular, basicamente cromosssomos, ácidos nucléicos e proteínas, além do Dogma Central da Biologia Molecular. Ainda, são descritos o telômero e a telomerase, que é objeto desse trabalho, assim como métodos de identificação de telomerase. Em seguida, são apresentados aspectos importantes de Bioinformática, como métodos de comparação de sequências, bancos de dados de sequências biológicas e *workflows*, que são utilizados em projetos de transcritomas.

Em seguida, no Capítulo 3, são apresentados conceitos de Aprendizado de Máquina e particularmente técnicas utilizadas comumente, como extração de características de sequências biológicas e métodos de aprendizado de máquina, como *Random Forest*, *Naive Bayes* Gaussiano, *Naive Bayes* de Bernoulli e Máquina de Vetores de Suporte.

Então, no Capítulo 4, é apresentado o método utilizado para extração de características, a serem utilizadas nos modelos de classificação de RNA telomerase em sequências biológicas.

Já no Capítulo 5, são discutidos os resultados obtidos a partir da aplicação do método proposto.

Finalmente, no Capítulo 6 concluímos o trabalho e apresentamos sugestões para trabalhos futuros.

Capítulo 2

Biologia Molecular e Bioinformática

Neste capítulo, apresentamos conceitos básicos de Biologia Molecular e de Bioinformática. Na Seção 2.1, conceitos como ácidos nucleicos, DNA, RNA e proteínas são abordados, assim como o Dogma Central da Biologia Molecular. Na Seção 2.3, são descritos telômeros e telomerase, sendo citados os componentes da telomerase, além de métodos para sua predição. Na Seção 2.2, são explorados conceitos básicos de Bioinformática, juntamente com artigos bibliográficos que embasam o presente trabalho.

2.1 Conceitos básicos

Biologia Molecular é o ramo da biologia que estuda os processos biológicos de um ponto de vista molecular, que envolve o estudo da estrutura e função do material genético e seus produtos expressos, as proteínas.

2.1.1 Ácidos nucleicos

Os ácidos nucleicos presentes no núcleo das células estão envolvidos no armazenamento, na transmissão e no processamento das informações genéticas de uma célula [25].

A estrutura dos ácidos nucleicos está relacionada à sua função. Essas moléculas são polímeros formados por cadeiras de nucleotídeos, cuja composição (tipo e sequência) determina suas características químicas. Essas características definem a sua interação com outras macromoléculas na célula, em particular, com as proteínas e sua conformação espacial (forma da molécula). A conformação espacial está diretamente relacionada à função e à atividade das macromoléculas na célula [26].

Organismos vivos contém dois tipos de ácidos nucleicos: ácido ribonucleico, abreviado por RNA e ácido desoxirribonucleico, ou DNA [27]. Ambos são biopolímeros, constituídos

de subunidades de nucleotídeos unidas por ligações fosfodiéster [25], conforme demostrado na Figura 2.1.

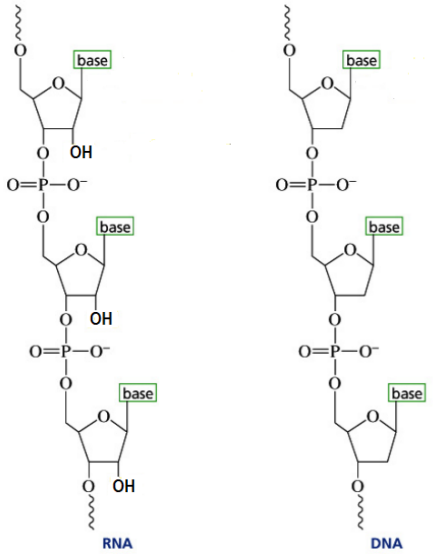

Figura 2.1: Ácidos nucleicos: à esquerda RNA e à direita DNA [1]

DNA

DNA se compõe de uma cadeia dupla, formada de duas cadeias simples, chamadas fitas. Ela tem uma espinha dorsal formada por repetições da mesma unidade básica. Essa unidade básica é formada por uma molécula de açúcar, chamada de desoxirribose, anexada a um resíduo de fosfato [27].

A molécula de açúcar contem cinco átomos de carbono, e eles são nomeados de 1' até 5', conforme Figura 2.2.

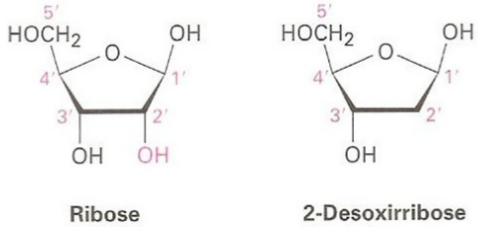

Figura 2.2: Estrutura da ribose e da desoxirribose [2].

A ligação que cria a espinha dorsal é entre o carbono 3' de uma unidade, o resíduo de fosfato, e o carbono 5' da próxima unidade. Por isso, as moléculas de DNA também têm uma orientação que, por convenção, começa com a ponta 5' e termina na ponta 3' [27].

Anexado a cada carbono 1' na espinha dorsal, existem outras moléculas chamadas bases. Há quatro tipos de bases: adenina (A), guanina (G), citosina (C) e tiamina (T).

Bases A e G pertencem a um grupo maior, chamado de purina, enquanto C e T pertencem às pirimidinas. Quando vemos uma unidade básica de molécula de DNA, consistindo de açúcar, fosfato e base, a chamamos de nucleotídeo [27].

Moléculas de DNA são cadeias duplas, que ficam juntas numa estrutura helicoidal. As cadeias ficam juntas pois cada base é pareada (ou ligada) a uma base na outra cadeia, conforme Figura 4.2. Base A pareia com T, sendo complementares, assim como a base C, que é complementar da G. Moléculas de DNA podem ser medidas através da quantidade de pares de bases que possuem [27].

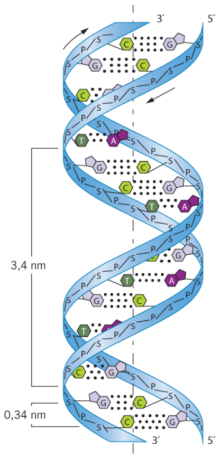

Figura 2.3: Estrutura de dupla hélice do DNA [3].

RNA

Moléculas de RNA são quimicamente bem semelhantes às de DNA. As principais diferenças são:

- No RNA, o açúcar é ribose, ao invés de desoxirribose, conforme Figura 2.2;
- No RNA, ao invés de timina (T), encontra-se uracila (U), que também pareia com adenina (A);
- RNA não forma dupla hélice [27].

Em um passado muito recente, os RNAs eram considerados meros intermediários entre o genoma e as proteínas. Descobertas recentes envolvendo uma variedade de novos genes de RNA não codificador (ncRNA), papéis biológicos e mecanismos de ação têm mostrado que a diversidade e a importância de ncRNAs foram subestimados. Atualmente, sabe-se que os RNAs funcionais que não codificam proteínas desempenham papéis importantes [28].

Eles estão envolvidos em várias atividades celulares, como silenciamento de genes, replicação, regulação da expressão gênica, transcrição, estabilidade cromossômica, estabilidade da proteína, translocação e localização e modificação do RNA, processamento e estabilidade [28].

2.1.2 Proteínas

A maior parte das substâncias nos seres vivos são proteínas, das quais existem muitos tipos. Proteínas estruturais atuam como blocos de construção de tecidos, enquanto outras, conhecidas como enzimas, agem como catalisadoras para para reações químicas [27].

Uma proteína é uma cadeia de moléculas mais simples, chamadas de aminoácidos. Todo aminoácido tem um átomo de carbono central, ao qual são ligados um átomo de hidrogênio, um grupo amino (NH_2), um grupo carboxi (COOH) e uma cadeia lateral, conforme Figura 2.4. É essa cadeia lateral que distingue um aminoácido de outro [27].

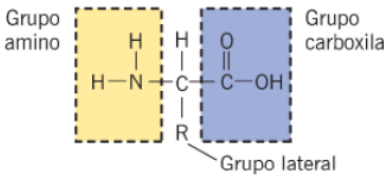

Figura 2.4: Estrutura química de um aminoácido, evidenciando um grupo amino, um grupo lateral e um grupo carboxila [3].

2.1.3 Dogma Central da Biologia Molecular - síntese de proteínas

O Dogma Central da Biologia foi proposto por Francis Crick em 1958 (publicado em 1970 na revista científica Nature [29]). Afirma que o DNA codifica a produção de RNA por transcrição, o RNA codifica a produção de proteínas por tradução e as proteínas não codificam a produção nem de proteínas nem de RNA nem de DNA. Crick afirmou que uma vez que a informação tenha passado para a proteína já não torna a sair [30].

O fluxo da informação genética nas células é, portanto, de DNA para RNA e deste para proteína, conforme Figura 2.5. Todas as células, desde a bactéria até os seres humanos, expressam sua informação genética dessa maneira [4].

No entanto, novas descobertas alteraram o Dogma Central e hoje sabe-se que por transcrição reversa a informação passa do RNA para o DNA, como no caso dos retrovírus, e que o DNA pode ser traduzido diretamente em proteínas, em sistemas in vitro usando ribossomas de E. coli [30].

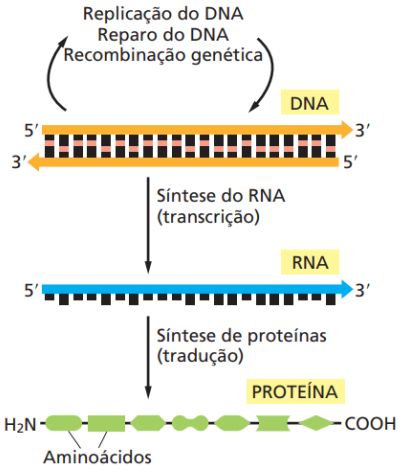

Figura 2.5: Dogma da Biologia Molecular: Replicação do DNA (evidenciando o reparo e recombinação genética), síntese do RNA (ou transcrição), síntese de proteínas (ou tradução), evidenciando os aminoácidos componentes delas [4].

Replicação

A capacidade das células de manter um alto grau de organização em um ambiente caótico depende da duplicação exata de grandes quantidades de informação genética armazenadas na forma química de DNA. Esse processo, denominado replicação do DNA, deve ocorrer antes de a célula produzir duas células-filhas geneticamente iguais [4].

O material genético de um organismo é transmitido da célula mãe para as células filhas durante a divisão celular. A transmissão fiel do material genético de uma célula ou um organismo para outro depende da capacidade de replicação das moléculas de DNA [3].

A manutenção da ordem também requer a vigilância contínua e o reparo dessa informação genética, uma vez que o DNA contido na célula é repetidamente danificado por compostos químicos e radiação oriundos do ambiente, por acidentes térmicos e por moléculas reativas [4].

A replicação do DNA é extraordinariamente exata. Moléculas constituídas de centenas de milhões de pares de nucleotídeos são duplicadas com poucos erros ou até mesmo sem erros [3].

Transcrição e retrotranscrição

Todo o RNA de uma célula é produzido a partir da transcrição do DNA, em um processo que apresenta certas similaridades em relação ao processo de replicação do DNA. A transcrição inicia com a abertura e a desespiralização de uma pequena porção da dupla-hélice de DNA, o que expõe as bases em cada fita do DNA [4]. Uma das duas fitas da dupla-hélice de DNA, então, serve como um molde para a síntese de uma molécula de

RNA. Assim como na replicação de DNA, a sequência de nucleotídeos da cadeia de RNA é determinada pelo pareamento de bases complementares entre os nucleotídeos a serem incorporados e o DNA-molde [4].

Uma transcriptase reversa é uma forma especial de polimerase que utiliza um molde de RNA para produzir uma fita de DNA [4]. Uma transcriptase reversa é uma enzima usada para gerar DNA complementar (cDNA) a partir de um molde de RNA, um processo denominado transcrição reversa, ou retrotranscrição [29]. Ao contrário de uma crença amplamente aceita, o processo não viola os fluxos de informação genética descritos pelo dogma central clássico, já que as transferências de informação do RNA para o DNA são explicitamente consideradas possíveis [29].

2.2 Conceitos básicos de Bioinformática

A Bioinformática é um campo interdisciplinar focado no desenvolvimento de ferramentas e métodos de *software* e *hardware* para apoiar o armazenamento, organização e análise de dados biológicos, particularmente relacionados ao sequenciamento genético [31].

A pesquisa em Biologia Molecular inicialmente era voltada quase que exclusivamente para a complexa tarefa de investigar DNA [27]. Porém, posteriormente, o RNA passou a ser investigado, pois descobriu-se que ele exercia papeis fundamentais nos mecanismos celulares [27].

O Projeto Genoma Humano foi iniciado em 1989 considerado finalizado em abril de 2003 e teve como objetivo o sequenciamento dos 3,1 bilhões de bases nitrogenadas do genoma humano. O grupo que se propôs a sequenciar o genoma humano consistiu em um consórcio público internacional, liderado pelo *National Human Genome Research Institute* (NHGRI), subordinado ao *National Institute of Health* (NIH) dos Estados Unidos. Reunindo equipes de pesquisa e laboratórios de vários países, seu objetivo central era o sequenciamento completo do genoma humano. O produto final do projeto consistiu no sequenciamento de um genoma-referência composto por genomas de diferentes povos [32, 33].

Transcritômica é uma área onde se estudam transcritomas de uma célula ou tecido. Um transcritoma é constituído por uma coleção de RNAs que são transcritos em um organismo. O transcritoma é dinâmico e representa um retrato interessante do estado celular de um organismo [34, 35].

Proteômica é o estudo das proteínas e suas interações em uma célula [36]. A proteômica envolve a aplicação de tecnologias para a identificação e quantificação do conteúdo total de proteínas presentes em uma célula, tecido ou organismo [37].

O primeiro projeto genoma humano foi considerado concluído, porém ainda em forma de rascunho, em 2001. Pouco tempo depois, as sequências do genoma de vários organismos modelo foram determinadas. Observamos também que o Instituto Nacional de Pesquisa do Genoma Humano (NGHRI) criou tecnologias avançadas para o sequenciamento de DNA, de 70 milhões de dólares, com o objetivo de sequenciar um genoma humano com US$ 1.000. Além disso, em anos recentes, temos o denominado sequenciamento de alto desempenho, que permitiu ampliar exponencialmente o número de genomas sequenciados de uma grande variedade de espécies [38].

Sequenciamento de alto desempenho refere-se a novas tecnologias usadas para sequenciamento de DNA e RNA. Pode sequenciar centenas de milhares de genes ou fragmentos de genomas em um curto período de tempo. Usa técnicas que permitem sequenciamento paralelo massivo.

Essas técnicas vem sendo utilizadas em medicina de precisão personalizada. por exemplo, as variantes/mutações de sequência detectadas nesses projetos têm sido amplamente utilizadas para diagnóstico de doenças, prognóstico, decisão terapêutica, tratamento de doenças e acompanhamento de pacientes [39, 40].

Por fim, os bancos de dados contendo dados gerados por sequenciamento de alto desempenho têm grande utilidade para a comunidade científica, pois uma grande parte deles é de acesso público. Esses repositórios de dados vêm acompanhados de ferramentas de busca, e constituem-se hoje em fontes importantes de informações que podem ser utilizados em projetos de pesquisa ao redor do mundo [41].

O Centro Nacional de Informação Biotecnológica (em inglês, *National Center for Biotechnology Information* - NCBI), é uma seção da Biblioteca Nacional de Medicina dos Estados Unidos (em inglês, *United States National Library of Medicine* - (NLM), um ramo dos Institutos Nacionais de Saúde (em inglês, *National Institutes of Health* [42]. O NCBI mantém bancos de dados de alguns aspectos importantes da biologia molecular de organismos como nucleotídeo, proteína, estrutura, taxonomia, genoma e expressão [41].

De particular interesse para este trabalho, citamos o banco de dados Telomerase [15], uma ferramenta *web* que fornece dados que podem ser utilizados em estudos de estrutura, função e evolução da ribonucleoproteína telomerase. O objetivo deste banco de dados é servir à comunidade de pesquisa, fornecendo uma compilação abrangente de informações conhecidas sobre a telomerase e seu substrato, os telômeros. A coleção de informações inclui sequências, alinhamentos, estruturas secundárias e terciárias, mutações conhecidas por causar doenças de deficiência de telomerase humana e pesquisadores atuais da telomerase.

2.3 Telômero e telomerase

Para evitar a perda de genes com o desgaste dos cromossomos, as extremidades dos cromossomos eucarióticos têm tampões de DNA especializados chamados de telômeros. Telômeros consistem em centenas ou milhares de uma curta sequência de DNA [43].

Telômeros são as extremidades dos cromossomos. Os telômeros contêm sequências nucleotídicas repetidas que permitem que as extremidades dos cromossomos sejam replicadas de maneira eficiente, conforme Figura 2.6. Os telômeros também desempenham uma outra função: as sequências de DNA repetidas, juntamente com as regiões adjacentes a elas, formam estruturas que evitam que as extremidades cromossômicas sejam confundidas com uma molécula de DNA quebrada que necessita de reparo pela célula [4]. Os telômeros formam estruturas especiais que cobrem as extremidades dos cromossomos para evitar a degradação por ataque nucleolítico e para distinguir os terminais dos cromossomos das quebras de fita dupla do DNA [12].

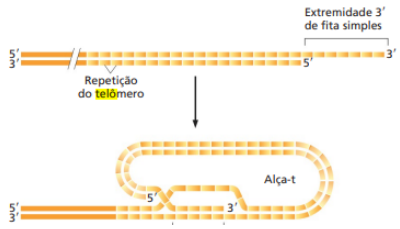

Figura 2.6: Repetição do telômero e estrutura da alça-t, evidenciando as extremidades da molécula e a colaboração delas na formação estrutural [4].

As repetições que formam os telômeros são consumidas lentamente ao longo dos vários ciclos de divisão, estabelecendo uma barreira que protege as regiões internas dos cromossomos que contêm os genes (ao menos por certo tempo). O encurtamento de telômeros foi relacionado ao envelhecimento de células e a perda progressiva dos telômeros pode explicar a razão pela qual as células somente podem se dividir um determinado número de vezes [43].

Com poucas exceções, os telômeros são compostos principalmente de DNA repetitivo associado a proteínas que interagem especificamente com o DNA telomérico de fita dupla ou simples ou entre si, formando complexos altamente ordenados e dinâmicos envolvidos na manutenção dos telômeros e na regulação do comprimento. Em células proliferativas e organismos unicelulares, o DNA telomérico é replicado pelas ações da telomerase, uma transcriptase reversa especializada [12].

Uma característica exclusiva da telomerase é que ela carrega seu próprio molde de RNA. A telomerase também possui vários outros domínios proteicos necessários à ligação da enzima às extremidades dos cromossomos [4].

A telomerase é responsável pela manutenção do comprimento do telômeros pela adição de sequências repetitivas ricas em guanina. A atividade dela é exibida em gametas e células-tronco e tumorais [44]. É uma enzima da ribonucleoproteína (RNP) que mantém a função dos telômeros e a integridade do genoma pela adição de repetições de DNA telomérico nas extremidades dos cromossomos, que é a solução mais comum para o problema de replicação final de cromossomos lineares em eucariotos [45].

As sequências de DNA telomérico são reconhecidas por proteínas ligadoras de DNA que reconhecem uma sequência específica de DNA e atraem uma enzima, chamada de telomerase, que repõe essas sequências cada vez que a célula se divide. A telomerase reconhece a extremidade de uma sequência telomérica de DNA existente e a estende na direção 5'-3', utilizando um molde de RNA que compõe a própria enzima para sintetizar novas cópias da repetição [4].

O complexo ribonucleoproteico da telomerase (RNP) é essencial para a estabilidade do genoma e desempenha esse papel por meio da adição de DNA repetitivo às extremidades dos cromossomos [13].

2.3.1 Componentes da telomerase

Ao contrário de outras transcriptases reversas, a telomerase é única por ser um complexo ribonucleoproteico, onde o componente de RNA, ou seja telomerase RNA, não apenas fornece o modelo para a síntese de repetições de DNA telomérico, mas também desempenha papéis essenciais na catálise, acumulação, processamento da ponta 3' da telomerase, localização e montagem de holoenzimas [46].

Os componentes principais da telomerase incluem a transcriptase reversa da telomerase (TERT) que catalisa a polimerização do DNA e o componente integral da telomerase RNA (TERC) que fornece o molde para a síntese do DNA telomérico [6].

O núcleo catalítico da telomerase é fornecido pelo polipeptídeo da transcriptase reversa da telomerase (TERT), que faz a transcrição reversa de um trecho curto da porção de RNA da telomerase (TERC) para estender as extremidades 3' do cromossomo [47].

TERC abriga domínios estruturais que conferem atividade enzimática da telomerase e servem como base para a ligação de uma variedade de proteínas acessórias [48]. O TERC atua como modelo para a adição das sequências de DNA às extremidades dos cromossomos, o que é conhecido como alongamento dos telômeros. É um componente da enzima telomerase e é essencial para o funcionamento adequado da telomerase [12, 49, 46].

Esses dois componentes trabalham juntos para adicionar sequências de DNA às extremidades dos cromossomos, um processo conhecido como alongamento dos telômeros. O RNA da telomerase atua como um modelo para a adição das sequências de DNA,

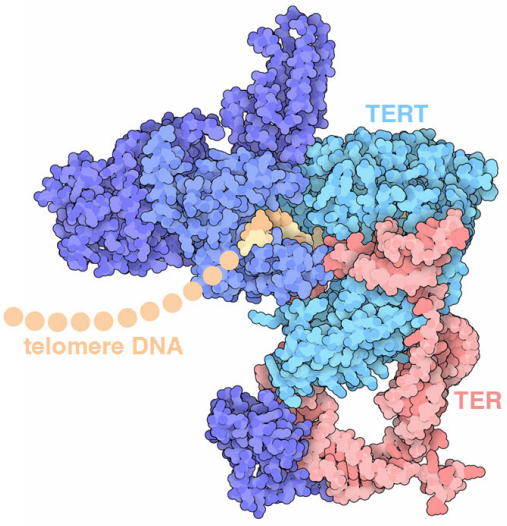

Figura 2.7: Estrutura do núcleo do telômero inclui uma transcriptase reversa (TERT) e proteínas associadas, um molde de RNA (TER) e um pequeno pedaço do DNA do telômero [5].

enquanto a transcriptase reversa da telomerase é responsável pela síntese real do novo DNA [46, 50].

Logeswaran et al. [6] mostraram que uma análise comparativa mostra que três domínios funcionais, template-pseudoknot (T-PK), CR4/5 e box H/ACA, são conservados entre vertebrados e linhagens basais de metazoários, indicando uma origem monofilética das telomerases animais com um mecanismo de biogênese relacionado ao snoRNA. No entanto, telomerases ao longo de linhagens animais separadas evoluíram com elementos estruturais divergentes nos domínios T-PK e CR4/5 [6], conforme Figura 4.2.

2.3.2 Métodos para predição de telomerase RNA

Os RNAs não-codificadores de proteínas (ncRNAs) são uma área de pesquisa bastante profícua em Bioinformática. Descobertas recentes revelaram novos tipos de ncRNAs desempenhando uma variedade de papéis nos mecanismos celulares, desde a regulação da expressão gênica até atividades catalíticas. Acredita-se também que outros tipos ainda serão encontrados. Métodos computacionais desenvolvidos para genes codificadores de proteínas geralmente falham na busca por ncRNAs. As funcionalidades dos RNAs não-codificadores geralmente dependem fortemente de sua estrutura secundária, o que torna a descoberta dos próprios ncRNAs e suas funcionalidades é muito diferente dos RNA codificadores de proteínas [28].

As abordagens bioquímicas para identificação de telomerase são muitas vezes desafiadoras e às vezes inviáveis devido à baixa abundância da telomerase, à falta de ferramentas

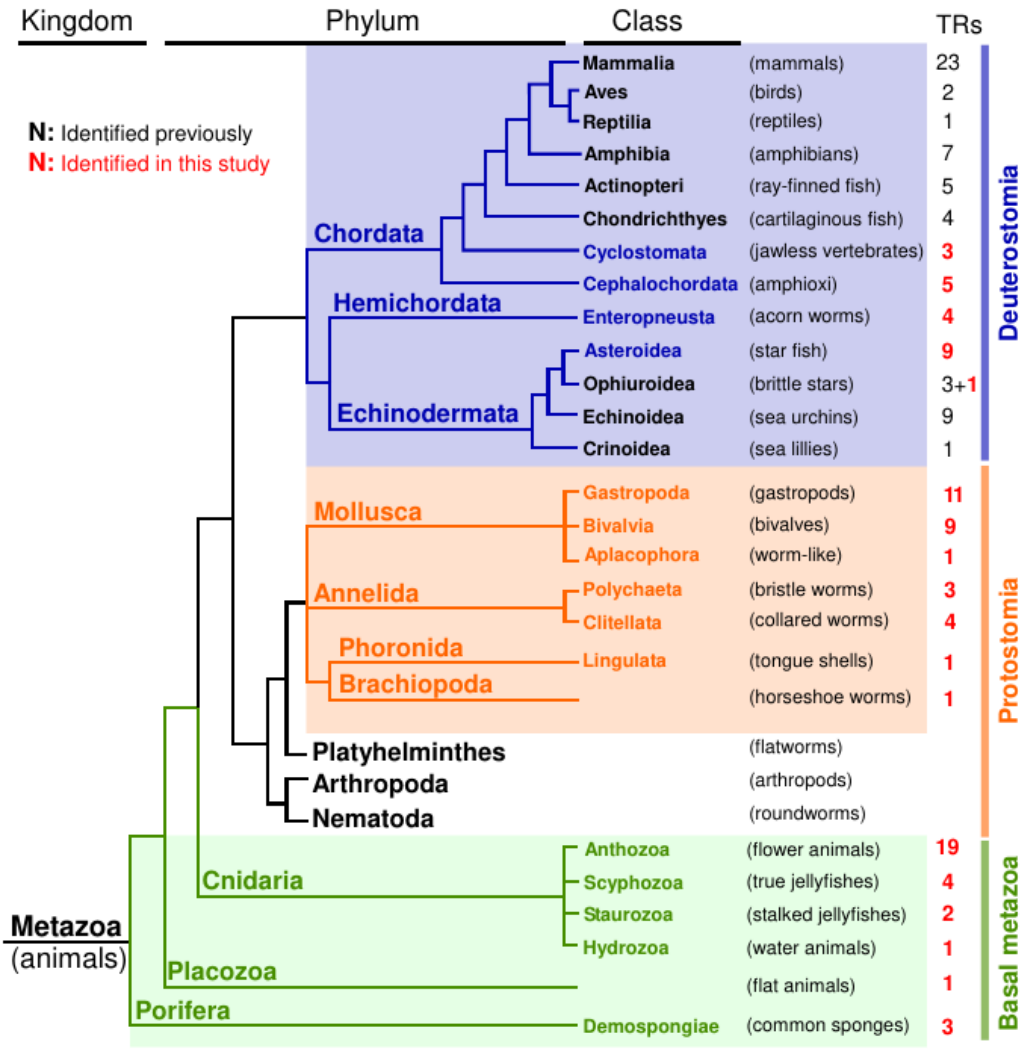

Figura 2.8: Relação evolutiva de filos metazoários basais e classes com RNA telomerase identificados no trabalho de [6].

de manipulação genética ou à falta de um procedimento de cultura escalável [6].

A telomerase varia dramaticamente em tamanho, sequência primária e estrutura secundária e caminhos de biogênese, apresentando baixa similaridade entre as espécies, mesmo próximas. Isso que dificulta bastante sua identificação através de métodos baseados em homologia [6].

Apesar de sua função primária profundamente conservada e semelhanças na sua estrutura que parecem se repetir em diferentes reinos dos organismos eucarióticos, as telomerases são muito difíceis de encontrar por pesquisas de homologia mesmo dentro de grupos filogeneticamente relativamente próximos [14].

Para entender a origem e a evolução da telomerase em todo o reino animal, Logeswaran et al. [6] empregou uma abordagem de bioinformática baseada em estrutura e guiada por

filogenia para identificar 82 novas telomerases de oito filos de metazoários anteriormente inexplorados, incluindo as esponjas de ramificação basal [6].

Em Waldl et al. [14], foram utilizados métodos baseados em homologia, aplicados em um grupo filogenético pequeno, o Saccharomycetales, para identificar 46 telomerases, das quais 27 não haviam sido relatadas antes. A ideia principal foi reduzir o espaço de busca usando sintenia[1], usando genes codificadores de proteínas que, sabidamente, envolvem regiões de DNA que provavelmente contém um gene de teomerase Essas regiões de sintenia foram usadas em diferentes sequências utilizando Blast, MEME, GLAM2 e INFERNAL [14].

2.4 Levantamento bibliográfico sobre RNA telomerase

Os trabalhos estudados para a elaboração desse projeto foram os de Bonidia [52], Logeswaran [6] e Waldl [14].

O trabalho de Bonidia [52] propôs um novo estudo de abordagens de extração de características baseadas em características matemáticas (mapeamento numérico com Fourier, entropia e Redes Complexas). Foram analisadas sequências de RNAs não-codificadores longos (em inglês, *long non-coding RNAs* - lncRNAs), tendo sido avaliada uma proposta envolvendo lncRNAs e mRNAs. Na proposta, foram validadas as características obtidas por métodos matemáticos em diferentes problemas de classificação, para prever a classe de lncRNA. Sua robustez foi analisada em cenários, com dados desbalanceados. Os resultados experimentais demonstraram um estudo aprofundado de várias características matemáticas, um novo *pipeline* de extração de recursos e seu bom desempenho e robustez para classificação de sequências de RNA distintas [52].

Já o trabalho de Logeswaran [6] teve o objetivo de entender a origem e a evolução do RNA telomerase em todo o reino animal, Neste projeto, foi utilizada uma abordagem de bioinformática baseada em estrutura e guiada por filogenia. Foram identificados 82 novos TRs de 8 filos de metazoários anteriormente inexplorados.

Waldl e colaboradores [14] desenvolveram um projeto para detectar RNA telomerase em genomas de Saccharomycotina. Os pesquisadores utilizaram métodos baseados em homologia, variando tamanho e estruturas secundárias. Foram identificados 27 novos RNAs telomerase no subgrupo Saccharomycetacea.

[1]Sintenia, que significa literalmente "na mesma fita", é a propriedade que as características ocorrem no mesmo cromossomo, e é frequentemente usada para significar que elas são contíguas dentro daquele cromossomo [51].

Capítulo 3

Aprendizado de Máquina

Nesse capítulo, são apresentados conceitos básicos de Aprendizado de Máquina, que serão utilizados neste trabalho. Na Seção 3.1, são abordados os paradigmas de aprendizado de máquina e, em seguida, as técnicas de desenvolvimento de modelos, como separação treinamento-teste, otimização de hiperparâmetros e validação cruzada. Além disso, são apresentados brevemente conceitos básicos de extração de características, especificamente para sequências biológicas. Na Seção 3.3, são descritos os algoritmos *Random Forest*, *Naive Bayes Gaussiano*, *Naive Bayes de Bernoulli* e Máquina de Vetores de Suporte (em inglês, *Support Vector Machine* - SVM), além das métricas utilizadas para medir *performance* dos modelos.

3.1 Conceitos básicos

Aprendizado de máquina é uma área da Ciência da Computação que visa propor algoritmos que aprendem a partir de uma coleção de exemplos de um certo fenômeno e são empregados em diferentes aplicações. Esses exemplos podem vir da natureza, serem feitos por humanos ou gerados por algum outro algoritmo [53]. A área trata de projetos de algoritmos que extraem automaticamente informações valiosas dos dados. A ênfase aqui é em "automático", ou seja, o aprendizado de máquina está preocupado com metodologias de propósito geral que podem ser aplicadas a muitos conjuntos de dados [54]. E ainda, o uso de computadores neste caso tenta simular o aprendizado humano, explorando padrões nos dados e aplicando o autoaperfeiçoamento para aprimorar continuamente o desempenho das tarefas de aprendizado [20].

O crescimento exponencial de dados em Biologia Molecular e Medicina nos últimos anos estimulou a aplicação de inúmeras técnicas de aprendizado de máquina para resolver problemas relevantes nessas áreas [20]. Consequentemente, novos métodos computacionais são necessários para analisar e extrair informação dessas sequências [23]. A combinação

de abordagens da Ciência da Computação com os princípios da evolução molecular revolucionou a área de pesquisa de Evolução Molecular [20].

Métodos de aprendizado de máquina têm mostrado grande aplicabilidade em bioinformática, pois auxiliam a extrair informações relevantes de conjuntos de dados biológicos diversos [23]. As técnicas de aprendizado de máquina são frequentemente integradas a métodos de bioinformática, bem como bancos de dados e redes biológicas, além de uma etapa de curadoria humana, para aprimorar o treinamento e a validação, para detectar as melhores características e permitir estudo de características e modelos [20].

Em seguida, descreveremos de forma breve os paradigmas de aprendizado de máquina, a saber, aprendizado de máquina supervisionado, aprendizado de máquina não-supervisionado, aprendizado de máquina semi-supervisionado e aprendizado por reforço.

3.1.1 Aprendizado de máquina supervisionado

Técnicas de aprendizado que extraem associações entre atributos independentes e um atributo dependente designado (o rótulo). O aprendizado supervisionado usa um conjunto de dados de treinamento para desenvolver um modelo de previsão consumindo dados de entrada e valores de saída. O modelo pode então fazer previsões dos valores de saída para um novo conjunto de dados. O desempenho dos modelos desenvolvidos usando aprendizado supervisionado depende do tamanho e variância do conjunto de dados de treinamento para obter melhor generalização e maior poder preditivo para novos conjuntos de dados [55]. Por exemplo, as técnicas de Máquina de Vetores de Suporte (em inglês, *Support Vector Machine* - SVM) e de *Random Forest* pertencem a este paradigma de aprendizado supervisionado.

3.1.2 Aprendizado de máquina não-supervisionado

Algoritmos de aprendizado de máquina não supervisionado são projetados para descobrir estruturas ocultas em conjuntos de dados não rotulados, nos quais a saída desejada é desconhecida. Esse mecanismo encontrou muitos usos nas áreas de compactação de dados, detecção de valores atípicos, ou *outliers*, classificação, aprendizado humano e assim por diante. A abordagem geral do aprendizado envolve treinamento por meio de modelos de dados probabilísticos. Alguns exemplos comumente utilizados de aprendizado não-supervisionado são agrupamento e redução de dimensionalidade, expectativa e minimização, clusterização k-médias, análise de componentes principais, entre outros [55].

3.1.3 Aprendizado de máquina semi-supervisionado

O aprendizado de máquina semi-supervisionado usa uma combinação de um pequeno número de conjuntos de dados rotulados e um grande número de conjuntos de dados não rotulados para gerar uma função de modelo ou classificador. Em contraste, os dados não rotulados são relativamente baratos e prontamente disponíveis. A metodologia de aprendizado de máquina semi-supervisionada opera em algum lugar entre as diretrizes de aprendizado não supervisionado (dados de treinamento não rotulados) e aprendizado supervisionado (dados de treinamento rotulados) e pode produzir uma melhoria considerável na precisão do aprendizado. O aprendizado de máquina semi-supervisionado recentemente ganhou maior destaque, devido à disponibilidade de grandes quantidades de dados não rotulados para diversas aplicações para dados da web, dados de mensagens, dados de estoque, dados de varejo, dados biológicos, imagens e assim por diante. Essa metodologia de aprendizado pode agregar valor de significado prático e teórico, especialmente em áreas relacionadas ao aprendizado humano, como fala, visão e caligrafia, que envolvem uma pequena quantidade de instrução direta e uma grande quantidade de experiência não rotulada [55]. Exemplos de métodos deste paradigma são modelos gerativos, separação de baixa densidade, métodos baseados em gráficos e abordagens heurísticas.

3.1.4 Aprendizado por reforço

A metodologia de aprendizado por reforço, do inglês *Reinforcement Learning* (RL) envolve a exploração de uma sequência adaptativa de ações ou comportamentos por um agente inteligente (agente RL) em um determinado ambiente com uma motivação para maximizar a recompensa cumulativa. A ação do agente inteligente desencadeia uma mudança observável no estado do ambiente. Essa técnica de aprendizagem sintetiza um modelo de adaptação treinando-se para um determinado conjunto de ações experimentais e respostas observadas ao estado do ambiente. Em geral, essa metodologia pode ser vista como um paradigma de aprendizado de tentativa e erro de teoria de controle com recompensas e punições associadas a uma sequência de ações. O agente RL muda sua política com base na experiência coletiva e consequentes recompensas. O aprendizado por reforço busca ações passadas que explorou e que resultaram em recompensas. Algumas aplicações que vêm sendo solucionadas por métodos deste paradigma são carros automáticos, softwares que traçam relatórios do mercado financeiro, projeções sobre determinado cenário e a ferramenta de indicação de vídeos do YouTube [55]. Métodos associados a este paradigma são critério de optimalidade, força bruta, valor de função (como o método de Monte Carlo), otimização estocástica e algoritmo de Dyna [55].

3.1.5 Treinamento e teste

No aprendizado de máquina, os dados são divididos em dados de treinamento e dados de teste. No início, parte dos dados é utilizada para desenvolver o modelo, ou seja, parte dos dados vão para a fase de treinamento [56].

Depois de desenvolver um modelo, com base nos dados de treinamento, e as métricas indicarem que estão bons o suficiente, pode-se testar o modelo com os dados restantes, conhecidos como dados de teste.

Quando os resultados obtidos a partir do modelo treinado com os dados de treinamento e com os dados de teste, o modelo de aprendizado de máquina estará pronto para ser utilizado, ou seja, consegue, a partir de novos dados, ainda não utilizados nas fases de treinamento e teste, tomar decisões corretas sobre a classificação (ou categorização) dos dados [56].

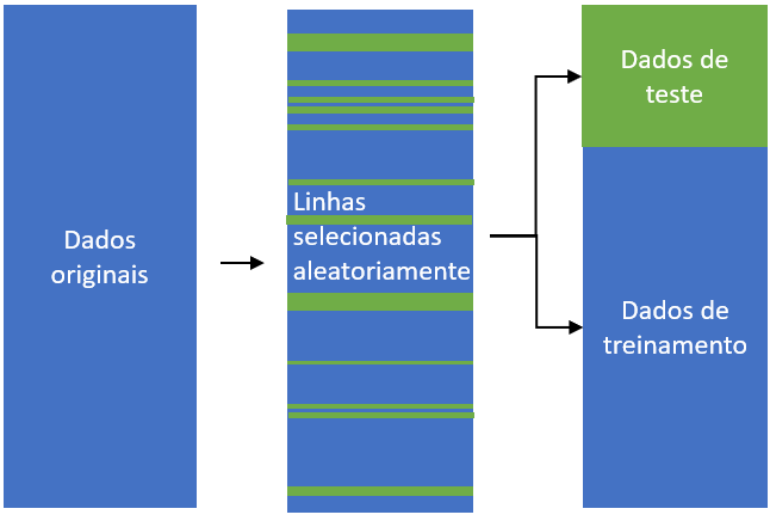

Figura 3.1: Esquema de divisão dos dados utilizados para criar o modelo de aprendizado de máquina, para as fases de treinamento e teste.

3.1.6 Otimização de hiperparâmetros e validação cruzada

A maioria dos algoritmos de aprendizado de máquina é configurada por um conjunto de hiperparâmetros cujos valores devem ser escolhidos com cuidado e que geralmente afetam consideravelmente o desempenho da eficácia do modelo. Para evitar um processo manual demorado e irreprodutível de tentativa e erro para encontrar configurações de hiperparâmetros de bom desempenho, vários métodos de otimização de hiperparâmetros podem ser empregados [57].

Métodos de aprendizado de máquina tentam construir modelos que capturam informações de interesse do objeto de estudo com base em dados fornecidos. Em geral, os algoritmos de aprendizado de máquina apresentam um conjunto de hiperparâmetros que devem ser determinados antes do início do treinamento [58]. A escolha de hiperparâmetros pode afetar significativamente o desempenho do modelo resultante, mas determinar um bom conjunto de valores pode ser complexo. Portanto, uma estratégia de busca disciplinada e teoricamente sólida é essencial [58].

O objetivo é encontrar um conjunto de valores de hiperparâmetros que nos forneça o melhor modelo para nossos dados em um período de tempo razoável [59].

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta técnica é amplamente empregada em problemas onde o objetivo da modelagem é a predição [60].

3.2 Extração de características

O problema de extrair características de dados é de importância crítica para a aplicação bem-sucedida do aprendizado de máquina. A extração de características, como geralmente é entendida, busca uma transformação ideal dos dados de entrada em um vetor de características (normalmente constituído por valores numéricos reais), que pode ser usado como entrada para um algoritmo de aprendizado de máquina [61].

A extração automática de características, e em seguida, a escolha e geração de modelos preditivos de aprendizado de máquina, são utilizado para estudar sistemas biológicos complexos, apresentando resultados satisfatórios [20]. Um dos principais objetivos da extração de características é aumentar a precisão dos modelos de aprendizado de máquina, pois auxilia a identificar características importantes dos dados de entrada, que constituem a entrada para os algoritmos de aprendizado de máquina, ao mesmo tempo, potencialmente removendo ruído e redundância dos dados de entrada [61].

Embora os algoritmos de aprendizado de máquina tenham sido aplicados com sucesso a um grande número de problemas relacionados à sequências genômicas, os resultados são fortemente afetados pelo tipo e número de características extraídas. Esse efeito tem motivado novos algoritmos e propostas de *pipelines* que envolvem uma etapa de extração de características, onde se tenta extrair informações relevantes de dados biológicos, o que é um desafio [52].

Para esta etapa, são utilizadas técnicas matemáticas que são capazes de extrair características de dados biológicos [62]. Por exemplo, um método relacionado à ponderação de características é a normalização, que redimensiona os dados em um intervalos adequados [61, 63]. A normalização é uma técnica de dimensionamento na qual os valores são

redimensionados para que fiquem entre 0 e 1. Também é conhecido como dimensionamento Min-Max e tem fórmula como apresentada na equação 3.1 [64, 63]:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

3.2.1 Extração de características para sequências biológicas

Considerando que a classificação de sequências biológicas é uma tarefa importante em bioinformática [65], a extração de características desempenha um papel crucial na classificação, pois é usada para converter as sequências do genoma em um conjunto de valores quantitativos [66]. A extração de características, a partir de transformações nos dados de entrada, gera um vetor de características [52]. Outro aspecto importante da extração de características é extrair informações importantes dos dados de entrada de forma compacta, removendo ruídos e redundância, o que aumenta a precisão dos modelos de aprendizado de máquina [61].

Bonidia et al. [52] propuseram um *pipeline* de extração de características utilizando as técnicas de mapeamento numérico, Fourier, entropia e redes complexas, descritas em seguida.

Transformada de Fourier e mapeamentos numéricos

A Transformada Discreta de Fourier é uma abordagem de processamento digital amplamente utilizada, que pode revelar periodicidades ocultas, ou seja, algum padrão de repetição que não é facilmente percebido. O método tem sido amplamente utilizado para estudar periodicidades e elementos repetitivos em sequências de DNA, genomas e estruturas de proteínas [67].

De acordo com Bonidia et al. [52], uma representação numérica deve ser usada para a transformação ou mapeamento de dados genômicos. Existem algumas técnicas de mapeamento como Voss, inteiro, real, curva Z, EIIP e números complexos [52].

- Representação de Voss Essa representação emprega quatro sequências indicadoras binárias para denotar a presença de um nucleotídeo de cada tipo [68].
- Representação de Inteiros Esta representação é unidimensional. Esse mapeamento pode ser obtido substituindo os quatro nucleotídeos (T, C, A, G) de uma sequência biológica por números inteiros (0, 1, 2, 3), respectivamente [52].
- Representação Real A representação real usa mapeamento real baseado na propriedade de complemento do mapeamento complexo. Este mapeamento aplica valores

decimais negativos para as purinas (A, G) e valores decimais positivos para as pirimidinas (C,T), conforme equação 3.2 [52].

$$r[n] = \begin{cases} -0,5, s[n] = G \\ -1,5, s[n] = A \\ 0,5, s[n] = C \\ 1,5, s[n] = T \end{cases} \tag{3.2}$$

- Representação de Curva-Z O esquema Z-curve é uma curva tridimensional para codificar sequências de DNA com semântica mais biológica [52]. O método da Curva Z foi criado como uma forma de mapear visualmente uma sequência de DNA ou RNA. Diferentes propriedades da Curva Z, como sua simetria e periodicidade, podem fornecer informações únicas sobre a sequência de DNA [46].
- EIIP Do inglês *electron-ion interaction pseudopotential* - EIIP). A energia de elétrons deslocalizados [1] em aminoácidos e nucleotídeos foi calculada como o EIIP [70]. É baseada em uma sequência numérica que representa a distribuição de energias de elétrons livres [52]. A equação 3.3 mostra o valor EIIP para os nucleotídeos [52, 70].

$$b[n] = \begin{cases} 0.0806, s[n] = G \\ 0.1260, s[n] = A \\ 0.1340, s[n] = C \\ 0.1335, s[n] = T \end{cases} \qquad n = 0, 1, ..., N-1. \tag{3.3}$$

- Representação de Números Complexos Este mapeamento numérico tem a vantagem de traduzir melhor algumas das características dos nucleotídeos em propriedades matemáticas e representa a natureza complementar dos pares AT e CG [52].

Entropia

A teoria da informação tem sido amplamente utilizada em bioinformática. Em particular, a entropia é uma medida da incerteza associada a um experimento probabilístico [65]. A teoria da informação pode ser definidas como blocos conectados compostos por uma fonte de mensagens, um codificador, um canal, um decodificador e um receptor [65].

[1] Elétrons deslocalizados são elétrons em uma molécula que não estão associados a um único átomo ou a uma ligação covalente. Elétrons deslocalizados são contidos dentro de um orbital que se estende ao longo de vários átomos adjacentes [69].

Com esse conceito de entropia, Bonidia et al. [23] extraem características relevantes para RNAs não-codificadores longos. Para gerar um experimento probabilístico, foi usada uma técnica de bioinformática chamada $k-mer$, que mapeia cada sequência na frequência das k bases vizinhas, gerando informações estatísticas [52].

Redes complexas

A pesquisa na área de Redes Complexas tem crescido constantemente devido ao seu potencial para representar, caracterizar e modelar uma ampla gama de sistemas e fenômenos naturais intrincados, sendo bastante explorada por matemáticos, cientistas da computação e biólogos [71]. O trabalho de Ito et al. [72] apresenta o BASiNET, uma ferramenta de alinhamento para classificar sequências biológicas com base na extração de características de medições de redes complexas [72]. Baseado nesse trabalho, Bonidia et al. [52] propuseram um modelo de extração de características baseado em redes complexas.

3.3 Técnicas de Aprendizado de Máquina

Essa seção discute alguns algoritmos de aprendizado de máquina supervisionados, que foram utilizados neste trabalho.

3.3.1 *Random Forest*

Random Forest é uma ferramenta popular de aprendizado de máquina baseada em árvores que pode ser facilmente adaptável a diferentes tipos de dados. Isso torna a técnica de *Random Forest* particularmente atraente para análise de dados genômicos de alta dimensão [73].

Algoritmos de *Random Forest* possuem vários hiperparâmetros que devem ser definidos pelo usuário, como quantidade de árvores na floresta, profundidade máxima das árvores e mínimo de amostras em um nó folha [74].

É bem conhecido que, na maioria dos casos, a técnica de *Random Forest* funciona razoavelmente bem com os valores padrão dos hiperparâmetros especificados nos softwares disponíveis com os algoritmos baseados nessa técnica. No entanto, ajustes de hiperparâmetros (em inglês, *hyperparameter tuning*), pode melhorar o desempenho dos algoritmos de *Random Forest* [74].

3.3.2 *Naive Bayes*

Naive Bayes é um algoritmo de aprendizado de máquina simples que utiliza o Teorema de Bayes, que dá a probabilidade de ocorrência do evento, juntamente com uma forte

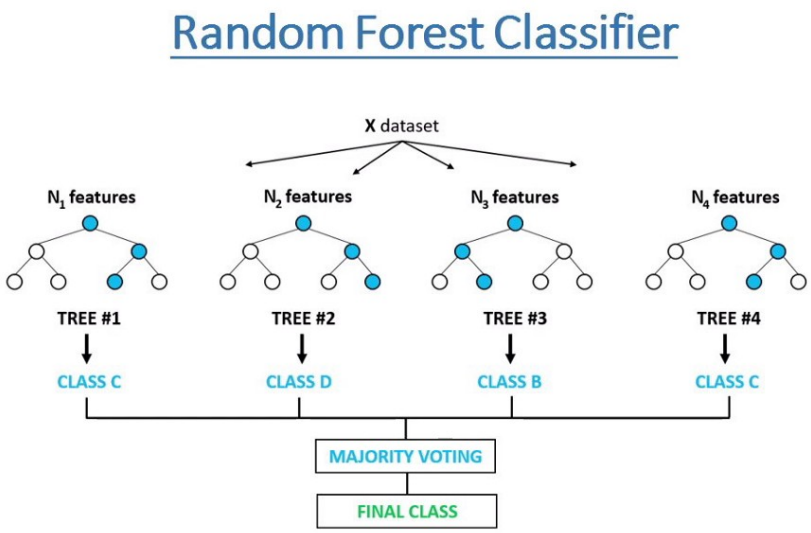

Figura 3.2: Random Forest, evidenciando os dados de entrada, as árvores e a classificação [7]

suposição de que os atributos são condicionalmente independentes dada a classe. Ou seja, que as características são todas igualmente importantes para o resultado. É um classificador probabilístico, o que significa que, dada uma entrada, ele prevê a probabilidade da entrada ser classificada para todas as classes [75].

O classificador *Naive Bayes* simplifica muito o aprendizado, assumindo que as características são independentes em determinada classe. Embora a independência seja geralmente uma suposição ruim, na prática, esse algoritmo frequentemente compete bem com classificadores mais sofisticados [76].

O teorema de Bayes se evidencia na equação 3.4.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.4}$$

onde
P(A|B) = Probabilidade de A dada a evidência de que B já ocorreu;
P(B|A) = Probabilidade de B dada a evidência de que A já ocorreu;
P(A) = Probabilidade de que A ocorrerá;
P(B) = Probabilidade de que B ocorrerá.

Naive Bayes Gaussiano

O algoritmo *Naive Bayes* Gaussiano implementa a classificação assumindo a probabilidade das características serem gaussianos. Devido à sua simplicidade e por ser extremamente

rápido em comparação com métodos mais sofisticados, o *Naive Bayes* Gaussiano também tem sido amplamente aplicado em problemas de predição em bioinformática [77].

Figura 3.3: *Naive Bayes* Gaussiano: curvas Gaussianas baseadas em médias e desvios padrão dos dados de características [8].

***Naive Bayes* de Bernoulli**

O algoritmo *Naive Bayes* de Bernoulli é utilizado para dados que são distribuídos de acordo com a distribuição multivariada de Bernoulli. É bem adequado para classificação binária [78]. Esse algoritmo pode ter um desempenho melhor em alguns conjuntos de dados, especialmente em menores conjuntos de dados [79].

Algumas vantagens de usar este algoritmo para classificação binária:

- É muito rápido em comparação com outros algoritmos de classificação [80].
- É rápido e também pode lidar facilmente com características irrelevantes [80].

A regra de decisão para *Naive Bayes* de Bernoulli segue conforme a equação 3.5.

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{3.5}$$

3.3.3 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (SVMs, do Inglês *Support Vector Machines*) constituem uma técnica de aprendizado que vem recebendo crescente atenção da comunidade de Aprendizado de Máquina. Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado [81, 82].

As SVMs são embasadas pela teoria de aprendizado estatístico. Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu [81, 82].

A SVM entrega como resultado o chamado hiperplano, que melhor separa os dois grupos. No espaço bidimensional, esta é uma linha simples, conforme Figura 3.4. Este plano é usado para decidir em qual classe um objeto de dados cai [9].

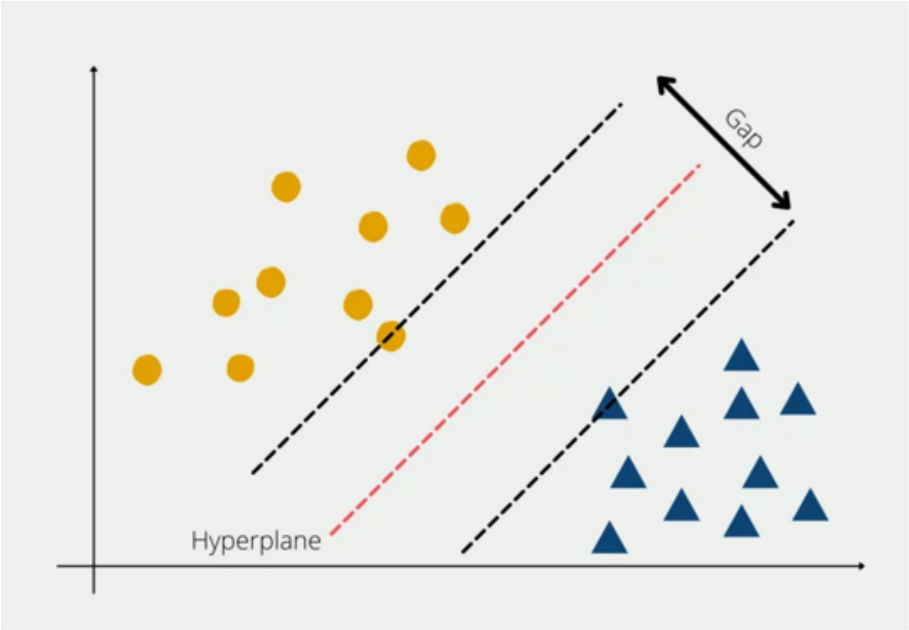

Figura 3.4: Melhor hiperplano gerado por uma máquina de vetores de suporte, separando os objetos de classe diferentes, evidenciados como símbolos diferentes (círculo e triângulo) [9].

3.3.4 Métricas

Uma tarefa importante na construção de qualquer modelo de aprendizado de máquina é a de avaliação de sua *performance* [83]. As métricas de avaliação conseguem medir a qualidades dos resultados apresentados por um modelo de aprendizado de máquina. Existem diferentes métricas para as tarefas de classificação e regressão [83].

Acurácia

A acurácia mede a frequência com a qual o classificador prevê corretamente, ou seja, rotula (categoriza) corretamente um dado de entrada [83]. É medida pela quantidade de acertos do modelo divido pelo total da amostra [84], ou seja, a razão entre o número de previsões corretas e o número total de previsões [83]. A equação 3.6 mostra como a acurácia é calculada, considerando verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN).

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.6}$$

A acurácia é útil quando as classes são bem balanceadas, ou seja, quando se tem uma quantidade dados parecida entre as classes. Se 99% do seu conjunto de dados pertence a uma classe, então, se o modelo considerar qualquer dado sendo de uma mesma classe, ele pode alcançar um nível de acurácia de 99% [83].

Como dados normalmente não tem classes balanceadas, como por exemplo, e-mail de spam, fraude de cartão de crédito e diagnóstico médico, para uma melhor avaliação do modelo e uma visão completa da avaliação do modelo, outras métricas, como *recall* e precisão, também devem ser consideradas [83].

Precisão

A precisão explica quantos dos casos previstos corretamente foram preditos como sendo positivos. No caso desse trabalho positivo é considerado o trecho de código genético identificado como codificador de telomerase. A precisão é útil nos casos em que o falso positivo é uma preocupação maior do que o falso negativo [83]. A precisão de um rótulo é definida como o número de verdadeiros positivos (VP) dividido pelo número de positivos previstos, ou seja verdadeiros positivos (VP) mais falsos positivos (FP) [83, 84], conforme equação 3.7.

$$precisão = \frac{VP}{VP + FP} \tag{3.7}$$

Recall

Recall explica quantos dos casos positivos reais, foram capazes de se prever corretamente com o modelo, ou seja, quais são verdadeiros positivos. É uma métrica útil nos casos em que o Falso Negativo (FN) é mais preocupante do que o Falso Positivo (FP). É importante em casos médicos em que não importa se disparamos um alarme falso, mas os casos positivos reais não devem passar despercebidos [83].

Representa qual a porcentagem de dados classificados como positivos comparado com a quantidade real de positivos que existem na amostra [84]. É definida como o número de verdadeiros positivos (VP) dividido pelo número total de positivos reais, ou seja, verdadeiros positivos (VP) mais falsos negativos(FN) [83], conforme equação 3.8.

$$recall = \frac{VP}{VP + FN} \tag{3.8}$$

F1-score

Essa métrica une precisão e *recall* afim de trazer um número único que determine a qualidade geral do nosso modelo [83].

Ela desfavorece mais os valores extremos e pode ser uma métrica de avaliação eficaz nos seguintes casos:

- Quando falsos positivos e falsos negativos são igualmente custosos;
- Quando o resultado não altera efetivamente ao adicionar mais dados;
- Quando verdadeiro negativo é alto [83].

Une precisão e *recall* afim de trazer um número único que determine a qualidade geral do modelo [84], conforme equação 3.9.

$$f1\ score = 2 \times \frac{precisao \times recall}{precisão + recall} \tag{3.9}$$

Área Sob a Curva (ROC-AUC)

A curva ROC, ou Característica de Operação do Receptor *(Receiver Operator Characteristic)*, é uma curva de probabilidade que plota a Taxa de Verdadeiro Positivo *(TPR-True Positive Rate)* contra a Taxa de Falso Positivo *(False Positive Rate-FPR)* em vários valores limite e separa o 'sinal' do 'ruído' [83].

A área sob a curva (AUC) é a medida da capacidade de um classificador para distinguir entre as classes. Quanto maior a AUC melhor é o desempenho do modelo em diferentes pontos de limiar entre classes positivas e negativas. Isso significa simplesmente que quando a AUC é igual a 1, o classificador é capaz de distinguir perfeitamente entre todos os pontos de classe positivos e negativos. Quando a AUC for igual a 0, o classificador estaria predizendo todos os Negativos como Positivos e vice-versa. Quando a AUC é 0,5, o classificador não é capaz de distinguir entre as classes Positivo e Negativo [83].

Matriz de confusão

A matriz de confusão é uma medida de desempenho para os problemas de classificação de aprendizado de máquina em que a saída pode ser duas ou mais classes. É uma tabela com combinações de valores previstos e reais [83], que indica os erros e acertos do seu modelo, comparando com o resultado esperado [84].

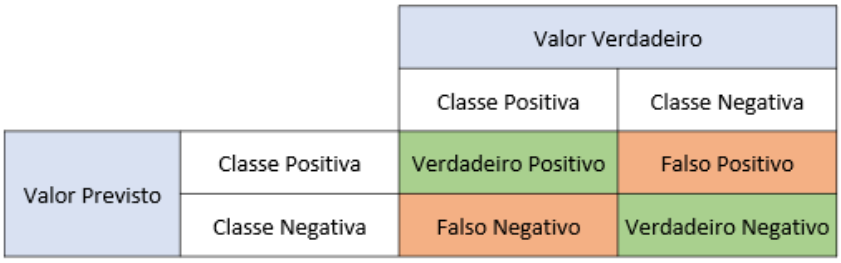

Figura 3.5: Matriz de confusão: visualização dos resultados obtidos em um modelo de classificação, relacionando os valores verdadeiros com os previstos.

Capítulo 4

Projeto

Esse capítulo apresenta o método utilizado para gerar o modelo de classificação de RNA telomerase, cujo objetivo é identificar, em fragmentos de genomas, presença ou não de RNA telomerase. Na Seção 4.1 é apresentado um fluxograma genérico do método, com as etapas utilizadas para a construção dos modelos. Na Seção 4.2, são descritos como foram obtidos os dados de entrada e em seguida como foram concatenados. Na Seção 4.3, é apresentado o método de extração de características, juntamente com as opções utilizadas. E finalmente, a Seção 4.4 apresenta como foram feitas as otimizações dos hiperparâmetros, os treinamentos e os testes para a geração dos modelos.

4.1 Método de geração de modelos

Nesta seção, é descrito o método genérico para classificar RNA telomerase, mostrado na Figura 4.1. O método tem duas etapas, sendo a primeira a "Extração de características" e a segunda o "Tuning de hiperparâmetros, treinamento e teste". Os dados de entrada são descritos na próxima seção.

4.2 Dados de entrada

Nesta seção, descreveremos com detalhes como foram extraídos e concatenados os dados a serem utilizados nas etapas de treinamento e teste para a geração dos modelos.

Sequências de DNA de RNA telomerase

Como arquivo de entrada para a etapa de "Extração de características", foi utilizado um arquivo *fasta* contendo 121 sequências de DNA de Telomerase RNA, identificadas em diferentes espécies de animais. Esse arquivo foi obtido banco de dados telomerase [15].

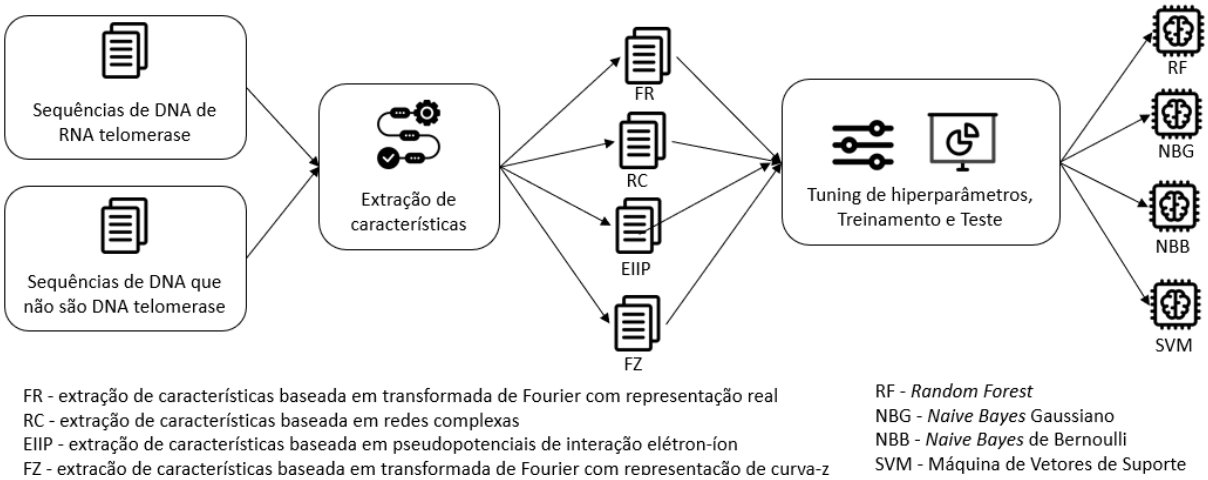

Figura 4.1: Fluxograma que representa o método. Inicialmente os conjuntos de dados positivo e negativo, obtidos dos bancos de dados, são passados por métodos de extração de características (*feature extraction*). Cada um dos arquivos obtidos por essa extração é utilizado para a criação de modelos de classificação de RNA telomerase, através de algoritmos de aprendizado de máquina.

Em média, as sequências de Telomerase RNA tinham 432 nucleotídeos, com um primeiro quartil de 370 e o terceiro quartil de 510 nucleotídeos. Esse arquivo e todos os demais do projeto encontram-se em repositório no github [85].

Sequências de DNA que não são de RNA telomerase

Para o conjunto negativo de dados, foram criados 10 arquivos, com sequências que não são de RNA telomerase. Foi utilizada a árvore filogenética mostrada na Figura 2.8 [6], que mostra a relação evolutiva de filos metazoários basais e classes com TR identificados. Animais de 8 filos diferentes foram selecionados, para que houvesse variabilidade, como segue, *Asterias rubens* (Echinodermata), *Homo sapiens* (Cordata), *Hydra vulgaris* (Cnidaria), *Lumbricus rubellus* (Nematoda), *Metaphire vulgaris* (Annelida), *Patella vulgata* (Mollusca), *Petrosia ficiformis* (Porifera), *Saccoglossus kowalevskii* (Hemicordata).

O genoma desses animais foi obtido no banco de genomas do *National Center for Biotechnology Information (NCBI)* [42]. Para cada genoma, foram extraídas 30 sequências aleatórias com tamanhos aleatórios entre 370 e 510. Esses valores foram determinados pela avaliação de tamanhos das sequências do conjunto positivo, sendo 370 o valor do primeiro quartil e 510 o do terceiro quartil.

A quantidade de 30 sequências foi determinada tomando como base a quantidade de sequências do conjunto positivo. Foram usadas 240 sequências provenientes de 8 animais, aproximadamente o dobro de 121, que é a quantidade de sequências do conjunto positivo. As 240 sequências foram concatenadas para formar o arquivo. Esse procedimento foi

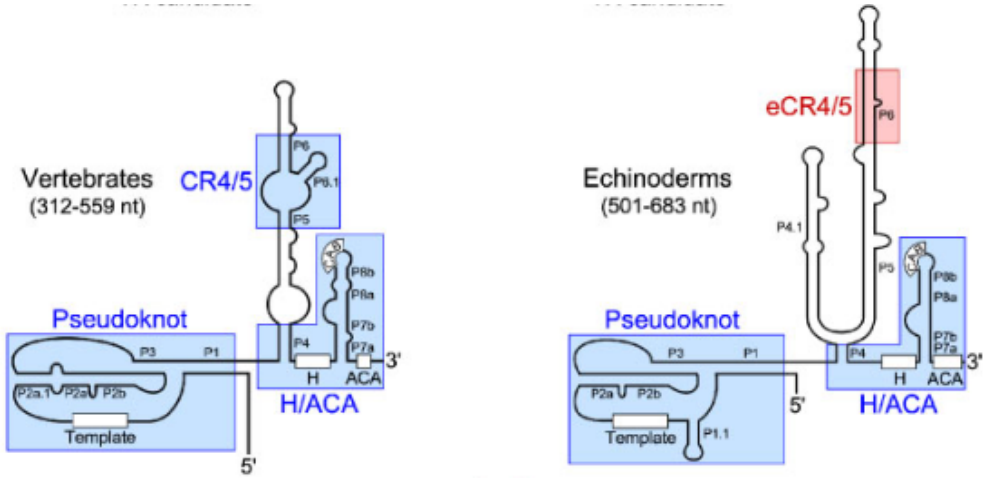

Figura 4.2: Comparação entre estruturas de telomerases de vertebrados e equinodermos. Dois domínios estruturais, T-PK e box H/ACA são conservados em ambos, enquanto o domínio CR4/5 dos vertebrados contém um stem loop P6.1, nos equinodermos ele é substituído por um domínio funcionalmente equivalente [6].

realizado 10 vezes, para que cada modelo fosse executado 10 vezes. Esse procedimento garantiu que as métricas obtidas pudessem ser avaliadas estatisticamente, dando maior robustez e confiabilidade aos resultados obtidos.

4.3 Extração de características

Seguindo o protocolo de extração de características indicado por Bonidia et al. [52], inicialmente os dados, tanto positivos quanto negativos, passaram por um filtro, para eliminar possíveis ruídos das sequências, tais como outras letras que não representam nucleotídeos, quais sejam, A, C, G, U ou T.

A próxima etapa foi a extração das características propriamente dito. Dentre os métodos, foram escolhidos:

1. Transformada de Fourier com representação real (FR);
2. Transformada de Fourier com representação de curva-z (FZ);
3. Transformada de Fourier com EIIP;
4. Redes Complexas (RC).

O método de transformada de Fourier foi escolhido por ser extensivamente utilizado em bioinformática, principalmente para análise de periodicidades e elementos repetitivos em sequências de DNA [52].

Na representação real, é usado um mapeamento baseado na propriedade de complemento do mapeamento complexo. Esse mapeamento aplica valores decimais negativos para as purinas (A, G) e valores decimais positivos para as pirimidinas (C, T) [52]. Esse método foi escolhido com a expectativa de que essa complementariedade pudesse colaborar na classificação.

O método da curva Z é um algoritmo utilizado em bioinformática para análise do genoma [46]. Esse método é capaz de codificar sequências de DNA com semântica mais biológica [52] por isso foi um dos escolhidos. E ainda, diferentes propriedades da curva Z, como sua simetria e periodicidade, podem fornecer informações únicas sobre a sequência de DNA [46].

O método de EIIP foi considerado devido a algumas características:

- A energia média dos elétrons deslocalizados do nucleotídeo é chamada de potencial de interação elétron-íon. Portanto, o EIIP de um nucleotídeo é uma quantidade física.
- É biologicamente mais significativo, pois representa uma propriedade física quando comparado aos valores indicadores que representam apenas a presença ou ausência de um nucleotídeo.
- Envolve apenas uma única sequência em vez de quatro no caso da representação de Voss, reduzindo assim a sobrecarga computacional em 75%.
- Na literatura, os valores de EIIP foram aplicados com sucesso para análise de DNA em vários estudos.
- Os valores EIIP foram divulgados para fornecer o mapeamento mais adequado para análise espectral de sequências de DNA [86].

O método de redes complexas foi escolhido porque leva em consideração as forças intermoleculares entre nucleotídeos vizinhos, característica importante na avaliação da estrutura espacial da molécula [52]. Conforme mencionado no Capítulo 2, essa é uma característica importante na identificação da telomerase [6].

Posteriormente foi feita a exclusão das *strings* que davam nome aos dados, para que ficassem somente números no conjunto de dados. Para determinar, no conjunto de dados, quais dados eram positivos e quais eram negativos, foram utilizados números 1 para dados positivos e 0 para dados negativos.

Foi feita também a normalização dos dados, para usar uma escala comum, sem distorcer as diferenças nos intervalos de valores nem perder informações. Foram redimensionados em um intervalo adequado entre 0 e 1. Conforme mencionado no Capítulo 3,

a normalização refere-se à centralização dos dados, redimensionando-os em um intervalo adequado [61, 63]

4.4 *Tuning* de hiperparâmetros, treinamento e teste

Foram feitas as etapas de *tuning* de hiperparâmetros, treinamento e teste para quatro algoritmos, com o objetivo de obter modelos de classificação diferentes, para serem avaliados quanto aos melhores resultados, sugerindo os melhores hiperparâmetros.

O *tuning* de hiperparâmetros foi feito com otimização randômica de hiperparâmetros, com validação cruzada, com a utilização da ferramenta de *Randomized Search Cross Validation*, da biblioteca *Scikitlearn* [87], biblioteca de aprendizado de máquina, de código aberto para a linguagem de programação Python [88]. A *Randomized Search Cross Validation* executa testes com variação randômica dos hiperparâmetros, dentro de uma lista de possibilidades indicada. Na prática, a utilização do *RandomizedSearchCV* permite encontrar valores muito próximos aos que mais otimizarão nossos estimadores, sem que seja necessário explorar todo o espaço de parâmetros [89]. Para cada algoritmo foi determinado um grupo de hiperparâmetros pertinente, descritos nas Subseções 4.4.1, 4.4.2, 4.4.3 e 4.4.4.

A própria ferramenta de busca randômica, *Randomized Search Cross Validation*, possui parâmetros de utilização, que guiam a maneira de uso. Foram utilizados:

- *param_distributions*: as opções possíveis de hiperparâmetros, conforme descritos nas subseções 4.4.1, 4.4.2, 4.4.3 e 4.4.4.
- $cv = 5$: Determina a estratégia de divisão de validação cruzada [87]. Nesse caso, validação cruzada de 5 *folds*.
- $verbose = 2$: Controla a verbosidade, os tipos e quantidades de mensagens a serem exibidas na execução [87].
- $n\_jobs = 4$: Número de tarefas a serem executadas em paralelo, refere-se ao uso dos processadores [87].
- $n\_iter = 50$: Número de configurações de parâmetro que são amostradas. Ele balanceia tempo de execução versus qualidade da solução [87].
- *scoring* = 'f1': Estratégia para avaliar o desempenho do modelo de validação cruzada no conjunto de teste [87]. Nesse caso foi escolhido o *f1-score*, pois essa métrica evidencia uma maior quantidade de verdadeiros positivos, o que é mais adequado para esse trabalho.

4.4.1 *Random Forest*

Como opções de hiperparâmetros para *Random Forest* foram utilizados:

- Número de árvores na floresta (*n_estimators*): variando entre 10 e 5000, com 10 opções;
- O número de features a serem considerados ao procurar a melhor divisão (*max_features*): *None*, sqrt, log2;
- Profundidade máxima das árvores (*max_depth*): 2 ou 4;
- Número mínimo de amostras necessárias para dividir um nó interno (*min_samples_split*): 2 ou 5;
- O número mínimo de amostras necessárias para estar em um nó folha (*min_samples_leaf*): 1 ou 2;
- Se as amostras de autoinicialização serão usadas na construção de árvores (*bootstrap*): verdadeiro ou falso;
- Critérios de partição (*criterion*): gini, entropia ou *log loss*;
- Se for fornecido um dicionário, as chaves são classes e os valores são pesos de classe correspondentes (*class_weight*): *balanced, balanced subsample* ou *None*.

4.4.2 *Naive Bayes* Gaussiano

No *Naive Bayes* Gaussiano, foi utilizada a *'var_smoothing'* como hiperparâmetro a se otimizar. Uma curva Gaussiana pode servir como um filtro, permitindo apenas as amostras próximas a ela. No contexto de *Naive Bayes*, assumir uma distribuição gaussiana é, essencialmente, dar mais pesos às amostras mais próximas da média da distribuição. Isso pode ou não ser apropriado, dependendo se o que se deseja prever segue uma distribuição normal. A variável *var_smoothing*, adiciona artificialmente um valor definido pelo usuário à variância da distribuição, cujo valor padrão é derivado do conjunto de dados de treinamento. Isso essencialmente amplia ou suaviza a curva e contabiliza mais amostras que estão mais distantes da média de distribuição. [87].

var_smoothing: *np.logspace*(0, -9, num = 100), ou seja, 100 números igualmente espaçados numa escala log, de 0 a -9.

4.4.3 *Naive Bayes* de Bernoulli

Como opções de hiperparâmetros para *Naive Bayes* de Bernoulli foram utilizados:

- *alpha*: 0 ou 1. Parâmetro de suavização aditivo (Laplace/Lidstone), técnica de suavização que ajuda a resolver o problema de probabilidade zero no algoritmo de aprendizado de máquina *Naive Bayes* [87].
- *force_alpha*: *True* ou *False*. Se *False* e alfa for menor que 1e-10, ele definirá alfa como 1e-10. Se *True*, alfa permanecerá inalterado.
- *binarize*: 0.0, 0.5 ou 1.0. Limite para binarização (mapeamento para booleanos) de recursos de amostra [87].
- *fit_prior*: *True* ou *False*. Se deve aprender as probabilidades anteriores da classe ou não [87].

4.4.4 Máquina de Vetores de Suporte

No método Máquina de Vetores de Suporte foram utilizadas as seguintes opções de hiperparâmetros:

- C: np.linspace(start = 0.001, stop = 1000, num = 20), ou seja, 20 números uniformemente espaçados em um intervalo de 0,001 e 1000. C é o parâmetro de penalidade, que representa erro de c classificação ou termo de erro. A classificação incorreta ou termo de erro informa à otimização da SVM quanto erro é suportável. É assim que você pode controlar a compensação entre o limite de decisão e o termo de classificação incorreta [87, 90].
- gamma: 1e-2, 1e-3, 1e-4, 1e-5. Define a distância de influência de um único ponto de treinamento [87, 90].
- kernel: *linear*, *sigmoid* ou *rbf*. A principal função do kernel é pegar espaço de entrada de baixa dimensão e transformá-lo em um espaço de dimensão superior, em caso de separação não linear [87, 90].

4.4.5 Treinamento e teste

Em todos os casos, a divisão treinamento-teste foi feita com os parâmetros "Proporção treinamento-teste de 80%-20% respectivamente" e "estratificação ativada". A estratificação visa preservar as mesmas proporções de exemplos em cada classe conforme observado no conjunto de dados original.

Com as obtenção dos modelos, algumas métricas foram utilizadas para verificar como estavam sendo classificados os dados e com qual eficiência. Foram utilizadas como métricas a acurácia, a precisão, a área sob a curva (AUC), o *F1-score*, além da matriz de confusão, que explicita a classificação (verdadeiros e falsos negativos, verdadeiros e falsos positivos).

Capítulo 5

Resultados e discussão

Nesse capítulo são discutidos os resultados obtidos a partir do método proposto. Na Seção 5.1 são apresentados os resultados obtidos a partir dos modelos baseados em *Random Forest*. Na Seção 5.2 são apresentados os resultados obtidos a partir dos modelos baseados em *Naive Bayes Gaussiano*. Na Seção 5.3 são apresentados os resultados obtidos a partir dos modelos baseados em *Naive Bayes de Bernoulli*. Na Seção 5.4, são apresentados os resultados obtidos a partir dos modelos baseados em Máquina de Vetores de Suporte. Por fim, na Seção 5.5, comparamos os resultados obtidos a partir dos diferentes modelos, para classificar RNA telomerase.

Após a coleta de dados, o pré-processamento para limpar ruídos, a extração de características e a geração de modelos através de algoritmos de aprendizado de máquina, utilizamos as métricas acurácia, precisão, *recall*, *F1 score*, área sob a curva e matriz de confusão foram calculadas para avaliar a *performance* de cada um dos modelos utilizados, conforme descrito no Capítulo 3. Os resultados completos obtidos estão descritos no Anexo I.

5.1 *Random Forest*

Nesta seção, são apresentados os resultados obtidos a partir dos modelos baseados em *Random Forest*. Neste experimento, o objetivo foi identificar, em amostras de sequenciamentos de DNA, quais apresentavam DNA relativo a RNA Telomerase. O mais interessante neste caso é saber quais são os verdadeiros positivos e evitar os falsos negativos. Por isso, as métricas que mais evidenciam essas características são a *F1 score* e a *recall*.

Na Tabela 5.1, pode-se observar que os métodos de extração de característica que geraram os melhores resultados, entre os modelos baseados em *Random Forest*, foram Transformada de Fourier com Representação Real e Pseudopotenciais de interação elétron-íon, seguidos de Redes Complexas. O método de curva-z teve um desempenho insatisfatório.

Verifica-se que, para todos eles, o desvio padrão foi pequeno, indicando consistência do modelo.

Tabela 5.1: Médias e desvios padrão dos resultados obtidos nos modelos de *Random Forest*, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z.

	FR	RC	EIIP	FZ
Acurácia média	0.830	0.790	0.787	0.711
desvio padrão	0.032	0.031	0.029	0.032
Precisão média	0.718	0.788	0.781	0.678
desvio padrão	0.048	0.125	0.075	0.163
Recall média	0.800	0.517	0.787	0.254
desvio padrão	0.078	0.063	0.082	0.060
AUC média	0.822	0.721	0.837	0.595
desvio padrão	0.040	0.033	0.044	0.034
F1 score média	0.755	0.618	0.781	0.364
desvio padrão	0.050	0.053	0.057	0.075

5.2 *Naive Bayes* Gaussiano

Na Tabela 5.2, com os dados referentes aos modelos gerados com *Naive Bayes* Gaussiano, pode ser observado através das métricas de *F1 score* e *recall*, que o método de extração de características que gerou o melhor modelo foi o da Transformada de Fourier com Representação Real, que também apresentou pequeno desvio padrão, indicando consistência do modelo. Os demais modelos também apresentaram consistência, evidenciada pelos desvios padrão pequenos. Mas não tiveram resultados satisfatórios, conforme observado na Tabela 5.2

5.3 *Naive Bayes* de Bernoulli

Os dados referentes aos modelos gerados com *Naive Bayes* de Bernoulli, representados na Tabela 5.3, mostram o método de extração de características de Redes Complexas conseguiu algum desempenho, no entanto, insuficiente. Os demais modelos, não conseguiram classificar, o que foi evidenciado ao verificar a matriz de confusão deles. Eles somente classificaram todas as sequências como negativas, gerando somente Verdadeiros Negativos e Falsos Negativos.

Tabela 5.2: Médias e desvios padrão dos resultados obtidos nos modelos de *Naive Bayes* Gaussiano, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z.

	FR	RC	EIIP	FZ
Acurácia média	0.790	0.776	0.776	0.682
desvio padrão	0.023	0.043	0.043	0.015
Precisão média	0.716	0.419	0.674	0.690
desvio padrão	0.060	0.057	0.126	0.244
Recall média	0.613	0.596	0.183	0.079
desvio padrão	0.044	0.164	0.022	0.013
AUC média	0.747	0.585	0.571	0.528
desvio padrão	0.021	0.036	0.013	0.014
F1 score média	0.658	0.479	0.287	0.141
desvio padrão	0.031	0.044	0.033	0.024

Tabela 5.3: Médias e desvios padrão dos resultados obtidos nos modelos de *Naive Bayes* de Bernoulli, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z.

	FR	RC	EIIP	FZ
Acurácia média	0.671	0.778	0.775	0.671
desvio padrão	0.000	0.040	0.047	0.000
Precisão média	0.000	0.555	0.000	0.000
desvio padrão	0.000	0.156	0.000	0.000
Recall média	0.000	0.388	0.000	0.000
desvio padrão	0.000	0.065	0.000	0.000
AUC média	0.500	0.621	0.500	0.500
desvio padrão	0.000	0.038	0.000	0.000
F1 score média	0.000	0.451	0.000	0.000
desvio padrão	0.000	0.083	0.000	0.000

5.4 Máquina de Vetores de Suporte

Os modelos gerados pela Máquina de Vetores de Suporte também apresentaram consistência, evidenciada pelos desvios padrão pequenos. Mas não tiveram resultados satisfatórios, conforme observado na Tabela 5.4. Os modelos baseados na extração de características da Transformada de Fourier e da Curva-z somente classificaram todas as sequências como negativas, gerando somente Verdadeiros Negativos e Falsos Negativos. E os modelos baseados na extração de características de Redes Complexas e Pseudopotenciais de interação elétron-íon (EIIP), apesar de conseguirem alguma classificação, não foi o suficiente para um bom desempenho.

Tabela 5.4: Médias e desvios padrão dos resultados obtidos nos modelos de Máquinas de Vetores de Suporte, com diferentes métodos de Extração de características, onde FR é Transformada de Fourier com Representação Real; RC é Redes Complexas; EIIP é Pseudopotenciais de interação elétron-íon; FZ é Transformada de Fourier com curva-z.

	FR	RC	EIIP	FZ
Acurácia média	0.671	0.782	0.776	0.671
desvio padrão	0.000	0.033	0.043	0.000
Precisão média	0.000	0.655	0.877	0.000
desvio padrão	0.000	0.094	0.165	0.000
Recall média	0.000	0.317	0.100	0.000
desvio padrão	0.000	0.029	0.022	0.000
AUC média	0.500	0.615	0.545	0.500
desvio padrão	0.000	0.022	0.009	0.000
F1 score média	0.000	0.425	0.178	0.000
desvio padrão	0.000	0.036	0.032	0.000

5.5 Observações gerais

A partir dos resultados descritos nas seções anteriores, avalia-se que os melhores modelos, entre todos os dezesseis, para classificar RNA telomerase foram baseados em *Random Forest (RF)* com Transformada de Fourier com Representação Real (RF FR), com Redes Complexas (RF RC) e com Pseudopotenciais de interação elétron-íon (RF EIIP); e aqueles baseados em *Naive Bayes* Gaussiano (NBG) com Transformada de Fourier com Representação Real (NBG FR).

Ao analisar somente esses quatro melhores modelos, utilizando suas matrizes de confusão, é possível observar que os modelos de RF RC e NBG FR são os que tiveram resultados menos satisfatórios, principalmente entre os verdadeiros positivos, que são de maior interesse para esse estudo.

E finalmente, observando os resultados dos modelos de RF FR e RF EIIP, verifica-se que os dados são bastante similares e não é possível indicar qual deles apresentarou os melhores resultados.

Tabela 5.5: Comparação das matrizes de confusão dos melhores modelos obtidos, onde RF FR refere-se ao modelo obtido com *Random Forest* e transformada de Fourier com representação real, RF RC refere-se ao modelo obtido com *Random Forest* e Redes Complexas, RF EIIP refere-se ao modelo obtido com *Random Forest* e método de extração de EIIP, NBG FR refere-se ao modelo obtido com *Naive Bayes* Gaussiano com transformada de Fourier e representação real.

	RF FR média	**RF RC** média	**RF EIIP** média	**NBG FR** média
Verdadeiros Positivos (VP)	19	12	19	15
Verdadeiros Negativos (VN)	41	45	43	43
Falsos Positivos (FP)	8	4	6	6
Falsos Negativos (FN)	5	12	5	9

Capítulo 6

Conclusões e trabalhos futuros

Neste trabalho, propusemos um método baseado em aprendizado de máquina supervisionado para classificar RNA telomerase. Em particular, inicialmente criamos um um repositório com sequências de RNA telomerase, a partir de informações coletadas no banco de dados *The Telomerase Database* [15], que pode ser encontrado em um repositório de github [85]. Em seguida, utilizamos técnicas de extração de características para identificar características (em inglês, *features*), a serem usadas em métodos de aprendizado de máquina. Por fim, usamos técnicas de aprendizado de máquina supervisionado (*Random Forest*, *Naive Bayes* Gaussiano, *Naive Bayes* de Bernoulli e Máquina de Vetor de Suporte) para classificar sequências de RNA telomerase.

Dos resultados obtidos com a aplicação do método, verificou-se que é possível criar modelos baseados em técnicas de aprendizado de máquina para classificar RNA telomerase. Os melhores resultados foram obtidos a partir dos modelos de *Random Forest* com Transformada de Fourier com Representação Real e *Random Forest* com método de extração de características baseado em Pseudopotenciais de interação elétron-íon.

Como o método apenas classifica trechos de DNA quanto à presença de RNA telomerase ou não, um possível trabalho futuro seria interessante na direção da identificação de RNA telomerase em genomas. Além disso, a busca de novos TERCs usando o modelo gerado.

Mais um trabalho possível é uma abordagem híbrida, levando em consideração algumas características (*features*) biológicas.

Outra direção de pesquisa interessante seria com a utilização de aprendizado semi-supervisionado e com a utilização de aprendizado de transferência (*transferlearning*).[1]

[1] O aprendizado de transferência é um problema de pesquisa em aprendizado de máquina que se concentra em armazenar o conhecimento adquirido ao resolver um problema e aplicá-lo a um problema diferente, mas relacionado [91].

E ainda, outro trabalho seria expandir o conjunto de treinamento com adição de trechos de DNA de RNA telomerase de outros reinos, como plantas e fungos. Por fim, o uso de outros algoritmos de aprendizado de máquina, tais como Redes Neurais, K-vizinhos mais próximos, do inglês *K-nearest neighbor* (KNN), Regressão Linear e Regressão Logistica, para geração de modelos classificadores de telomerase pode ser uma outra opção interessante de pesquisa.

Referências

[1] Bruice, P. Y.: *Química Orgânica - Vol. 2.* Pearson, 2006. ix, 6

[2] McMurry, J.: *Química Orgânica - vol. 2.* Thompson, 2011. ix, 6

[3] P., Snustad D. e M. J. Simmons: *Fundamentos de genética.* Guanabara Koogan, 2017. ix, 7, 8, 9

[4] Alberts, B., A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter, J. Wilson e T. Hunt: *Biologia Molecular da Célula.* Artmed, 2017. ix, 8, 9, 10, 12, 13

[5] Goodsell, D.: *Molecule of the month: Telomerase.* The RCSB PDB Molecule of the Month, 2018. ix, 14

[6] Logeswaran, D., Y. Li, J. Podlevsky e J. Chen: *Monophyletic Origin and Divergent Evolution of Animal Telomerase RNA.* Molecular biology and evolution, 38, agosto 2020. ix, x, 13, 14, 15, 16, 32, 33, 34

[7] Chauhan, A.: *Random Forest Classifier and its Hyperparameters*, 2021. `https://medium.com/analytics-vidhya/random-forest-classifier-and_its-hyperparameters-8467bec755f6`, acesso em 2023-02-01. ix, 25

[8] Starmer, J.: *Gaussian Naive Bayes, Clearly Explained*, 2020. `https://www.youtube.com/watch?v=H3EjCKtlVog`, acesso em 2023-02-11. ix, 26

[9] Lang, N.: *Support Vector Machine (SVM) – easily explained!*, 2023. `https://databasecamp.de/en/ml/svm-explained`, acesso em 2023-02-11. ix, 27

[10] Yang, J. X., R. H. Rastetter e D. Wilhelm: *Non-coding RNAs: an introduction.* Non-coding RNA and the Reproductive System, páginas 13–32, 2016. 1

[11] Eddy, S. R.: *Non-coding RNA genes and the modern RNA world.* Nature Reviews Genetics, 2(12):919–929, 2001. 1

[12] Giardini, M. A., M. Segatto, M. S. da Silva, V. S. Nunes e M. I. N. Cano: *Chapter One - Telomere and Telomerase Biology.* Em Calado, R. T (editor): *Telomeres in Health and Disease*, volume 125 de *Progress in Molecular Biology and Translational Science*, páginas 1–40. Academic Press, 2014. `https://www.sciencedirect.com/science/article/pii/B9780123978981000013`. 1, 12, 13

[13] Dew-Budd, K., J. Cheung, K. Palos, E. Forsythe e M. Beilstein: *Evolutionary and biochemical analyses reveal conservation of the Brassicaceae telomerase ribonucleoprotein complex.* PLOS ONE, 15:e0222687, abril 2020. 1, 13

[14] Waldl, M., M. C. Thiel, R. Ochsenreiter, A. Holzenleiter, J. V. A. Oliveira, M. E. M. T. Walter, M. T. Wolfinger e P. F. Stadler: *TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes*. Genes, 9(8), 2018, ISSN 2073-4425. `https://www.mdpi.com/2073-4425/9/8/372`. 1, 3, 15, 16

[15] Podlevsky, J. D., C. J. Bley, R. V. Omana, X. Qi e J. J. Chen: *The telomerase database*. Nucleic Acids Res, 36:D339–D343, 2008, ISSN 1362-4962. 2, 3, 11, 31, 43

[16] Petrova, O. A., A. B. Mantsyzov, E. V. Rodina, S. V. Efimov, J. Hackenberg, C.and Hakanpää, V. V. Klochkov, A. A. Lebedev, A. A. Chugunova, Alexander N. Malyavko, T. S. Zatsepin, A. V. Mishin, M. I. Zvereva, V. S. Lamzin, O. A. Dontsova e V. I. Polshakov: *Structure and function of the N-terminal domain of the yeast telomerase reverse transcriptase*. Nucleic Acids Research, 46(3):1525–1540, dezembro 2017, ISSN 0305-1048. `https://doi.org/10.1093/nar/gkx1275`. 2

[17] Schmidt, J. C. e T. R. T. R. Cech: *Human telomerase: biogenesis, trafficking, recruitment, and activation*. Genes Dev, 29(11):1095–105, 2015, ISSN 1549-5477. 2

[18] Larranaga, P., B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, José A Lozano, Rubén Armananzas, Guzmán Santafé, Aritz Pérez *et al.*: *Machine learning in bioinformatics*. Briefings in bioinformatics, 7(1):86–112, 2006. 2

[19] Ezziane, Z.: *Applications of artificial intelligence in bioinformatics: A review*. Expert Systems with Applications, 30(1):2–10, 2006. 2

[20] Auslander, N., A. B. Gussow e E. V. Koonin: *Incorporating Machine Learning into Established Bioinformatics Frameworks*. International Journal of Molecular Sciences, 22(6), 2021, ISSN 1422-0067. `https://www.mdpi.com/1422-0067/22/6/2903`. 2, 3, 17, 18, 21

[21] Kaplan, A.: *Artificial Intelligence, Business and Civilization: Our Fate Made in Machines*. Routledge, 2022. 2

[22] Mathé, C., M. F. Sagot, T. Schiex e P. Rouzé: *Current methods of gene prediction, their strengths and weaknesses*. Nucleic acids research, 30(19):4103–4117, 2002. 2

[23] Bonidia, R. P., L. D. H. Sampaio, F. M. Lopes, A. C. P. L. F. de Carvalho e D. S. Sanches: *Feature Extraction Approaches for Biological Sequences: A Comparative Study of Mathematical Models*. bioRxiv, 2020. `https://www.biorxiv.org/content/early/2020/06/09/2020.06.08.140368`. 2, 17, 18, 24

[24] Miracco, E. J., J. Jiang, D. D. Cash e J. Feigon: *Progress in structural studies of telomerase*. Current opinion in structural biology, 24:115–124, 2014. 3

[25] Barreiros, A. L. B. S. e M. L. Barreiros: *Química de Biomoléculas*. Universidade Federal de Sergipe / CESAD, 2012. 5, 6

[26] Zaha, A., H. B. Ferreira e L. M. P. Passaglia: *Biologia Molecular Básica*. Artmed, 2014. 5

[27] Setubal, J. e J. Meidanis: *Introduction to computational molecular biology*. PWS, 1997. 5, 6, 7, 8, 10

[28] Machado-Lima, A., H. A. del Portillo e A. M. Durham: *Computational methods in noncoding RNA research*. J Math Biol, 56(1-2):14–49, 2008. 7, 8, 14

[29] Crick, F.: *Central dogma of molecular biology*. Nature, 227(5258):561–3, 1970. 8, 10

[30] Moreira, C.: *Dogma Central da Biologia*. Revista de Ciência Elementar, 3(1):055, 2015. 8

[31] Lesk, A.: *Introduction to bioinformatics*. Oxford University Press, 2019. 10

[32] Goes, A. C. de Souza e B. V. X. Oliveira: *Projeto Genoma Humano: um retrato da construção do conhecimento científico sob a ótica da revista Ciência Hoje*. Ciência & Educação (Bauru), 20:561–577, 2014. 10

[33] Olson, M. V.: *The human genome project*. Proceedings of the National Academy of Sciences, 90(10):4338–4344, 1993. 10

[34] Srivastava, A., G. J. Joshy e K. M. K. Radha: *Transcriptome analysis*. Em Ranganathan, S., M. Gribskov, K. Nakai e C. Schönbach (editores): *Encyclopedia of Bioinformatics and Computational Biology*, páginas 792–805. Academic Press, Oxford, 2019, ISBN 978-0-12-811432-2. https://www.sciencedirect.com/science/article/pii/B9780128096338201611. 10

[35] LaRossa, R. A.: *Transcriptome*. Em M., Stanley e Kelly H. (editores): *Brenner's Encyclopedia of Genetics (Second Edition)*, páginas 101–103. Academic Press, San Diego, second edition edição, 2013, ISBN 978-0-08-096156-9. https://www.sciencedirect.com/science/article/pii/B9780123749840015539. 10

[36] Cho, W. C. S.: *Proteomics Technologies and Challenges*. Genomics, Proteomics & Bioinformatics, 5(2):77–85, 2007, ISSN 1672-0229. https://www.sciencedirect.com/science/article/pii/S1672022907600187. 10

[37] Aslam, B., M. Basit, M. A. Nisar, M. Khurshid e M. H. Rasool: *Proteomics: Technologies and Their Applications*. Journal of Chromatographic Science, 55(2):182–196, janeiro 2017, ISSN 0021-9665. https://doi.org/10.1093/chromsci/bmw167. 10

[38] Reuter, J. A., D. V. Spacek e M. P. Snyder: *High-Throughput Sequencing Technologies*. Molecular Cell, 58(4):586–597, 2015, ISSN 1097-2765. https://www.sciencedirect.com/science/article/pii/S1097276515003408. 11

[39] Qin, D.: *Next-generation sequencing and its clinical application*. Cancer biology & medicine, 16(1):4, 2019. 11

[40] Roy, S., C. Coldren, A. Karunamurthy, N. S. S. Kip, E. W. Klee, S. E. Lincoln, A. Leon, M. Pullambhatla, Temple Smolkin R. L., K. V. Voelkerding, C. Wang e A. B. Carter: *Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists*. The Journal of Molecular

Diagnostics, 20(1):4–27, 2018, ISSN 1525-1578. https://www.sciencedirect.com/science/article/pii/S1525157817303732. 11

[41] Libório, L. e V. H. Resende: *Introdução aos bancos de dados biológicos*. BIOINFO Revista Brasileira de Bioinformática, 2021. 11

[42] Medicine, National Library of: *National Center for Biotechnology Information*, 2023. https://www.ncbi.nlm.nih.gov/, acesso em 2023-02-08. 11, 32

[43] Rye, C., R. Wise, V. Jurukovski, J. DeSaix, J. Choi e Y. Avissar: *Biology*. OpenStax, 2016. 12

[44] Zvereva, M. I., D. M. Shcherbakova e O. A. Dontsova: *Telomerase: structure, functions, and activity regulation*. Biochemistry (Mosc), 75(13):1563–83, 2010. 13

[45] Shay, J. e W. Wright: *Telomeres and telomerase: three decades of progress*. Nature Reviews Genetics, 20, fevereiro 2019. 13

[46] Zhang, Q., N. Q. Kim e Feigon J.: *Architecture of human telomerase RNA*. Proceedings of the National Academy of Sciences, 108(51):20325–20332, 2011. https://www.pnas.org/doi/abs/10.1073/pnas.1100279108. 13, 14, 23, 34

[47] Förstemann, K. e J. Lingner: *Telomerase limits the extent of base pairing between template RNA and telomeric DNA*. EMBO reports, 6(4):361–366, 2005. 13

[48] Musgrove, C., L. I. Jansson e M. D. Stone: *New perspectives on telomerase RNA structure and function*. Wiley Interdiscip Rev RNA, 9(2):e1456, 2018. 13

[49] Jády, B. E., P. Richard, E. Bertrand e T. Kiss: *Cell Cycle-dependent Recruitment of Telomerase RNA and Cajal Bodies to Human Telomeres*. Molecular Biology of the Cell, 17(2):944–954, 2006. https://doi.org/10.1091/mbc.e05-09-0904, PMID: 16319170. 13

[50] Feng, J., W. D. Funk, S. S. Wang, S. L. Weinrich, A. A. Avilion, C. P. Chiu, R. R. Adams, E. Chang, R. C. Allsopp, J. Yu, S. Le, M. D. West, C. B. Harley, W. H. Andrews, C. W. Greider e B. Villeponteau: *The RNA Component of Human Telomerase*. Science, 269(5228):1236–1241, 1995. https://www.science.org/doi/abs/10.1126/science.7544491. 14

[51] Meyer, M., T. Munzner e H. Pfister: *MizBee: a multiscale synteny browser*. IEEE transactions on visualization and computer graphics, 15(6):897–904, 2009. 16

[52] Bonidia, R. P., L. D. H. Sampaio, D. S. Domingues, A. R. Paschoal, F. M. Lopes, A. C. P. L. F. de Carvalho e D. S. Sanches: *Feature extraction approaches for biological sequences: a comparative study of mathematical features*. Briefings in Bioinformatics, fevereiro 2021, ISSN 1477-4054. https://doi.org/10.1093/bib/bbab011, bbab011. 16, 21, 22, 23, 24, 33, 34

[53] Burkov, A.: *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2022. 17

[54] Deisenroth, M. P., A. A. Faisal e C. S. Ong: *Mathematics for Machine Learning*. Cambridge University Press, 2022. 17

[55] Awad, M. e R. Khanna: *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Springer nature, 2015. 18, 19

[56] Theobald, O.: *Machine learning for absolute beginners: a plain English introduction*, volume 157. Scatterplot press London, UK, 2017. 20

[57] Bischl, B., M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker e A. L. et al. Boulesteix: *Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, página e1484, 2021. 20

[58] Claesen, M. e B. De Moor: *Hyperparameter search in machine learning*. arXiv preprint arXiv:1502.02127, 2015. 21

[59] Andonie, R.: *Hyperparameter optimization in learning systems*. Journal of Membrane Computing, 1(4):279–291, 2019. 21

[60] Kohavi, R. *et al.*: *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Em *Ijcai*, volume 14, páginas 1137–1145. Montreal, Canada, 1995. 21

[61] Storcheus, D., A. Rostamizadeh e S. Kumar: *A Survey of Modern Questions and Challenges in Feature Extraction*. Em Storcheus, Dmitry, Afshin Rostamizadeh e Sanjiv Kumar (editores): *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 de *Proceedings of Machine Learning Research*, páginas 1–18, Montreal, Canada, 11 Dec 2015. PMLR. `https://proceedings.mlr.press/v44/storcheus2015survey.html`. 21, 22, 35

[62] Bonidia, R. P., D. S. Domingues, D. S. Sanches e A. C. P. L. F. de Carvalho: *MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors*. Briefings in Bioinformatics, 23(1), novembro 2021, ISSN 1477-4054. `https://doi.org/10.1093/bib/bbab434`, bbab434. 21

[63] Ali, P. J. M., R. H. Faraj, E. Koya, Peshawa J. M. Ali e R. H. Faraj: *Data normalization and standardization: a technical report*. Mach Learn Tech Rep, 1(1):1–6, 2014. 21, 22, 35

[64] Bhandari, A.: *Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization*, 2020. `https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/`, acesso em 2023-01-31. 22

[65] Vinga, S.: *Information theory applications for biological sequence analysis*. Briefings in Bioinformatics, 15(3):376–389, setembro 2013, ISSN 1467-5463. `https://doi.org/10.1093/bib/bbt068`. 22, 23

[66] M., Murugaiah e M. Ganesan: *A Novel Frequency Based Feature Extraction Technique for Classification of Corona Virus Genome and Discovery of COVID-19 Repeat Pattern*. Brazilian Archives of Biology and Technology, 64, 2021, ISSN 1678-4324. https://www.scielo.br/j/babt/a/fccQdMKwxND6g6CnRXBgC6F/?lang=en. 22

[67] Yin, c., Y. Chen e S. S. T. Yau: *A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering*. Journal of Theoretical Biology, 359:18–28, 2014, ISSN 0022-5193. https://www.sciencedirect.com/science/article/pii/S0022519314003324. 22

[68] Voss, R. F.: *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences*. Phys. Rev. Lett., 68:3805–3808, Jun 1992. https://link.aps.org/doi/10.1103/PhysRevLett.68.3805. 22

[69] McNaught, A. D. e A. et al. Wilkinson: *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford, 1997. 23

[70] Achuthsankar, N. S. e S. P. Sreenadhan: *A coding measure scheme employing electron-ion interaction pseudopotential (eiip)*. Bioinformation, 1(6):197, 2006. 23

[71] Costa, L. da F., F. A. Rodrigues e A. S. Cristino: *Complex networks: the key to systems biology*. Genetics and Molecular Biology, 31:591–601, 2008. 24

[72] Ito, E. A., I. Katahira, F. F. R. Vicente, L. F. P. Pereira e F. M. Lopes: *BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification*. Nucleic Acids Research, 46(16):e96–e96, junho 2018, ISSN 0305-1048. https://doi.org/10.1093/nar/gky462. 24

[73] Chen, X. e H. Ishwaran: *Random forests for genomic data analysis*. Genomics, 99(6):323–329, 2012, ISSN 0888-7543. https://www.sciencedirect.com/science/article/pii/S0888754312000626. 24

[74] Probst, P., M. N. Wright e A. L. Boulesteix: *Hyperparameters and tuning strategies for random forest*. WIREs Data Mining and Knowledge Discovery, 9(3):e1301, 2019. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1301. 24

[75] Webb, G.y I., E. Keogh e R. Miikkulainen: *Naive Bayes*. Encyclopedia of machine learning, 15:713–714, 2010. 25

[76] Rish, I. et al.: *An empirical study of the Naive Bayes classifier*. Em *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, páginas 41–46, 2001. 25

[77] Lou, W., X. Wang, F. Chen, Y. Chen, B. Jiang e H. Zhang: *Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes*. PloS one, 9(1):e86703, 2014. 26

[78] Artur, M.: *Review the performance of the Bernoulli Naive Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features*. Procedia Computer Science, 190:564–570, 2021. 26

[79] Bafjaish, S. S.: *Comparative analysis of Naive Bayesian techniques in health-related for classification task*. Journal of Soft Computing and Data Mining, 1(2):1–10, 2020. 26

[80] Kharwal, A.: *Bernoulli Naive Bayes in Machine Learning*, 2021. https://thecleverprogrammer.com/2021/07/27/bernoulli-naive-bayes-in-machine-learning/, acesso em 2023-02-11. 26

[81] Lorena, A. C. e A. C. P. L. F. Carvalho: *Uma introdução às support vector machines*. Revista de Informática Teórica e Aplicada, 14(2):43–67, 2007. 26

[82] Wang, H. e D. Hu: *Comparison of SVM and LS-SVM for regression*. Em *2005 International conference on neural networks and brain*, volume 1, páginas 279–283. IEEE, 2005. 26

[83] Agrawal, S. K.: *Metrics to Evaluate your Classification Model to take the right decisions*, 2021. https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/, acesso em 2023-02-01. 27, 28, 29

[84] Padua, M.: *Machine Learning -Métricas de avaliação: Acurácia, Precisão e Recall, F1-score*, 2020. https://medium.com/@mateuspdua/machine-learning-metricas-de-avaliacao-acuracia-precisao-e-recall-d44c72307959, acesso em 2023-02-01. 27, 28, 29

[85] Alvarez, A. L. S.: *Classificação de RNA telomerase usando extração de características e técnicas de aprendizado de máquina*, 2023. https://github.com/nalualvarez/telomerase. 32, 43

[86] Inbamalar, T. M. e R. Sivakumar: *Improved algorithm for analysis of DNA sequences using multiresolution transformation*. The Scientific World Journal, 2015, 2015. 34

[87] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay: *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. 35, 36, 37

[88] Van Rossum, G. e F. L. Drake: *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009, ISBN 1441412697. 35

[89] Silveira, G.: *Machine learning: validação de modelos*, 2022. 35

[90] Liu, C.: *SVM Hyperparameter Tuning using GridSearchCV*, 2020. https://www.vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv/, acesso em 2023-02-01. 37

[91] West, J., D. Ventura e S. Warnick: *Spring research presentation: A theoretical foundation for inductive transfer*. Brigham Young University, College of Physical and Mathematical Sciences, 1(08), 2007. 43

Anexo I

Resultados completos dos experimentos

Neste anexo, apresentamos os resultados completos, obtidos dos métodos de extração de características e dos modelos de aprendizado de máquina utilizados neste trabalho.

Tabela I.1: Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.849	0.767	0.822	0.792	0.822	0.849	0.849	0.822	0.877	0.849	0.830	0.032
Precisão	0.760	0.640	0.720	0.680	0.677	0.724	0.741	0.677	0.778	0.783	0.718	0.048
Recall	0.792	0.667	0.750	0.708	0.875	0.875	0.833	0.875	0.875	0.750	0.800	0.078
AUC	0.835	0.741	0.804	0.771	0.835	0.856	0.845	0.835	0.876	0.824	0.822	0.040
F1 score	0.776	0.653	0.735	0.694	0.764	0.792	0.784	0.764	0.824	0.766	0.755	0.050
VN	43	40	42	40	39	41	42	39	43	44	41	1.767
FN	5	8	6	7	3	3	4	3	3	6	5	1.874
FP	6	9	7	8	10	8	7	10	6	5	8	1.713
VP	19	16	18	17	21	21	20	21	21	18	19	1.874

Tabela I.2: Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.822	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.790	0.031
Precisão	0.789	0.857	0.700	1.000	0.765	0.688	0.688	0.706	0.684	1.000	0.788	0.125
Recall	0.625	0.500	0.583	0.542	0.542	0.458	0.458	0.500	0.542	0.417	0.517	0.063
AUC	0.772	0.730	0.730	0.771	0.730	0.678	0.678	0.699	0.710	0.708	0.721	0.033
F1 score	0.698	0.632	0.636	0.703	0.634	0.550	0.550	0.585	0.605	0.588	0.618	0.053
VN	45	47	43	48	45	44	44	44	43	49	45	2.098
FN	9	12	10	11	11	13	13	12	11	14	12	1.506
FP	4	2	6	0	4	5	5	5	6	0	4	2.263
VP	15	12	14	13	13	11	11	12	13	10	12	1.506

Tabela I.3: Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.795	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.787	0.029
Precisão	0.696	0.810	0.692	0.759	0.818	0.818	0.808	0.667	0.840	0.900	0.781	0.075
Recall	0.667	0.708	0.750	0.917	0.750	0.750	0.875	0.833	0.875	0.750	0.787	0.082
AUC	0.760	0.813	0.793	0.885	0.834	0.834	0.886	0.815	0.897	0.855	0.837	0.044
F1 score	0.681	0.756	0.720	0.830	0.783	0.783	0.840	0.741	0.857	0.818	0.781	0.057
VN	42	45	41	41	45	45	44	39	45	47	43	2.503
FN	8	7	6	2	6	6	3	4	3	6	5	1.969
FP	7	4	8	7	4	4	5	10	4	2	6	2.415
VP	16	17	18	22	18	18	21	20	21	18	19	1.969

Tabela I.4: Resultados completos obtidos com o modelo baseado em Random Forest com método de extração de Transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.712	0.658	0.740	0.722	0.671	0.699	0.767	0.699	0.712	0.726	0.711	0.032
Precisão	0.615	0.462	0.778	0.833	0.500	0.583	1.000	0.625	0.636	0.750	0.678	0.163
Recall	0.333	0.250	0.292	0.208	0.125	0.292	0.292	0.208	0.292	0.250	0.254	0.060
AUC	0.616	0.554	0.625	0.594	0.532	0.595	0.646	0.574	0.605	0.605	0.595	0.034
F1 score	0.432	0.324	0.424	0.333	0.200	0.389	0.452	0.313	0.400	0.375	0.364	0.075
VN	44	42	47	47	46	44	49	46	45	47	46	2.003
FN	16	18	17	19	21	17	17	19	17	18	18	1.449
FP	5	7	2	1	3	5	0	3	4	2	3	2.098
VP	8	6	7	5	3	7	7	5	7	6	6	1.449

Tabela I.5: Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de Transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.808	0.795	0.767	0.778	0.822	0.822	0.795	0.753	0.767	0.795	0.790	0.023
Precisão	0.727	0.714	0.667	0.667	0.789	0.789	0.765	0.625	0.652	0.765	0.716	0.060
Recall	0.667	0.625	0.583	0.667	0.625	0.625	0.542	0.625	0.625	0.542	0.613	0.044
AUC	0.772	0.751	0.720	0.750	0.772	0.772	0.751	0.721	0.731	0.730	0.747	0.021
F1 score	0.696	0.667	0.622	0.667	0.698	0.698	0.634	0.625	0.638	0.634	0.658	0.031
VN	43	43	42	40	45	45	45	40	41	45	43	2.079
FN	8	9	10	8	9	9	11	9	9	11	9	1.059
FP	6	6	7	8	4	4	4	9	8	4	6	1.944
VP	16	15	14	16	15	15	13	15	15	13	15	1.059

Tabela I.6: Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.685	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.776	0.043
Precisão	0.524	0.423	0.407	0.366	0.388	0.500	0.444	0.342	0.404	0.396	0.419	0.057
Recall	0.458	0.458	0.458	0.625	0.792	0.458	0.500	0.542	0.875	0.792	0.596	0.164
AUC	0.627	0.576	0.565	0.542	0.590	0.617	0.597	0.516	0.621	0.600	0.585	0.036
F1 score	0.489	0.440	0.431	0.462	0.521	0.478	0.471	0.419	0.553	0.528	0.479	0.044
VN	39	34	33	22	19	38	34	24	18	20	28	8.266
FN	13	13	13	9	5	13	12	11	3	5	10	3.945
FP	10	15	16	26	30	11	15	25	31	29	21	8.189
VP	11	11	11	15	19	11	12	13	21	19	14	3.945

Tabela I.7: Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.685	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.776	0.043
Precisão	0.556	0.500	0.714	0.500	0.800	0.833	0.667	0.667	0.667	0.833	0.674	0.126
Recall	0.208	0.167	0.208	0.167	0.167	0.208	0.167	0.167	0.167	0.208	0.183	0.022
AUC	0.563	0.563	0.563	0.573	0.573	0.594	0.563	0.563	0.563	0.594	0.571	0.013
F1 score	0.303	0.250	0.323	0.250	0.276	0.333	0.267	0.267	0.267	0.333	0.287	0.033
VN	45	47	45	47	48	48	47	47	47	48	47	1.101
FN	19	20	19	20	20	19	20	20	20	19	20	0.516
FP	4	4	2	4	1	1	2	2	2	1	2	1.252
VP	5	4	5	4	4	5	4	4	4	5	4	0.516

Tabela I.8: Resultados completos obtidos com o modelo baseado em Naive Bayes Gaussiano com método de extração de transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.685	0.658	0.685	0.694	0.658	0.685	0.699	0.685	0.671	0.699	0.682	0.015
Precisão	0.667	0.400	0.667	1.000	0.333	0.667	1.000	0.667	0.500	1.000	0.690	0.244
Recall	0.083	0.083	0.083	0.083	0.042	0.083	0.083	0.083	0.083	0.083	0.079	0.013
AUC	0.531	0.511	0.531	0.542	0.500	0.531	0.541	0.531	0.521	0.542	0.528	0.014
F1 score	0.148	0.138	0.148	0.154	0.074	0.148	0.154	0.148	0.143	0.154	0.141	0.024
VN	48	46	48	48	47	48	49	48	47	49	48	0.919
FN	22	22	22	22	23	22	22	22	22	22	22	0.316
FP	1	3	1	0	2	1	0	1	2	0	1	0.994
VP	2	2	2	2	1	2	2	2	2	2	2	0.316

Tabela I.9: Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.000
Precisão	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Recall	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AUC	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.000
F1 score	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VN	49	49	49	49	49	49	49	49	49	49	49	0.000
FN	24	24	24	24	24	24	24	24	24	24	24	0.000
FP	0	0	0	0	0	0	0	0	0	0	0	0.000
VP	0	0	0	0	0	0	0	0	0	0	0	0.000

Tabela I.10: Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.699	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.778	0.040
Precisão	0.556	0.643	0.267	0.667	0.556	0.476	0.667	0.400	0.500	0.818	0.555	0.156
Recall	0.417	0.375	0.211	0.417	0.417	0.417	0.417	0.417	0.417	0.375	0.388	0.065
AUC	0.627	0.636	0.575	0.667	0.627	0.596	0.657	0.555	0.606	0.667	0.621	0.038
F1 score	0.476	0.474	0.235	0.513	0.476	0.444	0.513	0.408	0.455	0.514	0.451	0.083
VN	41	44	38	44	41	38	44	34	39	47	41	3.859
FN	14	15	15	14	14	14	14	14	14	15	14	0.483
FP	8	5	11	5	8	11	5	15	10	2	8	3.859
VP	10	9	4	10	10	10	10	10	10	9	9	1.874

Tabela I.11: Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.671	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.775	0.047
Precisão	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Recall	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AUC	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.000
F1 score	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VN	49	49	49	49	49	49	49	49	49	49	49	0.000
FN	24	24	24	24	24	24	24	24	24	24	24	0.000
FP	0	0	0	0	0	0	0	0	0	0	0	0.000
VP	0	0	0	0	0	0	0	0	0	0	0	0.000

Tabela I.12: Resultados completos obtidos com o modelo baseado em Naive Bayes de Bernoulli com método de extração de transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.000
Precisão	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Recall	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AUC	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.000
F1 score	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VN	49	49	49	49	49	49	49	49	49	49	49	0.000
FN	24	24	24	24	24	24	24	24	24	24	24	0.000
FP	0	0	0	0	0	0	0	0	0	0	0	0.000
VP	0	0	0	0	0	0	0	0	0	0	0	0.000

Tabela I.13: Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de transformada de Fourier com representação real, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.000
Precisão	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Recall	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AUC	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.000
F1 score	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VN	49	49	49	49	49	49	49	49	49	49	49	0.000
FN	0	0	0	0	0	0	0	0	0	0	0	0.000
FP	24	24	24	24	24	24	24	24	24	24	24	0.000
VP	0	0	0	0	0	0	0	0	0	0	0	0.000

Tabela I.14: Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de Redes Complexas, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.740	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.782	0.033
Precisão	0.778	0.667	0.600	0.727	0.571	0.538	0.667	0.667	0.533	0.800	0.655	0.094
Recall	0.292	0.333	0.250	0.333	0.333	0.292	0.333	0.333	0.333	0.333	0.317	0.029
AUC	0.625	0.626	0.584	0.635	0.605	0.585	0.626	0.626	0.595	0.646	0.615	0.022
F1 score	0.424	0.444	0.353	0.457	0.421	0.378	0.444	0.444	0.410	0.471	0.425	0.036
VN	47	45	45	45	43	43	45	45	42	47	45	1.636
FN	17	16	18	16	16	17	16	16	16	16	16	0.699
FP	2	4	4	3	6	6	4	4	7	2	4	1.687
VP	7	8	6	8	8	7	8	8	8	8	8	0.699

Tabela I.15: Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de EIIP, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.740	0.808	0.781	0.847	0.795	0.753	0.753	0.767	0.767	0.808	0.782	0.033
Precisão	0.778	0.667	0.600	0.727	0.571	0.538	0.667	0.667	0.533	0.800	0.655	0.094
Recall	0.292	0.333	0.250	0.333	0.333	0.292	0.333	0.333	0.333	0.333	0.317	0.029
AUC	0.625	0.626	0.584	0.635	0.605	0.585	0.626	0.626	0.595	0.646	0.615	0.022
F1 score	0.424	0.444	0.353	0.457	0.421	0.378	0.444	0.444	0.410	0.471	0.425	0.036
VN	47	45	45	45	43	43	45	45	42	47	45	1.636
FN	17	16	18	16	16	17	16	16	16	16	16	0.699
FP	2	4	4	3	6	6	4	4	7	2	4	1.687
VP	7	8	6	8	8	7	8	8	8	8	8	0.699

Tabela I.16: Resultados completos obtidos com o modelo baseado em Máquina de Vetor de Suporte com método de extração de transformada de Fourier com curva-z, onde VN é verdadeiro negativo, FN é falso negativo, FP é falso positivo e VP é verdadeiro positivo.

	1	2	3	4	5	6	7	8	9	10	Média	DP
Acurácia	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.671	0.000
Precisão	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Recall	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AUC	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.000
F1 score	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VN	49	49	49	49	49	49	49	49	49	49	49	0.000
FN	24	24	24	24	24	24	24	24	24	24	24	0.000
FP	0	0	0	0	0	0	0	0	0	0	0	0.000
VP	0	0	0	0	0	0	0	0	0	0	0	0.000