UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Tiago Tambonis

Utilização do método Suvrel no ranqueamento de genes envolvidos em expressão diferencial e na análise de otimização de predição de epítopos lineares de células B

São José do Rio Preto
2020

Tiago Tambonis

Utilização do método Suvrel no ranqueamento de genes envolvidos em expressão diferencial e na análise de otimização de predição de epítopos lineares de células B

Tese apresentada como parte dos requisitos para obtenção do título de Doutor em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: CAPES

Orientador: Prof. Dr. Vitor Barbanti Pereira Leite

São José do Rio Preto
2020

T155u

Tambonis, Tiago

Utilização do método Suvrel no ranqueamento de genes envolvidos em expressão diferencial e na análise de otimização de predição de epítopos lineares de células B / Tiago Tambonis. -- São José do Rio Preto, 2020

93 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Vitor Barbanti Pereira Leite

1. Sequenciamento de nucleotídeos em larga escala. 2. Expressão gênica. 3. Reações antigeno-anticorpo. 4. Predição. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Tiago Tambonis

Utilização do método Suvrel no ranqueamento de genes envolvidos em expressão diferencial e na análise de otimização de predição de epítopos lineares de células B

Tese apresentada como parte dos requisitos para obtenção do título de Doutor em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: CAPES

Comissão Examinadora

Prof. Dr. Vitor Barbanti Pereira Leite
UNESP – São José do Rio Preto – SP
Orientador

Dr. Fabio Rogério de Moraes
UNESP – São José do Rio Preto – SP

Prof. Dr. Rodrigo Capobianco Guido
UNESP – São José do Rio Preto – SP

Prof. Dr. Michel Beleza Yamagishi
Embrapa Informática Agropecuária – SP

Prof. Dr. Renato Vicente
USP – São Paulo

São José do Rio Preto
24 de outubro de 2019

Dedico este trabalho a toda minha família, principalmente a minha mãe, Rosane, minha irmã, Priscila, as minhas avós, Yolanda e Mirtes e a minha esposa, Bianca.

Agradecimentos

A realização desta tese foi possível graças ao apoio e colaboração de muitas pessoas. De fato, foram muitas pessoas. A seguir vou agradecer as pessoas que lembro. Acho o diálogo muito valioso, e por isso talvez algumas pessoas não tenham percebido mas podem ter me ajudado sem terem percebido. Dada o exposto, é natural a existência da possibilidade de eu não conseguir lembrar de todos, porém, eu sou grato pela ajuda.

Agradeço a minha família por todo apoio, compreensão e amor. Agradeço sobretudo as mulheres que marcaram a minha vida: minha mãe, Rosane, minha irmã Priscila, minhas avós, Mirtes e Yolanda, e minha esposa, Bianca. Agradeço ao meu grande amigo Vitor e sua família, responsáveis por apoio indispensável em determinadas etapas da minha vida. Agradeço ao meu amigo Luiz Aurélio, pela amizade e por ter me mostrado a filosofia budista de Nitiren Daishonin, mudando definitivamente os rumos da minha vida. Agradeço ao meu amigo, Rafael Franco, por ter me apoiado quando não estava feliz no curso de engenharia de telecomunicações.

Agradeço ao meu orientador Vitor Barbanti Pereira Leite pela oportunidade cedida e por me orientar. Agradeço de maneira geral ao grupo de pesquisa que faço parte. Agradeço aos meus amigos do Departamento de Física que convivo diariamente por construírem um ambiente de trabalho muito sadio, aprazível e familiar: Renan, João, Ingrid, Zézinho, Paulo, Fernando, Carol, Gabi, Taísa, Antônio, Kaique, Rafael, Karol, Josimar, Coragem, Marcelo, Daniel, Goiás, Jesus, David, Luan e Guilherme e Keneth.

Agradeço também a todos os funcionários do IBILCE/UNESP por fazerem desta instituição um ótimo lugar para estudar, especialmente ao Bruno Rogério pelas valiosas dicas computacionais.

Agradeço a Daisaku Ikeda por difundir o budismo de Nitiren Daishonin, me proporcionando a oportunidade de conhecer essa filosofia e a todos meus companheiros da BSGI por me apoiarem.

Agradeço aos membros da banca do exame geral de qualificação de dou-

torado, Professor Doutor Rodrigo Capobianco Guido e o Doutor Fábio Rogério de Moraes, pelas sugestões apresentadas na minha qualificação e por tem aceitado participar da defesa. Agradeço por fim, ao Prof. Dr. Renato Vicente e o Prof. Dr. Michel Beleza Yamagishi, por terem aceitado participar da banca de defesa.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

RNA-seq. Na primeira parte desta tese, uma abordagem geométrica baseada no Método Suvrel ("Supervised Variational Relevance Learning") é comparada com pacotes R destinados a expressão diferencial. O método Suvrel procura determinar as relevâncias das características (por exemplo, genes ou transcritos) baseado em comparações de distâncias interclasses e intraclasses. A investigação foi realizada utilizando replicatas técnicas e biológicas. A análise utilizando replicatas técnicas foi realizada por meio de curvas ROC enquanto que as replicatas biológicas foram analisadas por meio de robustez. De forma geral, é mostrado que a análise proposta obteve melhores resultados na maioria dos casos. Além disso, é um método simples que não faz nenhuma suposição sobre a distribuição associada ao conjunto de dados de RNA-seq. Desta perspectiva, a relevância deste estudo foi mostrar que um método simples pode fornecer boa acurácia tanto quanto os métodos mais complexos.

Otimização de predição de epítopos lineares de células B. Epítopos são definidos como fragmentos constituintes de um antígeno (substância que ao entrar em um organismo é capaz de iniciar uma resposta imune) que interagem com receptores de células B, T e anticorpos. Após a ativação da resposta imune adaptativa é possível que o sistema imune crie memória, onde uma outra invasão será respondida de forma mais rápida. A identificação de epítopos utilizando processos experimentais ainda permanece difícil. Devido ao potencial de aplicação de ferramentas preditivas que podem auxiliar a identificação, uma série de algoritmos foram propostos. Dado o panorama exposto, a proposta da presente tese é a análise de possível melhora da acurácia de predição de epítopos lineares de células B associada ao preditor *LBTOPE* por meio do método Suvrel. A metodologia apresentada melhorou consideravelmente medidas estatísticas em dados gerados pelos autores do preditor e também em dados gerados por outros autores.

Palavras-chave: Expressão diferencial. RNA-seq. Abordagem Geométrica. Epítopos. Protozoários. Predição. Suvrel.

Abstract

RNA-seq. Although differential gene expression (DGE) profiling in RNA-seq is used by many researchers, new packages and pipelines are continuously being presented as a result of an ongoing investigation. In the present work, a geometric approach based on Supervised Variational Relevance Learning (Suvrel) was compared with DEpackages (edgeR, DESEq, baySeq, PoissonSeq, and limma) in the DGE profiling. The Suvrel method seeks to determine the relevance of characteristics (e.g., gene or transcript) based on intraclass and interclass distances. The comparison was performed using technical and biological replicates. The results showed that geometric approach had a better performance than the DEpackages. Therefore, the conclusion is that the geometric approach had a slight overall better performance than other methods. Moreover, it is a simple method that does not make any assumption about the distribution associated with RNA-seq data set. From this perspective, the relevance of this study was to show that a simple method can provide as good performance as more complex methods.

Prediction optimization of linear B-cell epitopes. Epitopes are defined as fragments of an antigen (a substance that invade an organism capable of start an immune response) that interact with B cell receptors, T and antibodies. After activation of the adaptive immune response it is possible for the immune system to generate memory where another invasion will be attacked faster. Identifying epitopes using experimental procedures still difficult. However, predictive tools can aid the identification. The purpose of this thesis is to analyze the possible improvement of the prediction performance of linear B-cell epitopes associated to predictor LBTOPE using the Suvrel method. The results show that the methodology improved statistical measures in data generated by the authors of predictor and also in data generated by other authors.

Keywords: Differential expression. RNA-seq. Geometric Approach. Epitopes. Protozoa. Prediction. Suvrel.

Lista de Figuras

Lista de Tabelas

Sumário

Parte I

Inferência de expressão diferencial de genes a partir de dados de RNA-seq usando um método baseado no Suvrel

Capítulo 1

Introdução

Ao analisar o intervalo de tempo compreendido entre meados da década de 90 até os dias atuais, pode-se constatar que a quantidade de informações geradas associadas a sequenciamento depositada em bibliotecas públicas cresceu exponencialmente. Entre as razões para este crescimento, podem-se elencar inúmeros fatores mas a diminuição considerável do custo de sequenciamento pode resumir a argumentação para explicar essa intensificação. A diminuição do custo de sequenciamento se tornou mais acentuada a partir dos anos 2006 e 2007 devido ao advento do sequenciamento de alto desempenho ("Next-Generation Sequencing", NGS), que por sua vez trouxe um aumento considerável no volume de dados gerados. A partir dessa mudança significativa na geração de dados, a hipótese reducionista pré-genômica da Biologia Molecular, que era aplicada em sistemas pequenos, foi modificada para a abordagem pós-genômica, que é aplicada em grandes sistemas e dirigida pela quantidade massiva de dados, atualmente na escala de armazenamento e processamento em terabytes [1–6].

Entre as modalidades de sequenciamento que foram impactadas pelo NGS, pode-se atentar ao sequenciamento de RNA (RNA-seq). Utilizando esta técnica associada à utilização de algoritmos computacionais, tornou-se possível a reconstrução de transcritomas completos com resolução melhor que de microarranjos [7]. Com as informações deste tipo de sequenciamento também é possível a análise de transcritos inteiros, aumentando a possibilidade de entendimento sobre dinâmica dos transcritomas [8–10]. Além da reconstrução de transcritoma, o avanço da tecnologia de RNA-seq tem propiciado outros desafios e oportunidades para pesquisadores, como a quantificação e análise de transcritos expressos diferencialmente [9].

A análise de expressão diferencial (EDG) é realizada por meio da análise das mudanças dos níveis de expressões dos genes. Entre as opções disponíveis, é possível obter informações para realizar esta tarefa por meio de RNA-seq. Para facilitar o de-

senvolvimento do texto, os passos associados à análise de EDG serão omitidos até a construção da tabela de contagens. Mais detalhes serão tratados no decorrer do texto. Na tabela mencionada, de forma sucinta, cada alocação dela sumariza uma medida de intensidade de expressão que um determinado gene teve em uma dada condição experimental. A partir desta tabela é possível identificar genes expressos diferencialmente analisando aqueles que tiveram sua abundância significativamente alterada entre as condições analisadas. Dependendo do objetivo do estudo, a geração de uma lista de ranqueamento de genes não é a conclusão final, mas um passo de uma análise mais ampla. Os dados gerados por RNA-Seq, como a lista de genes expressos diferencialmente, pode ser integrada a outras fontes de informações, como por exemplo, estudo de RNA de interferência, modificação de histona e metilação de DNA, para estabelecer um entendimento mais claro sobre mecanismos regulatórios [7]. RNA-seq também pode ser usado no campo da biologia de sistemas, a qual vem ganhando atenção de muitos pesquisadores. Nesta área, as entidades biológicas são interconectadas e integradas de forma a constituir redes moleculares. Dessa forma, tem-se o objetivo de, ao se utilizar múltiplas plataformas *ômicas*, conseguir compreensão sobre um sistema biológico que não se conseguiria utilizando somente uma plataforma. RNA-seq pode ser parte desta área a partir de uma lista de genes expressos diferencialmente, onde então identifica-se os genes que tiveram mudanças nas intensidades de expressão. Em seguida, utiliza-se estas informações para analisar em quais processos e caminhos biológicos tais genes estão envolvidos [11–13]. Considerando ainda integração de dados *ômicos*, RNA-seq também pode ser utilizado associado à metabolômica. Essa área se caracteriza pela análise de moléculas menores que 1200 Da e intermediários bioquímicos (metabólitos). Entre as áreas de estudo onde emprega-se a metabolômica, tem-se por exemplo o estudo de diabetes tipo 1 e câncer. Quando atenta-se a este estudo, o objetivo pode ser a identificação de biomarcadores associados ao início de uma doença, monitoramento de eficácia de tratamento e prognósticos. Dessa forma, a utilização de RNA-seq pode auxiliar alcançar tais objetivos por meio da exploração da interação entre perfis de expressão diferencial e mudanças nos níveis de metabólitos [14, 15].

Quando pretende-se identificar os genes que apresentaram mudanças estatisticamente significantes na abundância, deve-se atentar à existência de dois tipos de fontes de variações. O primeiro tipo é aquele que está ligado à variação relativa devido à causas biológicas, ao passo que o segundo tipo está ligado à incerteza associada à tecnologia de sequenciamento com o qual a abundância do transcrito é estimada. Portanto, um transcrito é declarado diferencialmente expresso se as quantidades diferem significativamente (atentando para os dois tipos de variações) em grupos distintos de amostras. A avaliação de expressão diferencial utilizando RNA-Seq é iniciada com o

sequenciamento das amostras, gerando bilhões de pequenas sequências, as quais posteriormente são mapeadas em um genoma de referência, sumarizadas e contadas. O próximo passo é conhecido como normalização, necessário devido a forma de sequenciamento por amostragem. Por fim, realiza-se a análise estatística que apontará o transcrito ou conjunto de transcritos que tiveram mudanças estatisticamente relevantes em suas abundâncias. Os dois últimos passos são realizados de forma conjunta. Dessa forma, na maioria das vezes, cada ferramenta de análise de EDG tem sua respectiva forma de normalizar. Embora existam pesquisas em todos os passos mencionados, ainda existe a necessidade de aperfeiçoamentos e melhorias. Nessa perspectiva, os refinamentos podem ir desde maiores estudos sobre qual a métrica para a sumarização mais adequada, até qual o efeito das suposições (é comum os pacotes destinados a expressão diferencial assumirem que as contagens seguem algum tipo de distribuição) feitas pelos métodos de inferência de expressão diferencial [7, 16]. A divisão desta parte da tese foi produzida da seguinte forma: descrição dos passos envolvidos em experimentos de expressão diferencial utilizando RNA-seq, posteriormente, apresenta-se os métodos de inferência que serão analisados, em seguida, descrição das ferramentas disponíveis para comparação de performance. Por fim, o manuscrito intitulado "Differential expression analysis in RNA-seq data using a geometric approach"publicado na revista "Journal of Computational Biology"[17] mostra os resultados obtidos utilizando a metodologia proposta.

Capítulo 2

Motivação e Objetivos

O uso de sequenciamento de alto rendimento produz um volume denso de informações, impondo a necessidade de metodologias adequadas que ajudem pesquisadores e correlatos a interpretarem de maneira apropriada as informações [7]. O RNA-seq de nova geração também é integrante deste cenário e os pacotes e as metodologias computacionais usadas ainda carecem de melhorias e aprimoramentos [7]. A análise de EDG é parte deste panorama, e por isso propõe-se a utilização de uma abordagem geométrica baseada no método Suvrel para análise de expressão diferencial de genes. Tal proposta também é motivada por estudos anteriores que mostraram que a metodologia chamada de Aprendizado de Relevância Variacional Supervisionado ("Supervised Variational Relevance Learning", Suvrel) melhorou a performance de ferramentas computacionais de predição de dados originados de microarranjos e concentrações de metabólitos medidos por ressonância magnética nuclear [18]. Dessa forma, tem-se o objetivo de analisar a possibilidade de o ranqueamento de genes envolvidos em expressão diferencial ser aperfeiçoado por meio da abordagem proposta.

Capítulo 3

Inferência de expressão diferencial por meio de experimentos de RNA-Seq

3.1 Visão geral dos passos envolvidos na análise de expressão diferencial

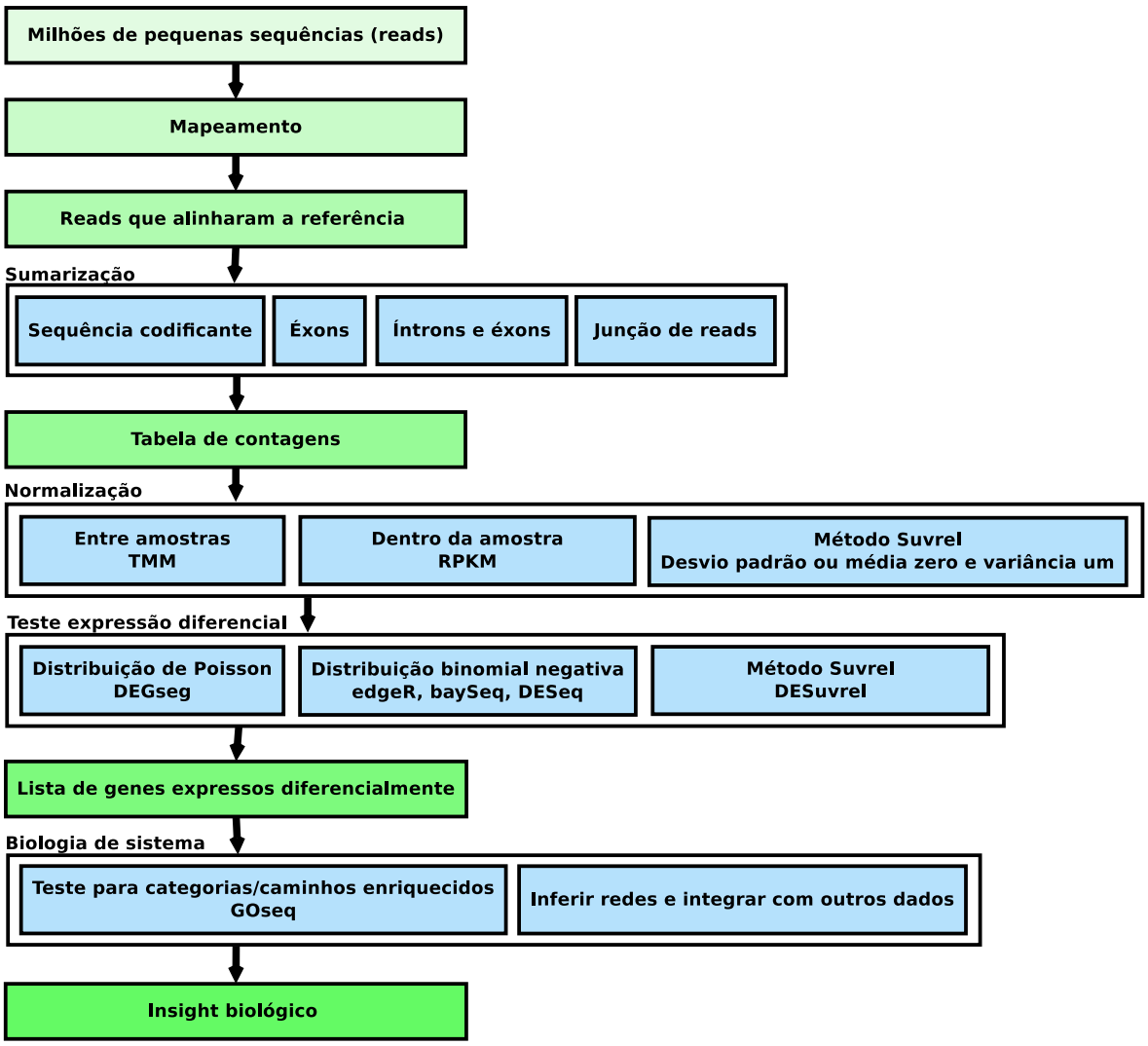
A visão geral dos passos envolvidos na análise de expressão diferencial por meio de RNA-seq é representada na Figura 3.1. Este tipo de experimento se inicia com o sequenciamento das amostras associadas a diferentes condições experimentais, passo este que gera de milhões a bilhões de pequenas sequências de nucleotídeos ("reads") [19]. O tamanho das sequências pode variar de acordo com o desenvolvimento da tecnologia envolvida e, no momento que me embasei para redigir este texto, elas possuíam tamanho de 25 a 55 e média de 33 nucleotídeos [9, 20]. Seguidamente, as sequências passarão por verificação de qualidade e, se preciso, algumas podem ser descartadas caso a qualidade seja baixa [7]. O próximo passo do experimento é o mapeamento dos "reads" em um genoma ou transcritoma de referência. Em seguida, as sequências são sumarizados e aglomeradas a partir de uma unidade de significado biológico, por exemplo, genes, transcritos ou éxons. Após esta etapa, gera-se uma tabela de contagens que necessita ser normalizada, geralmente, utilizando uma abordagem atrelada ao método que será usado para a inferência da expressão diferencial. Os passos precedentes são seguidos pela aplicação da abordagem de inferência de expressão diferencial, por exemplo, a presente no pacotes edgeR ou DESeq e a abordagem proposta nesta teste. Por fim, estes passos produzem uma lista de genes (ou qualquer outra característica

genômica de interesse) acompanhados pelos respectivos "fold-changes" e valores-p, na maioria dos casos. Dada as particularidades e características próprias de cada passo envolvido na análise de expressão, os tópicos a seguir tratarão com mais detalhes cada um deles.

3.2 Sequenciamento das amostras

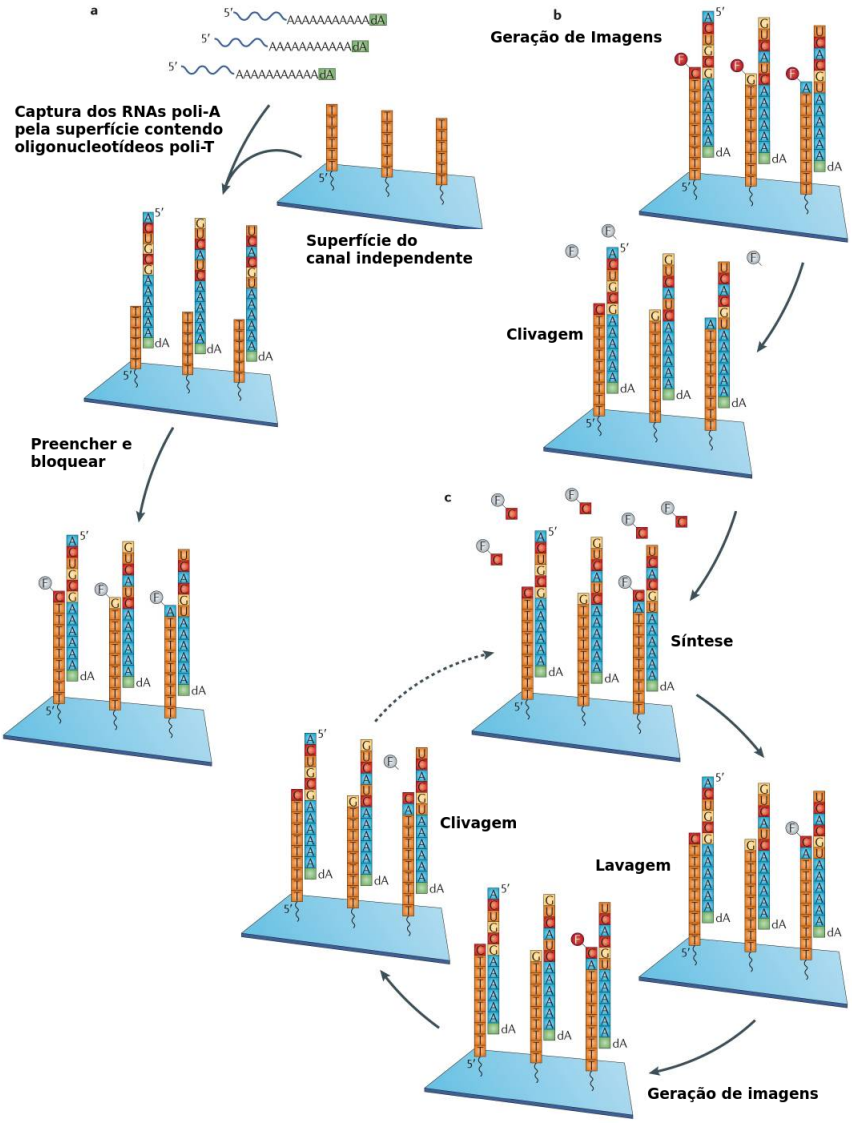
Amostras de RNA podem ser sequenciadas por meio de síntese de DNA complementar [8] ou por síntese [20]. Nesta tese, atenção será dada ao segundo caso devido ao seu amplo uso e potencial de crescimento. Para demonstrar os principais passos envolvidos no sequenciamento de RNA por síntese ilustrados na Figura 3.2, a tecnologia desenvolvida pela Helicos será usada como exemplo. Nesta máquina, os nucleotídeos poli-A são capturados pela calda 3' por oligonucleotídeos poli-dT ligados covalentemente pela cauda 5' a uma superfície de vidro ultra-limpo. O sequenciamento se incia em uma etapa denominada "preencher", integrante do par de etapa "preencher e travar". Este passo tem o objetivo de expor os pares hibridizados de RNA e poli-dT às polimerases e excesso de timinas naturais para que o sequenciamento não se inicie na cauda poli-A. No passo "travar", nucleotídeos adenina, citosina e guanina com seus respectivos grupos fluorescentes e quimicamente cliváveis são hibridizados às respectivas bases complementares, concluindo esta fase. A próxima etapa é denominada de lavagem, onde, então, os nucleotídeos não incorporados são descartados, assim como o excesso de timinas. Este ciclo de sequenciamento termina quando imagens são geradas e uma mistura que cliva o corante fluorescente do último nucleotídeo hibridizado é adicionada, tornando as fitas adequadas para outra ciclo de incorporação. O processo de sequenciamento prossegue com a adição de nucleotídeos distintos (no caso da Figura 3.2, C, citosina) com grupos corantes e cliváveis, seguida pela lavagem, geração de imagens e clivagem. A iteração deste último ciclo por muitas vezes com outras bases, fornece um grande conjunto de imagens, onde as incorporações das bases são detectadas e então usadas para gerar a informação de sequenciamento para cada *read*. A máquina usada como exemplo contém cerca de 50 canais independentes, onde cada utilização produzirá entre 800.000 a 12.000.000 "reads", com tamanho de 25 a 55 e média de 33 nucleotídeos, dependendo do tempo de execução e da qualidade da geração das imagens [9, 20]. Devido às tecnologias de sequenciamento continuarem em desenvolvimento estes números provavelmente cresceram e continuarão a crescer.

Figura 3.1: Fluxograma dos passos envolvidos em uma análise para detecção de expressão diferencial de genes. Milhões de "reads" são produzidos após amostras de RNA terem sido sequenciadas (caixa Milhões de pequenas sequências ("reads")). O mapeamento pode ser produzido com o auxílio de um genoma ou transcritoma de referência usando por exemplo o programa bowtie (caixa Mapeamento). Após os "reads" terem sido alinhados, a próxima etapa é a sumarização, que basicamente conta uma característica ligada ao objetivo do estudo, que no caso mais geral são os "reads" que alinharam ao respectivos genes de origem (caixa Sumarização). A partir da conclusão desta etapa, tem-se gerada a tabela de contagens que deverá ser normalizada (caixa Tabela de contagens e Normalização). A inferência de expressão diferencial é executada no teste de expressão diferencial, podendo ser realizada supondo que os "reads" seguem uma distribuição de Poisson, binomial negativa ou a abordagem geométrica proposta que não necessita de nenhuma suposição dessa natureza (Caixa Teste de expressão diferencial). Na conclusão desta etapa, tem-se gerada a lista de genes diferencialmente expressos (Caixa Lista de genes expressos diferencialmente). Os passos envolvidos na etapa chamada de biologia de sistemas são os estudos dos resultados obtidos com os passos anteriores, juntamente com outros dados auxiliares no sentido de obter um "insight" associado ao objetivo que motivou o estudo.



Fonte: obtido da Ref. [21] e adaptada da Ref. [7].

Figura 3.2: Esquema representativo da plataforma Helicos de sequenciamento. RNA poli-A por meio da calda 3' é capturado pela superfície contendo poli-dT. O passo posterior envolve uma etapa denominada de *"preencher-e-travar"* que introduz nucleotídeos A, C e G marcados quimicamente. Estes últimos passos garantem que o sequenciamento não se inicie na calda poli-A. Na parte b, imagens são geradas e o fluoróforo das bases que foram ligadas complementarmente é clivado. Na parte c, ciclos de introdução de bases distintas ligadas à fluoróforos, geração de imagens e clivagens são executados. Ao final do sequenciamento, são produzidas entre 800.000 a 12.000.000 "reads", com tamanho de 25 a 55 e média de 33 nucleotídeos dependendo do tempo de execução e da qualidade das gerações das imagens (as tecnologias de sequenciamento continuam em desenvolvimento e provavelmente estes números são maiores).



Fonte: obtido da Ref. [21] e adaptada da Ref. [20].

Figura 3.3: Representação do mapeamento ou alinhamento de “reads”. As sequências estão representadas pelos retângulos vermelhos e estão alinhadas a um gene de referência constituído por retângulos azuis e laranjas. A partir do alinhamento os “reads” são convertidos em medidas de quantificação de expressão. Essa medida de quantificação poderia ser quantidade de “reads” por éxons de um gene se os retângulos azuis simbolizassem os éxons do referido gene.

Fonte: adaptado de http://bio.lundberg.gu.se/courses/vt13/rnaseq.html.

3.3 Mapeamento ou alinhamento

Quando se tem o objetivo de produzir um experimento de análise de expressão diferencial por meio de RNA-seq, após as amostras terem sido sequenciadas, a próxima etapa é denominada mapeamento ou alinhamento. Nesta etapa, os “reads” são alinhados a um genoma ou transcritoma de referência e convertidos em medidas de quantificação de expressão (Figura 3.3) [7, 19, 22]. O mapeamento pode ser realizado por meio de inúmeras ferramentas [19] e elas podem ser divididas em duas categorias predominantes: as ferramentas baseadas em tabelas “hash” e as baseadas em transformação *Burrows–Wheeler* [23]. Uma tabela “hash” é uma estrutura de dados, usada para indexação de conjuntos de dados complexos para facilitar a busca rápida por “strings”. Transformação *Burrows–Wheeler* é um rearranjo reversível de “strings” que codifica o genoma em uma representação mais compacta, produzindo redundâncias de subsequências repetidas [22].

Grande parte dos algoritmos de alinhamento inicialmente executa um “matching” eurístico, o qual gera uma lista de possíveis localizações. Posteriormente, executa-se uma avaliação completa de todos os candidatos de alinhamento por meio de complexos algoritmos de alinhamento local exato [7, 22]. É possível que dadas as informações expostas ocorra a indução à conclusão de que a tarefa de encontrar uma localização única onde um “read” possua sequência idêntica à referência pareça ser fácil. Porém, isso não acontece [7]. A primeira consideração a respeito está relacionada à observação de que a referência nunca representa, de maneira perfeita, a fonte biológica do RNA sendo sequenciado. Expondo em outras palavras, o RNA sequenciado para a análise hipotética no texto não foi o utilizado na produção da referência utilizada no alinhamento. Possíveis entraves neste passo também podem vir por meio de atributos específicos da amostra, como a variação na sequência de DNA que afeta somente uma base (“Single Nucleotide Polymorphism”, SNP) e “indels” (inserções ou deleções). Desafios na conclusão desta etapa ainda podem ser originados do conhecimento de que os

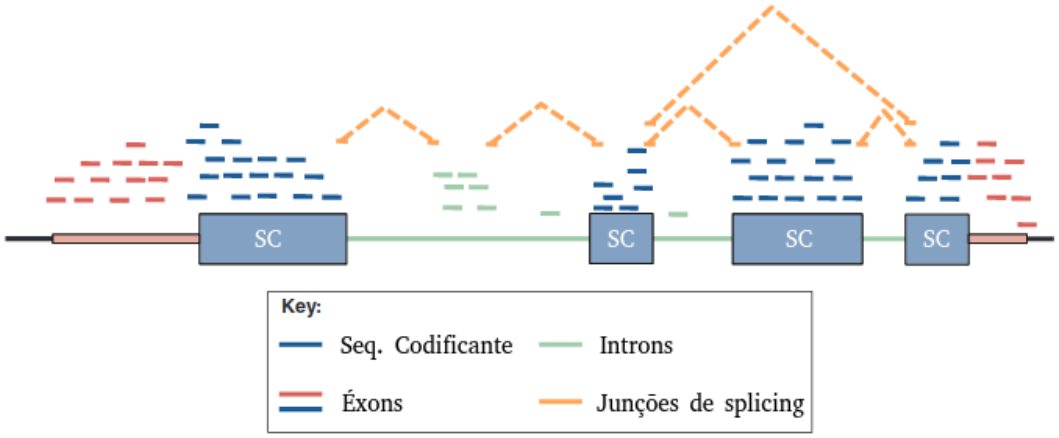
"reads" estão associados a transcritos, que por sua vez, possivelmente, passaram por "splicing". Ademais, as sequências podem se alinhar perfeitamente a múltiplas localizações e podem conter erros de sequenciamento que precisam ser levados em conta [7, 22]. Quando considera-se "reads" obtidos da transcrição reversa de DNA, onde os finais dos fragmentos de RNA podem ser sequenciados [8], os algoritmos de alinhamento diferem em como eles resolvem os "reads" que mapearam igualmente em várias regiões. Entre as opções possíveis, a ferramenta de mapeamento pode descartar tais "reads", alocá-los randomicamente ou utilizar métodos estatísticos que incorporam "scores" de alinhamento. Esta última situação pode trazer mais desafios quando somente um final dos fragmentos é sequenciado [7].

Por fim, existem pesquisas na análise da melhor maneira de alinhar as sequências. Contudo, todas as soluções envolvem, fatalmente, algum consenso entre os recursos computacionais utilizados pelos algoritmos e a incerteza permitida (vinda por exemplo ao ignorar informação de SNPs, negligenciar os "scores" de qualidades das bases, limitação do número de "mismatch" permitidos) no "matching" à referência [19].

3.4 Sumarização

Nesta etapa do experimento, os "reads" são sumarizadas sobre alguma unidade de significado biológico, como por exemplo, genes, éxons ou transcritos, utilizando ferramentas tais como HTSeq, Picard, BEDTools entre outras [7, 19]. A Figura 3.4 ilustra a representação de um gene, o qual é constituído por sequências codificantes, éxons e íntrons. Na figura também é possível visualizar junções de "splicing". Dependendo do objetivo associado à análise de expressão diferencial, pode-se sumarizar as sequências alinhadas contando o número delas que se alinharam, por exemplo, em éxons. Entretanto, é possível que o objetivo da análise de expressão diferencial por meio de éxons não seja o objetivo final ou talvez exista uma proporção significante dos "reads" que mapeiam em regiões genômicas fora dos éxons anotados, até mesmo em organismos bem anotados [22, 24]. É possível solucionar este entrave incluindo as sequências que alinharam em todo o comprimento do gene por meio da inclusão de íntrons na contagem. A solução também pode ser a utilização de éxons pobremente ou não anotados, assim como fronteiras de éxons variáveis. A partir do exposto, é possível verificar que existem muitas opções associadas a esta etapa. Porém, é necessário ter atenção na escolha da sumarização pois ela tem o potencial de mudar substancialmente as contagens de cada gene e levar a conclusões errôneas [7, 19].

Figura 3.4: Representação de um gene. Dependendo do objetivo do estudo, é possível sumarizar os "reads" que alinharam às sequências codificantes.



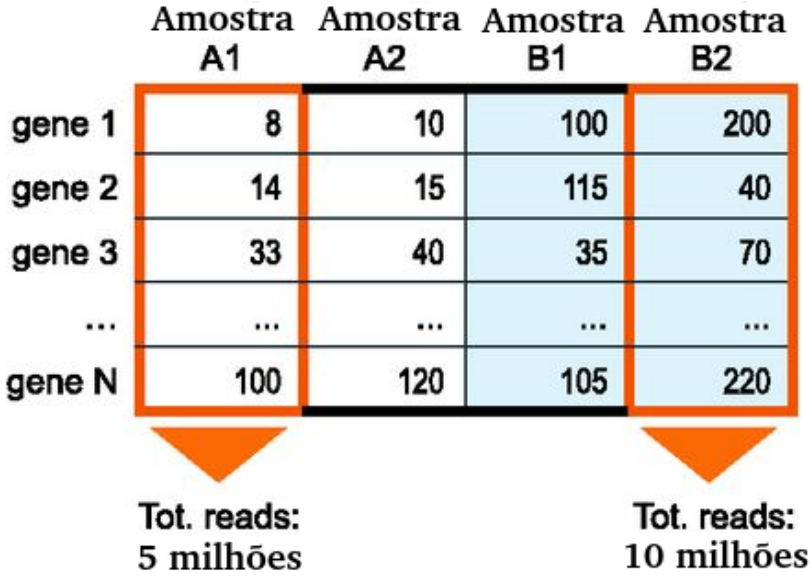
Fonte: adaptada da Ref. [7].

Após a conclusão dos passos descritos até aqui, tem-se a geração de uma matriz de contagem N de n *por* m, onde N_{ij} é o número de sequências atribuídas ao gene i no experimento de sequenciamento j, n é o número de genes analisados e m é o número de experimentos.

3.5 Normalização

O número de sequências que se alinham a um dado gene i não é uma medida direta da expressão do gene. A medida de contagem N_{ij} é proporcional a $l_i\mu_{ij}$, onde μ_{ij} e l_i são, respectivamente, a expressão esperada (número de fragmentos associados ao gene i) e o comprimento do gene, caracterizando possível viés de comprimento. A ocorrência deste fato pode abreviar a habilidade de detecção de expressão diferencial em genes com contagens pequenas devido a falta de cobertura. Nesta etapa, o possível problema exposto, assim como as diferenças entre os números totais de sequências entre diferentes corridas de sequenciamento, e presença de viés técnicos introduzidos por protocolos de preparação de bibliotecas devem ser analisados. Um exemplo de uma situação que pode ocorrer pode ser discutida e analisada por meio da Figura 3.5. Na ilustração, é mostrada uma tabela hipotética de contagens, constituída por medidas de expressões de N genes e 4 amostras divididas em duas condições A e B. Atentando para análise da coluna das amostras $A1$ e $B2$, é possível concluir que a partir do números totais de "reads" diferentes, 5 milhões e 10 milhões respectivamente, não é possível comparar diretamente as medidas de expressões. Portanto, o objetivo da normalização é tentar resolver ou amenizar efeitos indesejados como este para facilitar comparações

Figura 3.5: Representação de uma tabela de contagens hipotética obtida por meio de RNA-seq. As colunas listam as amostras e as linhas os genes. Assim, uma célula na tabela mostra a medida de expressão de um gene associada à amostra.



Fonte: adaptada da Ref. [22].

acuradas entre as amostras [25].

Suponha a existência de dois experimentos de RNA-Seq sem genes expressos diferencialmente. Considere ainda que o primeiro experimento gere duas vezes mais "reads" que o segundo. Assim, é possível concluir que as contagens do primeiro experimento são duplicadas e, assim, sermos inclinados a dizer que ocorre expressão diferencial. Este cenário, que pode levar a conclusões errôneas, acontece devido a diferentes profundidades de sequenciamento d_j (do inglês "sequencing depth", sendo definido como o número total de "reads" sequenciados ou mapeados). Para solucionar este problema, pesquisadores introduziram uma técnica que ficou conhecida como escalonamento global, onde as contagens em cada experimento j eram escalonadas por d_j [7, 22, 25].

Contudo, alguns pesquisadores optam por se atentarem mais a outro tipo de viés. A situação referida se caracteriza pela predominância de um conjunto restrito de genes altamente expressos que consomem boa parte dos "reads" disponíveis e, por consequência, alguns genes ficam subestimados. Na tentativa de resolver este entrave, Bullard et al. sugeriram uma normalização por quantil codificada no pacote baySeq [26] na forma de um escalonamento global, o qual ajusta as distribuições das contagens por meio de seus terceiros quartis. Robinson e Oshlack et al. propuseram outra forma de resolver, por meio do método de normalização denominado de "Trimmed Mean of M-values", (TMM), implementado no edgeR [27]. Neste método, um fator de

normalização é gerado após remover 30% dos genes que tem os mais extremos logs de "fold-changes", conhecidos como valores-M, para então usar este fator para corrigir as diferenças nos tamanhos das bibliotecas [22, 25].

Os possíveis vieses oriundos da utilização de RNA-seq foram descritos. Daqui por diante até o final desta seção, os métodos de normalização utilizados pelos pacotes que serão analisados na comparação com a abordagem proposta serão descritos.

Para investigar o método de normalização utilizado pelo pacote R DESEq [28], faz-se necessário relembrar a definição de N_{ij}, a contagem atribuída ao gene i no experimento j. Com base nesta definição, se supormos duas replicatas atribuídas com *j=1* e *j=2*, espera-se que um histograma das proporções

$$\frac{N_{i1}}{N_{i2}}, \tag{3.1}$$

considerando todos os genes i, mostre um pico bem definido e que sua mediana represente uma boa estimativa da proporção da diferença na profundidade de sequenciamento. Embora existam experimentos que utilizem uma única replicata por condição experimental, a maioria dos estudos utilizam mais do que duas replicatas. Neste caso, uma amostra de referência é produzida calculando a média geométrica f_i de um gene i considerando todas as amostras até ter este cálculo para todos os genes. Deste modo, o fator de escalonamento da amostra j é obtido a partir da mediana

$$s_j = \frac{N_{ij}}{f_i}, \tag{3.2}$$

considerando todos os genes i. Alicerçado na obtenção do valor de s_j, divide-se cada contagem por esse valor para se obter as contagens normalizadas [22, 25, 28].

O pacote R PoissonSeq [29] utiliza uma estimativa vinda de um ajuste de qualidade ("goodness-of-fit") para determinar um conjunto de genes que é diferenciável entre duas condições para então calcular os fatores de normalização [25].

O pacote R limma [30] inicialmente concebido para análise de expressão diferencial por meio de dados vindos de microarranjo, utiliza normalização por quantil. Uma nova versão foi desenvolvida para utilização em dados vindos de RNA-seq, e uma nova função de normalização chamada "voom" [31], desenvolvida especificamente para dados de RNA-Seq foi implementada. De uma forma geral, este método realiza uma regressão "LOWESS" para estimar a relação entre a média e a variância, e transforma as contagens na escala log para executar uma transformação linear [25].

Tabela 3.1: Tabela informativa que lista os métodos de normalização executados pelos pacotes que serão estudados juntamente com as informações da abordagem proposta.

Pacote	**Método**
baySeq	Escalonamento global que ajusta as distribuições das contagens de um experimento a partir do terceiro quantil.
edgeR	Escalonamento global utilizando um fator que foi calculado após 30% dos genes com os mais extremos logs de "fold-changes" terem sido retirados.
DESeq	Utiliza escalonamento global a partir de um fator s_j que é calculado como a mediana de $\frac{N_{ij}}{f_i}$, onde N_{ij} é a contagem atribuída ao gene i no experimento j e, f_i é a média geométrica de um gene i considerando todas as amostras
PoissonSeq	Utiliza uma estimativa vinda de um ajuste de qualidade para definir um conjunto de genes que é pelo menos diferenciável entre duas condições para então calcular os fatores de normalização das bibliotecas.
limmaVoom	Realiza uma regressão "LOWESS" para estimar a relação entre a média e a variância e transformar as contagens na escala log para obter uma transformação linear.
Abordagem geométrica (proposta)	"Reads" por milhão. A normalização utilizada na abordagem proposta nesta tese, chama-se "reads" por milhão, em inglês "Reads per million", (RPM). Esta normalização consiste em somar todos "reads" associados a um experimento i. Posteriormente, cada gene tem seu valor multiplicado por 1 milhão, seguido da divisão pela valor da soma anteriormente citada.

Fonte: adaptada da Ref. [21].

A normalização utilizada na abordagem proposta nesta tese, chama-se "reads" por milhão, em inglês "Reads per million", (RPM). Esta normalização consiste em somar todos "reads" associados a um experimento i. Posteriormente, cada gene tem seu valor multiplicado por 1 milhão, seguido da divisão pelo valor da soma anteriormente citada.

Informações resumidas sobre essa seção podem ser encontradas na Tabela 3.1.

3.6 Modelagem estatística de expressão gênica

Executada a normalização, a inferência de expressão diferencial pode ser realizada. Para efetuar tal tarefa, os pacotes assumem distribuições que os "reads" podem seguir para, posteriormente, realizam testes estatísticos. Nesta seção, as distri-

buições que tais pacotes utilizam e as modelagens estatísticas associadas a eles serão discutidas.

As distribuições estatísticas associadas a estes processos mais utilizadas são as distribuições de Poisson e binomial negativa, que são apresentadas no Apêndice **.

3.6.1 Distribuição de Poisson

Para discorrer sobre a distribuição de Poisson utilizou-se o material disponível na Ref. [32]. Este tipo de distribuição pode ser empregada como modelo em uma ampla gama de assuntos associados à espera de ocorrência, por exemplo, espera de chegada de um ônibus ou chegada de clientes em um banco. Para utilizar essa distribuição é necessário que a suposição associada a ela seja satisfeita: no caso do exemplo do banco, a probabilidade da chegada é proporcional ao tamanho do tempo de espera. Seja x uma variável aleatória de valores inteiros não negativos, tem-se uma distribuição de Poisson se

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \tag{3.3}$$

onde $x = 0, 1, ...$, e o parâmetro λ é algumas vezes chamado de parâmetro de intensidade. Neste tipo de distribuição a média de x é igual ao valor esperado,

$$E(X) = \lambda \tag{3.4}$$

e a variância é dada por

$$Var(X) = \lambda. \tag{3.5}$$

Ao utilizar essa distribuição supõe-se, então, que a variância é igual a média.

3.6.2 Distribuição binomial negativa

Assim como na dissertação sobre distribuição de Poisson, o material utilizado para discorrer sobre distribuição binomial negativa foi encontrado na Ref [32]. A distribuição binomial negativa deriva da binomial e por isso a exposição será iniciada pela distribuição binomial. Contudo, preliminarmente é necessário definir as tentativas de Bernoulli: número de sucessos numa sequência de n tentativas independentes, onde cada tentativa resulta somente nas possibilidades de sucesso ou fracasso. A distribuição binomial conta o número de sucessos numa sequência de n tentativas de Bernoulli. Seja X a sequência de tentativas independentes de Bernoulli, a qual denota a tentativa na

qual o r-ésimo sucesso ocorre, é dita seguir uma distribuição binomial negativa se (r, p):

$$P(X = x|r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \tag{3.6}$$

onde $x = r, r+1,...$ e r é um inteiro fixo. Pode-se definir a distribuição binomial negativa em termos da variável randômica Y = *número de falhas antes do r-ésimo sucesso*. Esta formulação é estatisticamente equivalente aquela dada em termos de X = *número de tentativas no qual o r-ésimo sucesso ocorre*, desde que $Y = X$ - r. Assim, usando a relação entre Y e X ($Y = X$ - r), a distribuição binomial também pode ser escrita como

$$P(Y = y) = (-1)^y \binom{-r}{y} p^r (1-p)^y. \tag{3.7}$$

Utilizando essa distribuição, o valor esperado de Y é dado pela equação:

$$E(Y) = r\frac{(1-p)}{p}, \tag{3.8}$$

e a variância:

$$Var(Y) = \frac{r(1-p)}{p^2}. \tag{3.9}$$

Se definimos a média como $\mu = \frac{r(1-p)}{p}$, consequentemente $E(Y) = \mu$, e a variância pode ser calculada como:

$$Var(Y) = \mu + \frac{1}{r}\mu^2. \tag{3.10}$$

Ao utilizar essa distribuição supõe-se, então, que a variância é uma função quadrática da média.

Como exposto na seção 3.2, os fragmentos de RNA a serem sequenciados são capturados pelos oligonucleotídeos poli-dT, e devido a nem todos serem capturados, os "reads" sequenciados representam uma amostragem do "estado real" de todos os fragmentos. Suponha que a mesma amostra seja sequenciada duas vezes, é plausível esperar a partir da constatação anterior que as contagens sejam ligeiramente diferentes. A razão desta ocorrência se deve ao número finito de "reads" que podem ser capturados pelo equipamento de sequenciamento e este é o motivo de obter-se uma amostragem do "estado real" da fonte a qual eles foram originados. Sabendo-se, então, que os experimentos de sequenciamento podem ser considerados como uma amostragem randômica produzida pelos "reads" a partir de um "pool" de fragmentos, a representação das contagens pode ser realizada por meio da distribuição de Poisson:

$$f(n, \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \tag{3.11}$$

onde n é o número da contagem dos "reads" associado ao um gene e λ é o valor esperado do número total de "reads" gerados pelos fragmentos dos transcritos que se alinhariam ao gene. A constatação de que até mesmo sequenciando duas amostras diferentes as contagens são diferentes pode ser associada a um ruído técnico, e este por sua vez pode ser denominado como "shot noise". Esta variabilidade na maioria das vezes pode ser bem associada ao ruído de Poisson em replicatas técnicas. Entretanto, existem casos que a variância nas contagens associadas a um gene na maioria das vezes é maior que a média, e tal circunstância não encoraja o uso da distribuição de Poisson, a qual é adequada a situações onde a variância é igual a média. Nesta caso, é apropriado o uso da distribuição binomial negativa, a qual também pode substituir a distribuição de Poisson no acaso anterior [33]. Na distribuição binomial negativa a variância é maior que a média e é calculada da seguinte forma:

$$\nu = \mu + \alpha\mu^2, \tag{3.12}$$

onde α é o fator de dispersão e μ é a média [22, 25].

A estimativa do fator α é uma das diferenças entre os pacotes edgeR e DESeq. O cálculo da estimativa realizado pelo DESeq decompõe a estimativa da variância em uma parte associada à estimativa da expressão média do gene, e a uma segunda parte que é associada à modelagem de um termo relacionado à variabilidade da expressão biológica. O cálculo da estimativa no edgeR é realizada a partir da combinação ponderada de dois componentes: um efeito da dispersão específica para cada gene e um efeito de dispersão comum que afeta todos os genes [22, 25]. A abordagem presente no PoissonSeq consiste em modelar as contagens dos genes N_{ij} como uma variável de Poisson. Neste caso, a média μ_{ij} da distribuição é representada por uma relação log linear $\log \mu_{ij} = \log d_j + \log \beta_i + \gamma_i \delta_j$, onde d_j representa o tamanho da biblioteca normalizada, β_i é o nível de expressão do gene i e γ_i é a correlação do gene i com a condição δ_j. Se não houver diferença significante na expressão do gene entre duas condições então γ_i é zero. O pacote baySeq implementa uma abordagem bayseana na modelagem por meio de distribuições binomiais negativas e os parâmetros da probabilidade a *priori* são estimados por amostragem numérica. Por fim, como dito, o pacote limma foi atualizado para analisar dados vindos de RNA-seq incorporando um método de normalização apropriado para inferir expressão diferencial a partir de modelos lineares [22, 25].

A abordagem proposta é discutida no manuscrito presente na referência [17]. Contudo, cabe informar que o método proposto não faz nenhuma suposição quanto possíveis distribuições que os "reads" podem seguir.

Tabela 3.2: Tabela comparativa dos métodos de modelagem estatística da expressão de genes executados pelos pacotes que serão estudados juntamente com as informações da abordagem geométrica.

Pacote	**Método**
baySeq	Utiliza uma abordagem bayseana na modelagem de distribuições binomiais negativas.
edgeR	Assume que os "reads" seguem uma distribuição binomial negativa.
DESeq	Assume que os "reads" seguem uma distribuição binomial negativa.
CuffDiff	Quando o gene possuir uma única isoforma é assumido que os genes seguem uma distribuição binomial negativa e no caso de houver múltiplas isoformas é usado uma distribuição binomial negativa com parâmetros da distribuição beta como pesos.
PoissonSeq	Modela as contagens dos genes N_{ij} como uma variável de Poisson.
limmaQN	Utiliza modelos lineares.
limmaVoom	Utiliza modelos lineares.
Abordagem proposta	As relevâncias dos genes são obtidas a partir das contagens sem nenhum tipo de suposição.

Fonte: adaptada da Ref. [21].

Informações resumidas sobre esta seção podem ser obtidas na Tabela 3.2.

3.7 Teste para expressão diferencial

A intenção deste capítulo é discorrer sobre os passos envolvidos para inferir genes expressos diferencialmente por meio de dados de RNA-seq. Tratou-se do sequenciamento das amostras, mapeamento, sumarização e as possíveis modelagens que podem ser assumidas após a sumarização. Esta seção trata propriamente da obtenção da lista de genes expressos diferencialmente. A inferência de expressão diferencial é obtida a partir dos cálculos dos parâmetros dos respectivos modelos estatísticos entre, dependendo da ferramenta computacional utilizada, duas condições ou mais. Para executar esta tarefa, o pacote limma usa um teste t moderado com erros padrões e graus de liberdades modificados para calcular o valor-p. Os pacotes edgeR e DESEq usam uma variação do teste exato de Fisher adequado à distribuição binomial negativa para calcular valores-p. O pacote baySeq realiza inferência de expressão diferencial por meio de uma abordagem bayseana. Nela, dois modelos são estimados, um assumindo que ocorre expressão diferencial e outro assumindo não que ocorre. Em seguida, uma função de verossimilhança adequada ao modelo é utilizada para identificar genes expressos diferencialmente. A forma de realizar inferência presente no PoissonSeq constitui-se na realização de um teste de significância do termo γ_i, o qual é definido como a correlação

Tabela 3.3: Tabela comparativa dos métodos de normalização executados pelos pacotes que serão estudados juntamente com as informações da abordagem geométrica.

Pacote	**Método**
baySeq	No algoritmo inicialmente é estimado dois modelos para todos os genes: um assumindo que não ocorre expressão diferencial e outro assumindo que existe, e em seguida a partir dos dados a função verossimilhança adequada ao modelo é usada para identificar os genes expressos diferencialmente.
edgeR	Variação do teste exato de Fisher.
DESeq	Variação do teste exato de Fisher.
PoissonSeq	Realiza um teste de significância do termo γ_i, que é a correlação da expressão do gene i entre duas condições, calculado a patir de "scores" estatísticos [25].
limmaVoom	Utiliza um teste t moderado com erros padrões e graus de liberdades modificados para calcular o valor-p [25].
Abordagem geométrica	As relevâncias dos genes são obtidas a partir das contagens sem nenhum tipo de suposição.

Fonte: adaptada da Ref. [21].

da expressão do gene i entre duas condições, calculado a partir de "scores" estatísticos. Todos os pacotes analisados aqui possuem implementados em seus respectivos algoritmos métodos padrões para correção de hipótese múltipla. A única exceção é o pacote PoissonSeq, o qual implementa uma nova forma de calcular a estimativa da taxa de falsa descoberta. As informações utilizadas nesta seção podem ser encontradas na Ref. [25]. Descrições sobre a metodologia da abordagem geométrica podem ser encontradas na referência [17]. Resumo sobre a discussão desta seção pode ser visto na Tabela 3.3.

Capítulo 4

Metodologia

Nesta seção a metodologia utilizada para análise de desempenho da inferência de expressão diferencial entre os pacotes DESeq, edgeR, PoissonSeq, limma-Voom, baySeq e a abordagem proposta nesta tese será descrita. A análise será realizada por meio de curvas ROC (do inglês, "Receiver Operation Characteristic"), medidas estatísticas como sensitividade, especificidade, acurácia e CCM (Coeficiente de Correlação de Matthews, no inglês, Matthews Correlation Coefficient). As informações contidas nesta seção também serão utilizadas na segunda parte da tese.

4.1 Curvas ROC

Os primeiros usos de curvas ROC datam da Segunda Guerra Mundial. Naquela época, elas eram utilizadas para quantificar a habilidade ("Receiver Operating Characteristic", ROC) dos operadores de radar ("receiver operators") em distinguir um sinal de um ruído. Em uma situação de guerra, tal medida se mostrou importante pois tornava possível medir a habilidade de um operador decidir corretamente se um sinal no radar era um avião inimigo (sinal) ou algum outro objetivo irrelevante (ruído). Posteriormente, as curvas ROC foram empregadas em psicologia experimental. Nos anos 70, foram amplamente usadas na classificação de pessoas doentes na área de pesquisa biomédica. Na década de 90, as curvas ROC passaram a ser empregadas em aprendizado de máquina ("machine learning") e desses estudos foi demonstrada a importância delas na avaliação e comparação de algoritmos [34].

Ao se estudar uma instância (gene) de um assunto (expressão diferencial), cada instância pode ser positiva p (expresso diferencialmente) ou negativa n (não expresso diferencialmente) gerando as classes $\vec{p}$ e $\vec{n}$. Com auxílio de um método de

inferência, as instâncias podem ser classificadas como positiva Y ou negativa N, gerando as classes preditivas $\vec{Y}$ e $\vec{N}$. A predição de cada instância pode ser realizada por um classificador por meio de valores contínuos e a predição de qual classe preditiva a instância pertencerá será feita usando um limiar. No caso de o classificador ser discreto, ele predirá a classe do elemento sem uso de limiar [35].

Para discorrer mais profundamente sobre curvas ROC, suponha a necessidade de predizer duas classes. Dado um classificador e uma instância, existem quatro desfechos possíveis:

- a instância é positiva e é classificada como positiva, assim é denominada como um verdadeiro positivo, VP;
- a instância é positiva e é classificada como negativa, ela é denominada como um falso negativo, FN;
- a instância é negativa e é classificada como negativa, ela é então denominada como um verdadeiro negativo, VN;
- a instância é negativa e é classificada como positiva, ela é então denominada como um falso positivo, FP.

Tais estabelecimentos podem ser utilizados nas definições abaixo [35]:

Taxa de verdadeiro positivo:

$$taxa\_vp = \frac{\text{positivos corretamente classificados}}{\text{total de positivos}}. \tag{4.1}$$

Taxa de falso positivo:

$$taxa\_fp = \frac{\text{negativos incorretamente classificados}}{\text{total de negativos}}. \tag{4.2}$$

Sensitividade = $taxa\_vp$, e,

$$\text{Especifidade} = \frac{\text{verdadeiros negativos}}{\text{falsos positivos} + \text{verdadeiros negativos}}. \tag{4.3}$$

A partir deste ponto faz-se necessário definir um espaço bidimensional ROC como sendo constituído pela taxa de verdadeiro positivo no eixo Y e a taxa de falso positivo sendo o eixo X. No momento que um classificador discreto é aplicado a um conjunto teste, ele produz somente uma classe de decisão Y ou N, gerando um ponto no espaço ROC. Contudo, um classificador probabilístico pode fornecer uma probabilidade

Tabela 4.1: Tabela de "scores" de um classificador probabilístico hipotético, onde p e n representam uma classe positiva e negativa, respectivamente

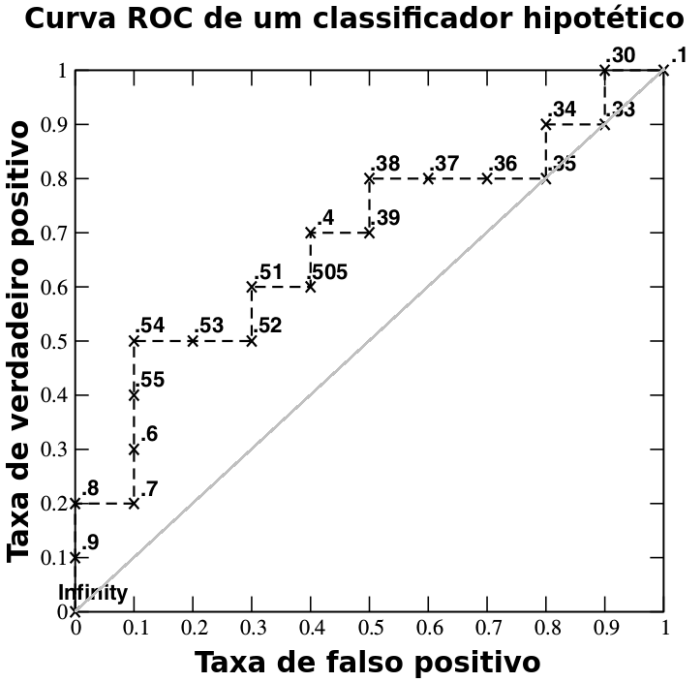
Instância	**Classe**	"Score"
1	**p**	0,9
2	**p**	0,8
3	**n**	0,7
4	**p**	0,6
5	**p**	0,55
6	**p**	0,54
7	**n**	0,53
8	**n**	0,52
9	**p**	0,51
10	**n**	0,505
11	**p**	0,4
12	**n**	0,39
13	**p**	0,38
14	**n**	0,37
15	**n**	0,36
16	**n**	0,35
17	**p**	0,34
18	**n**	0,33
19	**p**	0,30
20	**n**	0,1

Fonte: adaptada da Ref [35].

ou "score", indicando um valor numérico que representa o grau ao qual a instância é membro de uma classe. É oportuno neste ponto notar que existe diferença entre classificadores que geram probabilidades e os que geram "scores", embora possuam o mesmo nome. Um classificador gera probabilidade se responder estritamente aos teoremas padrões da probabilidade. Porém, se a saída do classificador não responder, ele então gera um "score", onde este último exemplo apenas apontará que um "score" mais alto corresponde a uma probabilidade mais alta [35].

A Tabela 4.1, adaptada da referência [35], lista os resultados de um classificador probabilístico hipotético. Na tabela, a primeira coluna **Instância** identifica a instância, a segunda coluna, nomeada de **Classe**, indica qual classe originalmente a instância pertence (p e n representa positiva e negativa, respectivamente) e a terceira e última coluna, **Score**, indica o "score" associado à classificação desta instância. Devido a forma de geração da curva *ROC*, os "scores" devem estar ordenados de forma decrescente. Uma curva ROC como a da Figura 4.1 produzida a partir dos dados da Tabela 4.1, se inicia considerando um nível auxiliar de $+\infty$ gerando o ponto (0;0). Este

Figura 4.1: Curva ROC gerada a partir da Tabela 4.1.



Fonte: adaptada da Ref [35].

nível decresce até encontrar a primeira instância positiva com "score" 0,9 gerando o ponto (0;0,1). Em um espaço bidimensional ROC, um ponto é definido como *(taxa de falso positivo; taxa de verdadeiro positivo)*. Dado que a primeira instância é positiva, então o valor 0,1 é atribuído ao eixo da taxa de verdadeiro positivo e o seu valor é obtido com o auxílio da equação 4.1 a partir do seguinte cálculo [35]:

$$taxa\_vp = \frac{1}{10} = 0,1. \tag{4.4}$$

O nível auxiliar é cada vez mais decrescido. A ideia por trás da construção da curva continua a mesma até todas as instâncias terem sido examinadas quando o nível chega a 0,1 e o ponto (1;1) é plotado. Qualquer curva ROC é, na verdade, uma função degrau, pois é gerada a partir de um conjunto finito de instâncias e ela se aproximará de uma curva verdadeira a medida que o número de instâncias se aproximar do infinito. Na Figura 4.1 ainda é possível ver uma linha auxiliar que representa o resultado que um classificador aleatório obteria [35].

A análise visual das performances de vários classificadores por meio de curvas pode se tornar subjetiva ou até mesmo desafiadora. Para contornar este problema, utiliza-se um método que reduz a curva à um valor escalar por meio do cálculo da área abaixo da curva 'ROC" (ou, abreviadamente, AAC). Dado que o valor gerado a partir

do cálculo está associado a uma porção da área de um quadrado unitário, seu valor sempre estará entre 0 e 1, sendo que um classificador randômico deve possuir um valor por volta de 0,5 e um classificador perfeito gera um igual a 1 [35].

4.2 Medidas estatísticas

A partir das definições da seção 4.1, é possível desdobrar as seguintes medidas estatísticas:

$$\text{Acurácia} = \frac{\text{Total de instâncias corretamente classificadas}}{\text{Total de classificações}}, \tag{4.5}$$

coeficiente de correlação de Mathews [36]:

$$CCM = \frac{(VP)(VN) - (FP)(FN)}{\sqrt{[VP + FP][VP + FN][VN + FP][VN + FN]}}, \tag{4.6}$$

onde, os valores podem ser de -1 (classificação foi completamente incorreta) a $+1$ (classificação totalmente correta). O CCM é um caso especial do coeficiente de correlação de Pearson [37]. Valor preditivo positivo [36]:

$$PPV = \frac{VP}{VP + FP}, \tag{4.7}$$

e medida F1 [38]:

$$F1 = \frac{2VP}{2VP + FN + FP}. \tag{4.8}$$

Para análise detalhada da performance essas últimas equações serão utilizadas juntamente com valores abaixo da curva ROC e com as seguintes medidadas já descritas na seção 4.1:

$$\text{Sensitividade} = \frac{\text{positivos corretamente classificados}}{\text{total de positivos}} = taxa\_vp, \tag{4.9}$$

e,

$$\text{Especifidade} = \frac{\text{verdadeiros negativos}}{\text{falsos positivos + verdadeiros negativos}}. \tag{4.10}$$

Parte II

Otimização da predição de epítopos lineares de células B

Capítulo 5

Introdução e Motivação

Estamos diariamente expostos a um número incontável de microrganismos, como bactérias, parasitas e vírus. Felizmente, existe o sistema imunológico (imunitário), o qual tem como função defender o organismo do ataque de agentes externos, principalmente prevenindo infecções e combatendo aquelas que já estão em curso. Portanto, tal sistema é de fundamental importância para nossa sobrevivência. Tal relevância também pode ser vista quando algum indivíduo apresenta decréscimo na qualidade de vida por ter sido acometido por doença ou falha nesse sistema.

A imunologia é o domínio da biologia que estuda o sistema imunológico. A palavra imunologia é derivada do latim *immunitas* ou *immunis* cujo significado é "isento de carga", onde, entre as possíveis referências à carga, a relevante para este estudo é enfermidade. Sendo assim, a derivação pode ser associada com o termo "isento de enfermidade". A definição moderna de imunologia, entre pequenas derivações, pode ser entendida utilizando a seguinte frase extraída da referência [39]: "é uma ciência experimental na qual explicações do fenômeno imunológico são baseadas em observações experimentais e as conclusões são obtidas a partir delas." A pesquisa associada ao sistema imunológico é importante para uma série de áreas, como, por exemplo: estudo de vacinas, tratamento de câncer, transplante de órgãos, tratamento de esclerose múltipla, doenças auto-imunes, entre outras.

Epítopos, também conhecidos como determinantes antigênicos, são definidos como fragmentos constituintes de um antígeno (substância que ao entrar em um organismo é capaz de iniciar uma resposta imune) que interagem com receptores de células B, T e anticorpos. No caso deste trabalho, os fragmentos são sequências de aminoácidos. A interação entre essas sequências peptídicas e receptores de células B ou receptores de células T é o principal passo na ativação de uma resposta imune adaptativa [40, 41]. Após a ativação da resposta imune adaptativa é possível que o sistema

imune crie memória, onde uma outra invasão será respondida de forma mais rápida. O sistema imune é composto por células de uma considerável gama de classes que atuam cooperativamente com certa redundância por meio de muitos mecanismos. De fato, trata-se de um sistema complexo e o foco do presente estudo, a predição de epítopos, é um assunto restrito dentro desta ampla área. Dessa forma, somente tópicos necessários para o entendimento do trabalho serão abordados.

A identificação de epítopos utilizando processos experimentais ainda permanece difícil devido às seguintes razões: 1) as técnicas ainda continuam incertas; 2) consomem muito tempo; e 3) é necessário um alto grau de sofisticação, de conhecimento e habilidades nos processos envolvidos [42, 43]. Métodos de predição possuem a capacidade de acelerar a descoberta, assumindo a forma de uma ferramenta de assistência às técnicas experimentais, sendo essa uma das razões para a aplicação de tais métodos [42]. Além dos entraves experimentais, existem casos onde as técnicas experimentais apresentam-se limitadas quando comparadas com à predição por ferramentas computacionais. Uma situação que representa essa constatação é exibido na referência [44], onde a análise de mais de 4.000 proteínas associadas à bactéria patogênica *M. tuberculosis* se tornaria inviável utilizando técnicas experimentais sem o auxílio de ferramentas de predição na descoberta de epítopos [44]. Devido ao potencial de aplicação de ferramentas preditivas, uma série de algoritmos, baseados em diferentes características físico-químicas e estruturais da proteínas-alvo, como hidrofobicidade, flexibilidade, acessibilidade e estrutura secundária, estão sendo produzidos ou aprimorados [44]. Contudo, embora existam muitas ferramentas para predição de epítopos, os preditores ainda possuem limitações [40, 45–47].

Dado o panorama exposto, a proposta da presente tese é a análise de possível melhora da performance de predição de epítopos lineares de células B por meio do método Suvrel. O trabalho apresentado por Marcelo Boareto et al. mostra que a metodologia chamada de Aprendizado Supervisionado de Relevância por Método Variacional ("Supervised Variational Relevance Learning", Suvrel), melhorou a performance de ferramentas computacionais de predição de dados originados de microarranjos e concentrações de metabólitos medidos por ressonância magnética nuclear [18]. O referido estudo apresenta um tensor métrico que pode ser usado para atribuir relevância estatística ao vetor de característica, e os resultados apresentados se baseiam na utilização da diagonal principal. Para realizar a análise aqui proposta, escolheu-se baseá-la na utilização do tensor métrico assim como resultados e dados fornecidos em um artigo onde o preditor *LBTOPE* [48] foi apresentado. Esta última escolha se deve ao número de sequências analisadas ser elevado ao ser comparado com estudos associados a outros preditores de epítopos lineares de células B. A escolha também se deve a característica

Tabela 5.1: Principais resultados apresentados no estudo do LBTOPE [48]. *Limiar* se refere a um limiar encontrado pelos autores que maximiza o Coeficiente de Correlação de Matthews (CCM), *Sensit.* se refere a sensitividade, *Espec.* se refere a especificade, *Acur.* se refere a acurácia e AAC se refere a área abaixo da curva *ROC*.

	Sensit.	**Especif.**	**Acurác.**	**CCM**	**AAC**
$LBTOPE\_FIX$	80,5	81,67	81,24	0,61	0,88
$LBTOPE\_VAR$	75,18	76,34	75,89	0,51	0,83

Fonte: adaptada da referência [48].

comum de o *Suvrel* ser geométrico e o algoritmo de *Máquinas de Vetores Suporte*, o qual o $LBTOPE$ se baseia, também ser. Na tabela 5.1 são mostrados os principais resultados apontados pelos autores presentes no referido artigo. $LBTOPE\_FIX$ representa o modelo gerado pelos autores que utiliza sequências de 20 aminoácidos de comprimento e $LBTOPE\_VAR$ é o modelo gerado quando o comprimento das sequências pode variar. A partir de tais resultados, é possível concluir que há espaço para aumento de performance de predição. Ainda assim, os preditores de epítopos lineares de células B estão tendo dificuldades de generalização. No caso deste preditor, embora possua bons resultados de performance, falha consideravelmente quando comparados com bancos de dados utilizados por outros preditores. Outros preditores também falham no conjunto de dados do $LBTOPE$ [48]. Dado o exposto, existe uma situação para análise de possível melhora do preditor $LBTOPE$ assim como seu poder de generalização.

Esta parte da tese está estruturada da seguinte forma. Inicialmente, uma visão geral do sistema imunológico é apresentada. Em seguida, descreve-se como é possível obter sequências imunogênicas, e posteriormente apresenta-se um dos algoritmos mais utilizados para redução de redundância, o *CD-hit*. Posteriormente descreve-se a Composição de Dipeptídeos, a qual foi utilizada para geração do vetor de características. Em seguida, tem-se a descrição do Suvrel e Máquinas de Vetores de Suporte. Por fim, os resultados e conclusões são apresentados. Optou-se por dividir a tese desta forma na intenção de seguir a estrutura dos passos associados à predição: obtenção das sequências, redução da redundância, mapeamento das sequência utilizando vetores de características, seleção de característica, treinamento e avaliação.

Capítulo 6

Objetivos

A presente pesquisa propõe-se a analisar a possível melhora da performance de predição de epítopos lineares de células B associada ao preditor *LBTOPE* por meio do método *Suvrel*. Para realizar tal objetivo, a análise se baseará nos resultados e dados apesentados em um estudo que gerou o preditor *LBTOPE*. A partir da exposição na seção 5, é possível concluir que há espaço para aumento de performance de predição. Soma-se também o fato de que o preditor *LBTOPE* falha na predição em dados de outros preditores assim como outros preditores falham em dados associados ao estudo *LBTOPE*. Assim, os objetivos desta parte da tese podem ser divididos em:

- Investigar a possível melhora de predição de epítopos lineares de células B utilizando os dados gerados pelos autores do *LBTOPE*;
- Análisar a possível melhora de predição comparada com o *LBTOPE* em dados de outros autores.

Capítulo 7

Sistema imunológico

7.1 Visão geral

Complementarmente ao exposto anteriormente, é possível inferir que a imunidade é uma derivação da palavra latina *immunitas*, a qual, no Império Romano, era usada para designar quando a colonia, comunidade ou indivíduo estavam protegidos (livres) da carga de impostos locais ou imperiais, de serviço nas legiões ou executar outras tarefas necessárias [49]. Seguindo esta descrição, pode-se dizer que imunidade é uma proteção, ou mais precisamente, uma reação contra microrganismo e seus componentes, proteínas, polissacarídeos e pequenos agentes químicos que são identificados como estranhos independentemente da consequência fisiológica ou patológica. As células e moléculas responsáveis pela imunidade constituem o sistema imune. O sistema imune é composto por células de uma considerável gama de classes que atuam cooperativamente com certa redundância por meio de muitos mecanismos. Devido a isso, somente será detalhado o que é necessário para o entendimento do restante do trabalho [39].

Existem dois tipos de imunidade: a inata (também chamada de natural ou nativa) e a adaptativa (também chamada de específica ou adquirida). A imunidade inata é constituída por células e barreiras físicas: pele, moléculas antimicrobianas, epitélio mucoso, proteínas do sistema do complemento, macrófagos, neutrófilos, células "natural killers" (NK), entre outras. Este tipo de imunidade é a 1ª linha de defesa contra microrganismos. Ela está presente mesmo antes da infecção ocorrer e é preparada para reagir rapidamente aos microrganismos e seus produtos fundamentalmente da mesma forma. A reação associada à imunidade natural está baseada em mecanismos específicos que interagem com estruturas que são comuns a grupos de microrganismos relacionados e podem não discernir pequenas diferenças [39].

A imunidade adaptativa é assim chamada pois desenvolve e adapta a resposta de defesa à infecção. Esta proteção é divida em dois tipos: 1) a imunidade humoral, composta pelos linfócitos B e seus produtos secretados (antianticorpos); 2) a imunidade celular, composta principalmente pelos linfócitos T auxiliares e citotóxicos. A ativação deste tipo de resposta está condicionada não exclusivamente, mas principalmente, à presença de antígenos. Estes últimos são definidos como substâncias estranhas que possuem a capacidade de se ligar a receptores específicos nos linfócitos, gerando ou não ativação da resposta imune adaptativa. Uma substância que é estranha ao organismo e ativa a resposta imune é chamada de imunógeno. A imunidade adaptativa reconhece um grande número de substâncias microbianas e não microbianas. Além disso, ela é específica, ou seja, possui a habilidade de distinguir entre diferentes substâncias. É possível que uma substância induza este tipo de resposta mais do que uma vez e, devido à memória associada a esta imunidade, cada nova exposição é atacada mais vigorosamente [39].

A resposta nativa e a adaptativa são partes de um sistema interligado de defesa. Os mecanismos de defesas iniciais estão à cargo da resposta inata e microrganismos que evoluíram seus mecanismos geram a necessidade de atuação da resposta adaptativa. Embora pareça que as atuações dos dois tipos de respostas sejam independentes, há conexões entre os dois sistemas. A resposta inata pode dirigir a natureza da resposta adaptativa assim como impulsioná-la, onde esta última, por sua vez, aumenta os mecanismos protetores da imunidade inata, tornando-os mais capazes de combater os microrganismos patogênicos. A imunidade que protege um indivíduo contra um microrganismo é geralmente induzida pela resposta do hospedeiro ao microrganismo. Dentro do âmbito da imunidade, diz-se que ela é ativa quando foi gerada pela exposição a um antígeno, onde, o indivíduo imunizado tem papel ativo na resposta ao antígeno. No caso de um indivíduo responder a um antígeno e permanecer protegido à exposições subsequentes, ele é considerado imune. Por fim, um indivíduo que não foi exposto a um antígeno particular é dito ser inativo ou não imune a esse antígeno [39].

O reconhecimento de um antígeno por parte da resposta adaptativa depende significativamente da interação entre sequências presentes no antígeno, conhecidas como epítopos, e os receptores presentes nas células T e B. Epítopos, também denominados como determinantes antigênicos, são definidos como fragmentos de aminoácidos associados a um antígeno que interagem com receptores de células B ou de células T por meio da complementariedade entre a região de ligação do receptor e o fragmento. Nesta tese, os fragmentos são sequências de aminoácidos associados à antígenos proteicos. A interação pode ou não resultar na ativação de uma resposta imune, sendo assim, o reconhecimento de epítopos o passo principal para ativação da resposta adaptativa [40, 41].

A seguir, o reconhecimento dos epítopos por partes das células B e T será detalhado.

7.2 Reconhecimento antigênico na imunidade humoral

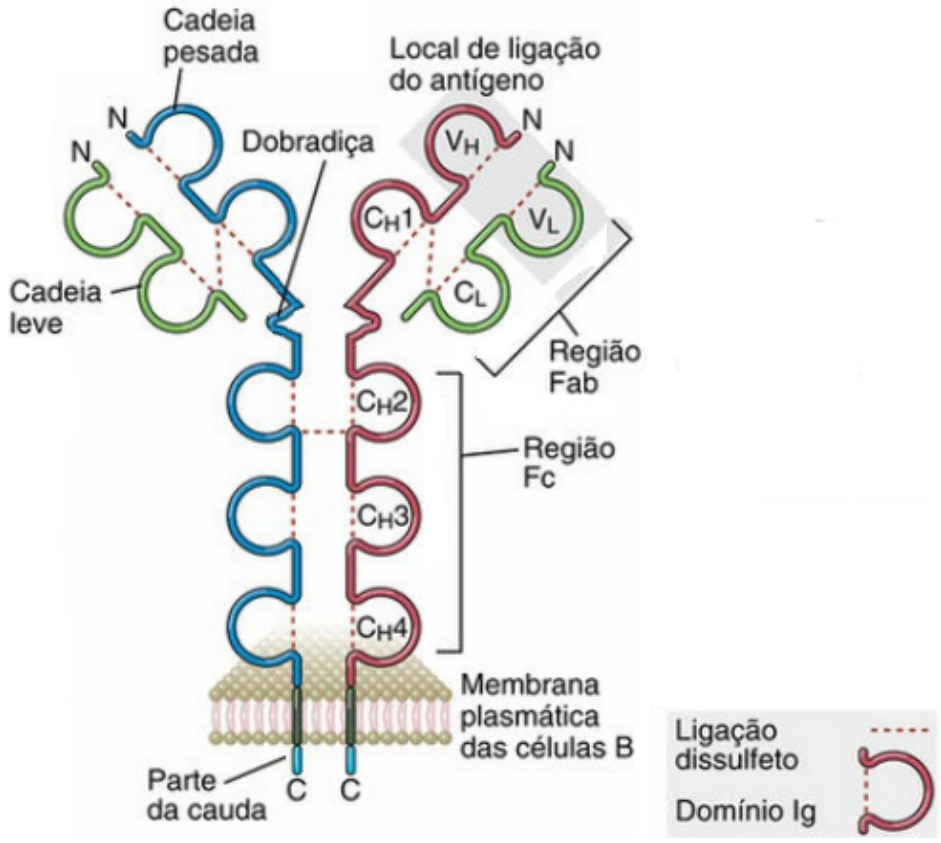
Como apresentado anteriormente, a imunidade humoral é composta pelos linfócitos B e pelos anticorpos. Além disso, é o mecanismo de defesa mais importante contra microrganismos extracelulares e suas toxinas. O linfócito B é produzido na medula óssea e é responsável pela secreção de anticorpo. Essas moléculas, dentro da referida importância, por meio de receptores de membrana (anticorpos), reconhecem antígenos extracelulares solúveis. Cada grupo de clone B tem, em sua membrana, anticorpos que reconhecem antígenos com uma única especificidade. Por sua vez, outros clones terão diferentes especificidades, dessarte, diz-se que os receptores são clonalmente distribuídos, permitindo que o organismo reconheça e responda a milhões de antígenos estranhos. A partir do reconhecimento, as células B imaturas tornam-se ativas e, uma vez ativadas, elas se diferenciam em plasmócitos que secretam anticorpos de mesma especificidade. Uma vez secretados, os anticorpos podem ser encontrados no plasma, no fluido intersticial dos tecidos e nas secreções mucosas. Sendo assim, os anticorpos podem ser encontrados ligados à membrana dos linfócitos B ou na forma secretada, onde, nesta última forma atuam na neutralização de toxinas, na prevenção da entrada e espalhamento dos patógenos e eliminação de microrganismos [39].

Os anticorpos existem em grande diversidade, o que proporciona a eles alta especificidade no reconhecimento de estruturas moleculares estranhas. Sabendo que as classes destas moléculas partilham semelhanças em características estruturais básicas (regiões constantes, F_c) e investigando estritamente o reconhecimento de epítopos, serão descritos os detalhes sobre a forma do anticorpo ligado à membrana do linfócito B (Figura 7.1) [39].

Um anticorpo é composto de duas cadeias leves idênticas (segmentos verdes na Figura 7.1) e duas cadeias pesadas idênticas (segmentos azul e vermelho na Figura 7.1). A molécula em questão é composta por unidades homólogas repetidas com cerca de 110 resíduos dobradas de forma independente em um motivo globular, conhecidas como domínios Ig. Cada domínio Ig contém duas camadas de folhas β-pregueadas, onde cada camada é composta por cerca de três a cinco fitas de cadeia antiparalela. Por fim, as duas camadas são mantidas unidas por ponte dissulfeto e faixas adjacentes de cada folha β são conectadas por pequenas alças. A justaposição dos domínios V_L e V_H é o local onde ocorre o reconhecimento antigênico através da interação com o epítopo

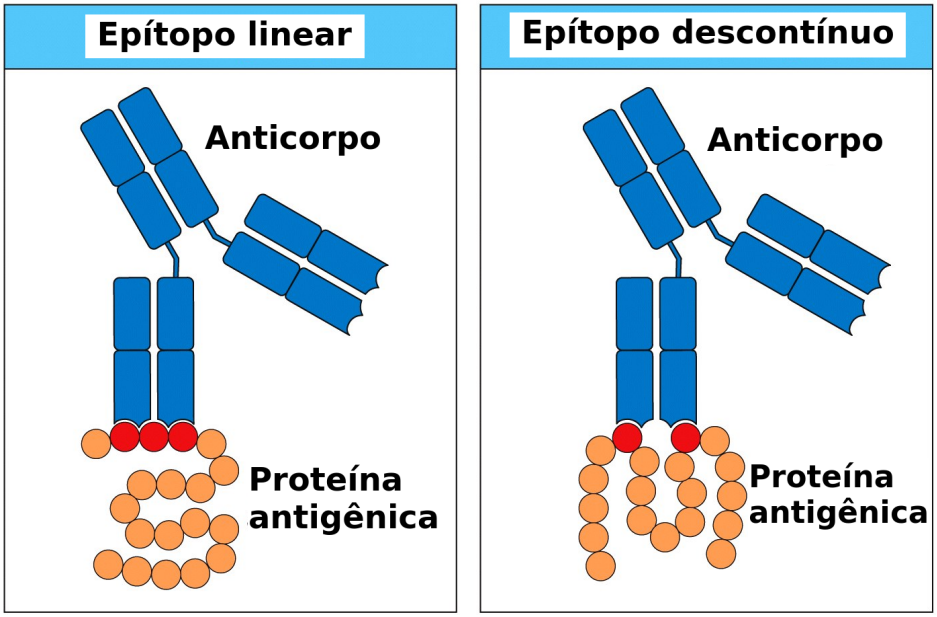
(Figura 7.1 e 7.2). A alta especificidade de ligação entre anticorpos e antígenos é atribuída à região onde ocorre o reconhecimento. A grande variabilidade nos domínios V_L e V_H é o motivo pelo qual diferentes anticorpos se ligam a um grande número de antígenos estruturalmente variados. Por fim, funções efetoras e as propriedades físico-químicas dos anticorpos estão associadas às regiões constantes que constituem uma dada classe, as quais apresentam relativamente poucas variações entre diferentes anticorpos.

Figura 7.1: Estrutura do anticorpo ligado à membrana da célula B. Um anticorpo é composto de duas cadeias leves idênticas e duas cadeias pesadas idênticas. A molécula em questão é composta por unidades homólogas repetidas, com cerca de 110 resíduos, dobradas de forma independente em um motivo globular, conhecidas como domínios Ig. Cada domínio Ig contém duas camadas de folhas β-pregueadas, onde cada camada é composta por cerca de três a cinco fitas de cadeia antiparalela. Por fim, as duas camadas são mantidas unidas por ponte dissulfeto e faixas adjacentes de cada folha β são conectadas por pequenas alças. A justaposição dos domínios V_L e V_H é o local onde ocorre a ligação antigênica.



Fonte: obtida da referência [39].

Figura 7.2: Representação da interação entre anticorpos e os dois tipos de epítopos. Do lado esquerdo é mostrada a interação entre a região variável de um anticorpo e um epítopo linear. Do lado direito é mostrada a interação entre a região variável e um epítopo descontínuo.



Fonte : adaptada da Ref. [50].

7.3 Reconhecimento antigênico na imunidade adaptativa celular

A imunidade adaptativa celular, também denominada imunidade celular, é constituída pelos linfócitos T, moléculas do complexo principal de histocompatibilidade (MHC no idioma inglês, "major histocompatibility complex") e, por fim, as células apresentadoras de antígenos (APC no idioma inglês, "antigen-presenting cell"). Existem situações onde microrganismos, por exemplo vírus e algumas bactérias, se alojam e se proliferam dentro de fagócitos ou outras células e a defesa fica a cargo deste domínio de imunidade. A defesa associada à imunidade celular também pode, por meio dos linfócitos T, atacar microrganismos extracelulares por meio do recrutamento de leucócitos ou auxiliando as células B na produção de anticorpos.

Os linfócitos T são distribuídos dentro de populações com funções diferentes, onde, as mais relevantes para este estudo são associadas as auxiliares (também denominados CD4+) e as citotóxicas (CD8+). A classificação CD vem do método de nomenclatura do agrupamento de diferenciação (do idioma inglês, "cluster of differentiation"). Este método utiliza moléculas da superfície celular características de uma linhagem celular em particular ou que diferenciam estágios. Dessa forma, a maioria

Tabela 7.1: Características das moléculas do MHC da classe I e da classe II.

Característica	**MHC de classe I**	**MHC de classe II**
Cadeias polipeptídicas	α, microglobulina β_2	α e β
Localização dos resíduos polimórficos	Domínios α_1 e α_2	Dom ínios α_1 e β_1
Local de ligação do correceptor da célula T	CD8 se liga principalmente ao domínio α_3	CD4 se liga a um bolsão criado por partes dos domínios α_2 e β_2
Tamanho da fenda de ligação do peptídio	Acomoda peptídios de 8-11 resíduos	Acomoda peptídios de 10-30 resíduos ou mais
Nomenclatura		
Humanos	HLA-A, HLA-B, HLA-C	HLA-DR, HLA-DQ, HLA-DP
Camundongos	H-2K, H-2D, H-2L	I-A, I-E

Fonte: obtida da Ref. [51].

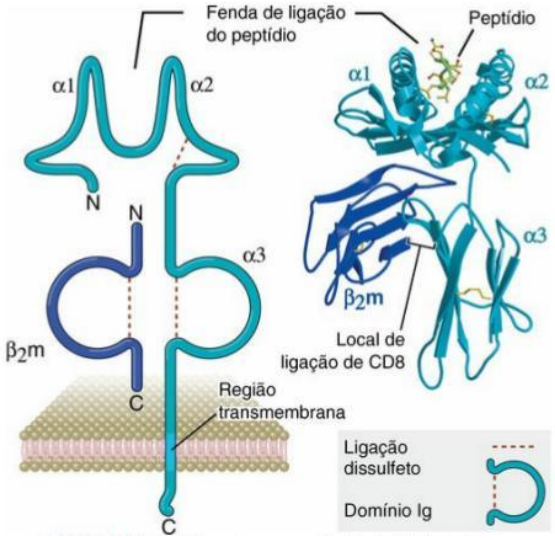
dos linfócitos T auxiliares CD4+ expressam a proteína CD4 na superfície celular e essa proteína é utilizada para distinguir tais linfócitos, analogamente aos linfócitos T citotóxicos. A partir de um reconhecimento antigênico, é possível que os linfócitos auxiliares secretem citocinas, as quais são responsáveis por muitas das respostas celulares da imunidade inata e adaptativa. Dentre essas respostas, as citocinas estimulam a diferenciação e proliferação das próprias células T, células B, macrófagos e outros leucócitos. Quando um linfócito T citotóxico é estimulado por um antígeno, ele pode induzir células infectadas por vírus e microrganismo intracelulares à morte.

Os linfócitos T expressam, em suas membranas, receptores de célula T (TCR no idioma inglês, "T-cell receptor") $\gamma\beta$, os quais atuam como mediadores da imunidade adaptativa celular.

A captura, processamento e apresentação dos antígenos é feita pelas células apresentadoras de antígenos. As células dendríticas, por exemplo, capturam antígenos microbianos presentes no exterior das células, se deslocam até os órgãos linfoides e apresentam os antígenos por meio do complexo principal de histocompatibilidade (MHC do idioma inglês, "major histocompatibility complex") aos linfócitos T imaturos.

As moléculas do MHC formam um conjunto de proteínas nas membranas das células apresentadoras de antígenos que possuem a função de ligar antígenos e apresentá-los aos linfócitos T. Dentro deste trabalho, a atenção deve ser dada às moléculas do MHC da classe I e da classe II, onde as principais características podem ser visualizadas na Tabela 7.1. A molécula de classe I é constituída por uma cadeia α polimórfica

Figura 7.3: A molécula de classe I é constituída por uma cadeia α polimórfica ligada de forma não covalente a uma cadeia β polimórfica. À direita é mostrado um "cartoon" de fitas da porção extracelular de uma molécula do MHC de classe I (HLA-B27).



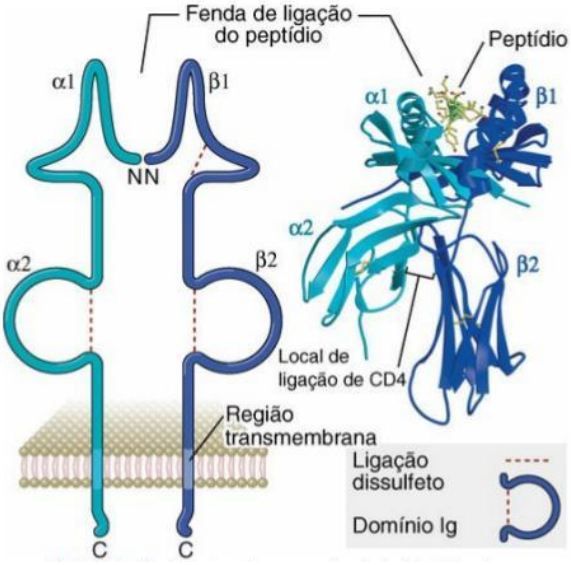
Fonte: obtida da Ref. [51].

ligada de forma não covalente a uma cadeia β polimórfica e uma fenda de ligação do peptídio (Figura 7.3). A molécula do MHC de classe II é constituída também por cadeia α polimórfica ligada de forma não covalente a uma cadeia β e uma fenda de ligação do peptídio (Figura 7.4).

O principal passo na ativação da resposta imune adaptativa ocorre quando uma molécula apresentadora após ter processado um dado antígeno (no caso desta tese, antígeno proteico), apresenta uma sequência peptídica por meio do MHC ao receptor da célula T (Figura 7.5). A partir da ativação, uma série de mecanismos efetores e de reforço da reposta também são ativados.

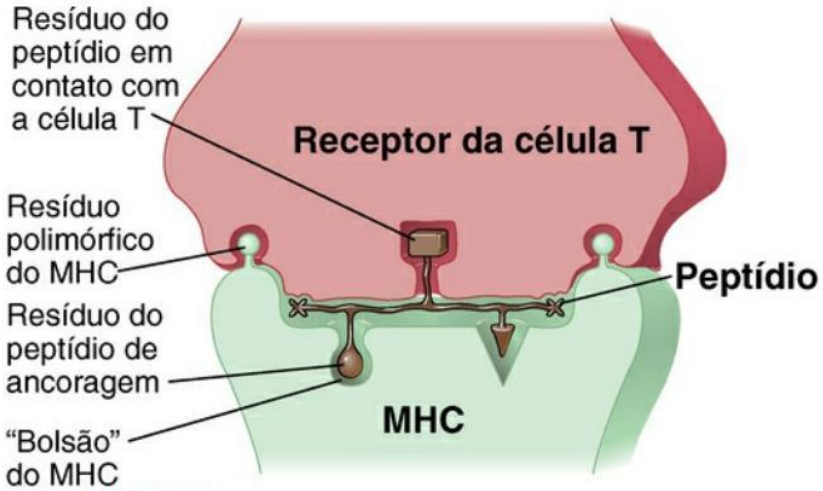
Cada clone T tem em sua membrana receptores que reconhecem antígenos com uma única especificidade. Por sua vez, outros clones terão diferentes especificidades, dessarte, diz-se que os receptores são clonalmente distribuídos, permitindo que o organismo reconheça e responda a milhões de antígenos estranhos.

Figura 7.4: A molécula do MHC de classe II é constituída por cadeia α polimórfica ligada de forma não covalente a uma cadeia β e uma fenda de ligação do peptídio. À direita é mostrado um "cartoon" de fitas da porção extracelular de uma molécula do MHC de clase II (HLA-DR1).



Fonte: obtida da Ref. [51].

Figura 7.5: Representação esquemática do reconhecimento de um peptídio na imunidade adaptativa. Tal fenômeno ocorre quando uma molécula apresentadora, após ter processado um antígeno, apresenta uma sequência peptídica por meio do MHC ao receptor da célula T.



Fonte: obtida de Ref. [51].

Capítulo 8

Metodologia

8.1 Banco de dados - IEDB

O IEDB ("Immune Epitope Database and Analysis Resource") é uma base de dados gratuita, fundada pelo Instituto Nacional de Alergia e Doenças Infecciosas dos Estados Unidos. Neste banco de dados, é possível pesquisar dados experimentais de caracterização de anticorpos e epítopos de células T e B de humanos, primatas e outras espécies, além de epítopos envolvidos em doenças infecciosas, alergia, autoimunidade e transplantes. A Tabela 8.1 [52] mostra informações sobre este banco de dados.

Tabela 8.1: Resumo dos dados contidos no IEDB em 2019.

Epítopos peptídicos	582.306
Epítopos não-peptídicos	2.762
Ensaios de células T	355.740
Ensaios de células B	477.857
Ensaios de ligantes de MHC	1.184.612
Organismos fontes de epítopos	3.698
Alelos de restrição de MHC	783
Referências	20.0327

Fonte: dados obtidos na Ref. [52].

8.2 Redução da redundância

A partir da obtenção do conjunto de sequências imunogênicas e não imunogênicas, é necessário se atentar à possibilidade de redundância. Para evitar tal pro-

blema, utilizam-se algoritmos. Na maior parte da literatura a redução de redundância é realizada utilizado a ferramenta computacional chamada CD-HIT [53, 54]. Tal algoritmo incrementalmente inicia a análise a partir da maior sequência, definindo-a como representativa e, então, processa as sequências subsequentes da maior para a menor, classificando-as como redundantes ou representativas por meio da análise de limiares de identidade com outras sequências representativas. O usuário pode definir limiares de identidade das sequências dentro dos valores de 0.4 a 1.0. Também é necessário definir os tamanhos das "palavras" (*k-mers*) que o algoritmo utiliza internamente, os quais são definidos a partir dos limiares de identidade (Tabela 8.2).

Tabela 8.2: Tabela de parâmetros do programa cd-hit.

Tamanho da palavra	**Limiar de identidade**
5	0.7 - 1.0
4	0.6 - 0.7
3	0.5 - 0.6
2	0.4 - 0.5

Fonte: guia do usuário do CD-hit disponível em https://github.com/weizhongli/cdhit.

8.3 Descritor - Composição de Dipeptídeos

A composição de dipeptídeos [55] transcreverá uma sequência de aminoácido por meio da fração de cada par de aminoácido do tipo i, j presente nela, ou seja:

$$CD = \left(\frac{N_{i,j}}{N-1}\right)_{i,j=1,2,3,\ldots,20}, \tag{8.1}$$

onde N é o tamanho da sequência e $N_{i,j}$ é a quantidade de pares de aminoácidos i, j.

8.4 Seleção de características - Método *Suvrel*

Os passos para obtenção do tensor métrico ligado a características de interesse (por exemplo, pares associados à Composição de Dipeptídeos) seguem a referência [18], que origina o Método Suvrel. Os passos são estruturados a partir da criação de uma métrica seguida pela definição de uma função custo, tornando possível o cálculo do tensor métrico. Afim de iniciar a metodologia, define-se a distância d_{ij}^2 entre dois

peptídios i e j como:

$$d_{ij}^2 = \sum_{\mu\nu} g_{\mu\nu} \Delta x_{ij\mu} \Delta x_{ij\nu}, \tag{8.2}$$

onde μ, ν se referem a características distintas e $\Delta x_{ij\nu} = x_{i\nu} - x_{j\nu}$. Em muitos estudos envolvendo experimentos associados a um conjunto de características, têm-se o objetivo de encontrar as mais relevantes. No contexto desta tese, o objetivo é obter os pares de peptídeos mais relevantes vindos da composição de dipeptídeo a partir de um conjunto de sequências imunogênicas e não-imunogênicas. Logo, a inserção da variável $g_{\mu\nu}$ no cálculo da distância, que ainda necessita ser definida, está associada a esta constatação e é ela que fornecerá o grau de relevância de um determinado par de características.

Para continuar o desenvolvimento de uma metodologia que forneça tais valores, define-se uma função objetivo:

$$E(\{g_{\mu\nu}\}; \mathcal{L}) = \sum_{a \in C} \langle d_{ij}^2 \rangle_{a,a} - \gamma \sum_{\langle a \neq b \rangle} \langle d_{ij}^2 \rangle_{a,b}, \tag{8.3}$$

onde $\mathcal{L}$ se refere ao conjunto treinamento, γ está associado ao peso das distâncias interclasses (imunogênicas e não imunogênicas) comparadas com as intraclasses, a e b estão associados aos rótulos das classes, $\langle a \neq b \rangle$ designa que pares de classes diferentes são adicionados no cálculo somente uma vez e, por fim,

$$\langle d_{ij}^2 \rangle_{a,b} = \frac{1}{n_{ab}} \sum_{i \in a, j \in b} d_{ij}^2. \tag{8.4}$$

Na equação 8.3, o primeiro termo está associado às distâncias das sequências dentro da mesma condição e o segundo termo calcula as distâncias entre condições diferentes. A Equação 8.3 pode ser reescrita como:

$$E(\{g_{\mu\nu}\}; \mathcal{L}) = \sum_{\mu\nu} g_{\mu\nu} \epsilon_{\mu\nu}, \tag{8.5}$$

onde,

$$\epsilon_{\mu\nu} = e_{\mu\nu}^{dentro} - \gamma e_{\mu\nu}^{fora}. \tag{8.6}$$

Na Equação 8.6, o primeiro termo pode ser escrito como:

$$e_{\mu\nu}^{dentro} = \sum_{a\in C} \frac{1}{n_{aa}} \sum_{i,j\in a} \Delta x_{ij\mu} \Delta x_{ij\nu}$$

$$= \sum_{a\in C} \frac{1}{n_a^2} \left(n_a \sum_{i\in a} x_{i\mu} x_{i\nu} + n_a \sum_{j\in a} x_{j\mu} x_{j\nu} - 2 \sum_{i\in a} x_{i\mu} \sum_{j\in a} x_{j\nu} \right)$$

$$= 2 \sum_{a\in C} \mathrm{cov}(\vec{x}_\mu^a; \vec{x}_\nu^a). \tag{8.7}$$

Na Equação 8.6, o segundo termo pode ser escrito como:

$$e_{\mu\nu}^{fora} = \sum_{\langle a\neq b\rangle\in C} \frac{1}{n_{ab}} \sum_{i\in a, j\in b} \Delta x_{ij\mu} \Delta x_{ij\nu}$$

$$= \sum_{\langle a\neq b\rangle\in C} \frac{1}{n_a n_b} \left(n_b \sum_{i\in a} x_{iu} x_{i\nu} + n_a \sum_{j\in b} x_{j\mu} x_{j\nu} - \sum_{i\in a} x_{i\mu} \sum_{j\in b} x_{j\nu} - \sum_{i\in a} x_{i\nu} \sum_{j\in b} x_{j\mu} \right)$$

$$= \sum_{\langle a\neq b\rangle\in C} [\mathrm{cov}(\vec{x}_\mu^a; \vec{x}_\nu^a) + \mathrm{cov}(\vec{x}_\mu^b; \vec{x}_\nu^b)] + (m_{a\mu} - m_{b\mu})(m_{a\nu} - m_{b\nu})$$

$$= (K-1) \sum_{a\in C} \mathrm{cov}(\vec{x}_\mu^a; \vec{x}_\nu^a) + \sum_{\langle a\neq b\rangle\in C} (m_{a\mu} - m_{b\mu})(m_{a\nu} - m_{b\nu})$$

$$= (K-1) \sum_{a\in C} \mathrm{cov}(\vec{x}_\mu^a; \vec{x}_\nu^a) + K^2 \mathrm{cov}(\vec{m}_u; \vec{m}_\nu) \tag{8.8}$$

K é o número total de classes, $\vec{x}_{mu}^a$ é o conjunto de amostras associadas à característica μ que pertencem a classe a, $\vec{m}_\mu = \{m_{a\mu}\}$ e

$$\mathrm{cov}(\vec{m_u}; \vec{m_\nu}) = \frac{1}{K} \sum_{a\in C} m_{a\mu} m_{a\nu} - \left(\frac{1}{K} \sum_{a\in C} m_{a\mu} \right) \left(\frac{1}{K} \sum_{a\in C} m_{a\nu} \right)$$

Com o auxílio das Equações 8.7 e 8.8, é possível reescrever a Equação 8.6 como:

$$\epsilon_{\mu\nu} = [2 - (K-1)\gamma] \sum_{a\in C} \mathrm{cov}(\vec{x}_\mu^a; \vec{x}_\nu^a) - \gamma K^2 \mathrm{cov}(\vec{m}_u; \vec{m}_\nu). \tag{8.9}$$

É possível obter o tensor métrico $g_{\mu\nu}$ a partir da observação de que quando a Equação 8.3 for extrema e negativa, os peptídios são próximos entre a mesma condição e distantes entre condições diferentes, juntamente com o vínculo $\sum_{\mu\nu} g_{\mu\nu}^2 = 1$, forma-se um conjunto de observações que pode ser associado aos multiplicadores de Lagrange:

$$\frac{\partial}{\partial g_{\mu\nu}} = \left(E + \Theta \left(\sum_{\mu\nu} g_{\mu\nu}^2 - 1 \right) \right) = 0 \tag{8.10}$$

onde a solução é:

$$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu'\nu'} \epsilon^2_{\mu'\nu'}}}. \tag{8.11}$$

Para que o tensor métrico seja definido positivamente deve-se satisfazer $\gamma > \gamma^*$, onde $\gamma^* = \frac{2}{K-1}$. A aplicação do Método Suvrel baseia-se na transformação dos dados por meio de $\sqrt{g_{\mu\nu}}$, onde, após isto, as dimensões menos relevantes diminuem e as mais relevantes aumentam. Mais detalhes sobre os passos realizados para obter a Equação 8.11 podem ser obtidos nos apêndices A e C na Ref. [21].

8.5 Preditores

Na literatura, atualmente os preditores de epítopos estão estruturados na utilização de redes neurais, modelos de Markov ocultos e Máquinas de Vetores de Suporte (MVS, em inglês, "Support Vector Machines", SVM). Modelos de Markov ocultos e Máquinas de Vetores de Suporte são adequados quando existem informações estruturais e tem aplicações na predição de peptídios que se ligam à moléculas MCH-I. Redes neurais artificiais e MVS são treinadas e utilizadas na predição, por exemplo, de epítopos de células B. Breves informações adicionais de preditores podem ser visualizadas na Tabela 8.3 [44, 45].

8.5.1 Máquinas de Vetores de Suporte (Support Vector Machines)

Antes de discorrer sobre MVS, alguns aspectos sobre Teoria de Aprendizado Estatístico (TAE) serão abordados. Decidiu-se por tal abordagem baseada na Ref. [56] e por ela levar de forma fluída ao conceito de margem, base fundamental das MVS [57, 58].

Teoria de Aprendizado Estatístico

Inicialmente, no que diz respeito ao aprendizado de máquina, deve-se distinguir duas categorias denominadas de supervisionado e não-supervisionado. No aprendizado supervisionado, exemplos x_i são associados a um rótulo y_i, onde no contexto desta tese, $y_i \in \{+1, -1\}$. Em contrapartida, no aprendizado não-supervisionado os exemplos não possuem rótulos associados. No aprendizado supervisionado o objetivo é classificar de forma correta novas entradas, ao passo que no não-supervisionado o

Tabela 8.3: Lista dos preditores de epítopos de células B e T assim como as ferramentas usadas na predição.

Predição de epítopos de células T.	**Descrição**
MHCPred 20	Preditor de epítopos de células T de humanos e ratos.
MHC-THREAD	Preditor de peptídeos que se ligam à moléculas de MHC-II.
BIMAS	Preditor de epítopos que se ligam à moléculas de HLA.
PREDDEP	Preditor de epítopos que se ligam à moléculas de MHC-I.
RANKPEP	Preditor de peptídeos que se ligam à moléculas de MHC-I e MHC-II.
SVMHC	Preditor de peptídeos que se ligam à moléculas de MHC-I.
Imtech	Técnica de otimização de matriz para predição de epítopos que se ligam à MHC.
Predição de epítopos de células B.	
Bcepred	Preditor de epítopos baseado em propriedades físicas e químicas.
ABCpred	Preditor de epítopos que utiliza redes neurais artificiais.
Discotope	Preditor de epítopos descontínuos a partir da estrutura tridimensional de proteínas.
LBTOPE	Preditor de epítopos lineares de células B por meio de MVS.
BCpred	Implementa MVS por meio do emprego de uma série de métodos de kernel de *string*, radial e subsequência.
CEP	Preditor de epítopos conformacionais.

Fonte: adaptada da referência [44].

objetivo é analisar padrões e tendências. Dado que o objetivo do estudo proposto é predição, o restante do texto se baseará em aprendizado supervisionado. Neste tipo de tarefa, o objetivo por meio de treinamento, é gerar um classificador apto a predizer rótulos novos. De um ponto de vista mais rigoroso, é possível definir aprendizado supervisionado como um algoritmo que gera um classificador particular $\hat{f} \in F$, onde F é o conjunto dos classificadores possíveis de serem gerados pelo algoritmo, sendo este capaz de predizer corretamente novos rótulos. A fase de treinamento é definida como a geração do classificador particular $\hat{f}$, enquanto que a fase de teste corresponde à análise da capacidade de predizer rótulos novos [56].

Como dito anteriormente, o objetivo desta seção é discorrer sobre a im-

portância do conceito de margem e isto será feito por meio da análise de risco esperado, risco empírico e risco marginal. No sentido de iniciar a dissertação desta seção, assuma que os dados utilizados para o aprendizado são produzidos por uma distribuição de probabilidade $P(\vec{x}, y)$ desconhecida, a qual, por sua vez, gera os dados de forma independente e identicamente distribuída. O risco esperado $R(f)$ é uma medida de generalização de f, e este pode ser calculado a partir de um conjunto de dados teste por meio da Equação 8.12:

$$R(f) = \int c(f(\vec{x}, y))dP(\vec{x}, y), \tag{8.12}$$

onde c é uma função custo, e por exemplo pode ser calculada por meio da Equação 8.13:

$$c(f(\vec{x}, y)) = \frac{1}{2}|y - f(\vec{x})| = \begin{cases} 0, \ se\ \vec{x}\ foi\ corretamente\ classificado \\ 1, \ se\ \vec{x}\ foi\ incorretamente\ classificado. \end{cases} \tag{8.13}$$

A tarefa de buscar um classificador f poderia ser realizada por meio da minimização do risco esperado, porém a função $P(\vec{x}, y)$ geralmente é desconhecida.

Contudo, é possível continuar a dissertação por meio da análise do risco empírico, o qual é baseado na premissa de que a minimização dos erros no conjunto de treinamento minimiza os erros no conjunto teste. Tal risco pode ser calculado pela Equação 8.14:

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} c(f(\vec{x_i}, y_i)). \tag{8.14}$$

Quando $n \to \infty$, esta expressão converge para o risco esperado. Contudo, quando $n << \infty$ não existe convergência e este fato pode ocasionar a ocorrência de um modelo superajustado (com alta complexidade), onde o $R(f)_{emp} = 0$ e $R(f) = 0,5$. Dessa forma, a utilidade desta última equação se limita a conclusão de que deve-se levar em conta a complexidade da classe de extração a qual $\hat{f}$ será obtida. O limite abaixo (Equação 8.15) garantido com probabilidade $1 - \theta$ descreve esta conclusão:

$$R(f) \leqslant R_{emp}(f) + \sqrt{\frac{h(ln(\frac{2n}{h}) + 1) - ln(\frac{\theta}{4})}{n}}, \tag{8.15}$$

onde h é denominado de dimensão Vapnik-Chervonenkis (VC). Neste caso, do conjunto de funções F, n é o número de exemplos e o termo na raiz quadrada é denominado termo de capacidade. A dimensão VC é um índice escalar que mensura a capacidade (complexidade) do conjunto de funções F, onde, quanto maior seu valor, mais comple-

xos são os classificadores que podem ser obtidos a partir de F. Por exemplo, no caso de funções lineares no $\mathbb{R}^2$, a dimensão VC é 3 devido a este ser o número máximo de amostras que podem ser corretamente classificadas por uma reta para qualquer que seja a configuração associada a uma rotulação binária [59]. Analisando esta última equação conclui-se que a escolha adequada de $\hat{f}$ é aquela que minimiza o risco empírico $R(f)_{emp}$ e possui baixa dimensão VC (baixa complexidade). Contudo, a utilidade prática desta equação está limitada a esta conclusão, pois o cálculo de h não é trivial e ainda o seu valor pode ser infinito ou desconhecido.

Ainda assim, ao lançar mão de um classificador linear na forma $f(\vec{x}) = \vec{w} \cdot \vec{x}$, é possível continuar a dissertação desta seção por meio da relação entre erro marginal $R_p(f)$ e o risco esperado. Define-se a margem de confiança $\delta(f(\vec{x_i}, y))$ como:

$$\delta(f(\vec{x_i}, y)) = y_i f(x_i) \begin{cases} > 0 \ se \ y_i \ foi \ corretamente \ classificado \\ < 0 \ se \ y_i \ foi \ incorretamente \ classificado. \end{cases} \tag{8.16}$$

Com o auxílio da Equação 8.16 é possível definir o erro marginal $R_p(f)$ de uma função f (Equação 8.17) como a proporção de exemplos no treinamento que possuem a margem de confiança menor que um dado valor $p > 0$:

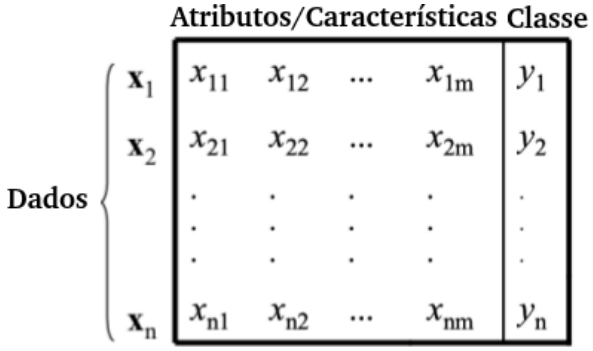
$$R_p(f) = \frac{1}{n} \sum_{i}^{n} I(y_i f(\vec{x}_i) < p), \tag{8.17}$$

onde $I(q) = 1$ se verdadeiro e $I(q) = 0$ se falso. Com o auxílio desta última equação é possível atingir o objetivo proposto no início da seção por meio do limite da relação:

$$R(f) \leqslant R_p(f) + \sqrt{\frac{c}{n}(\frac{R^2}{p^2} log^2(\frac{n}{p}) + log(\frac{1}{\theta})} \tag{8.18}$$

garantido com probabilidade $1 - \theta \in [0, 1]$. Nesta última relação o segundo termo do lado direito está associado à complexidade do conjunto de funções F, c é uma constante, $p > 0$, F esta associado à classe de funções lineares $f(\vec{x}) = \vec{w} \cdot \vec{x}$ com $|x| \leqslant R$ e $|w| \leqslant 1$. A análise desta última relação mostra que com maior valor de p tem-se maior $R_p(f)$ e menor complexidade, ao passo que com menor valor de p tem-se complexidade maior e $R_p(f)$ menor. Portanto, a minimização do risco esperado pode ser buscada por meio de um classificador que siga um compromisso entre maximização da margem e minimização de erro marginal. Esta conclusão leva a definição de hiperplano separador ótimo, sendo este definido como aquele que minimiza o erro sobre os dados de treinamento e teste e possui margem p elevada.

Figura 8.1: Representação gráfica das definições de dados, atributos e classes.



Fonte: adaptada da Ref. [56].

Máquinas de Vetores de Suporte

Para a produção desta seção utilizou-se o material apresentado em [56] como referência. Porém, existem situações que recorreu-se a outras fontes de informações complementares e nesses casos a origem será indicada.

Para iniciar a dissertação sobre Máquinas de Vetores de Suporte (MVS), define-se X como o espaço dos dados, Y como sendo os rótulos das classes $Y = \{+1, -1\}$, T um conjunto de treinamento com n dados $\vec{x_i} \in X$ e seus respectivos rótulos $y_i \in Y$. Uma representação gráfica pode ser vista na Figura 8.1.

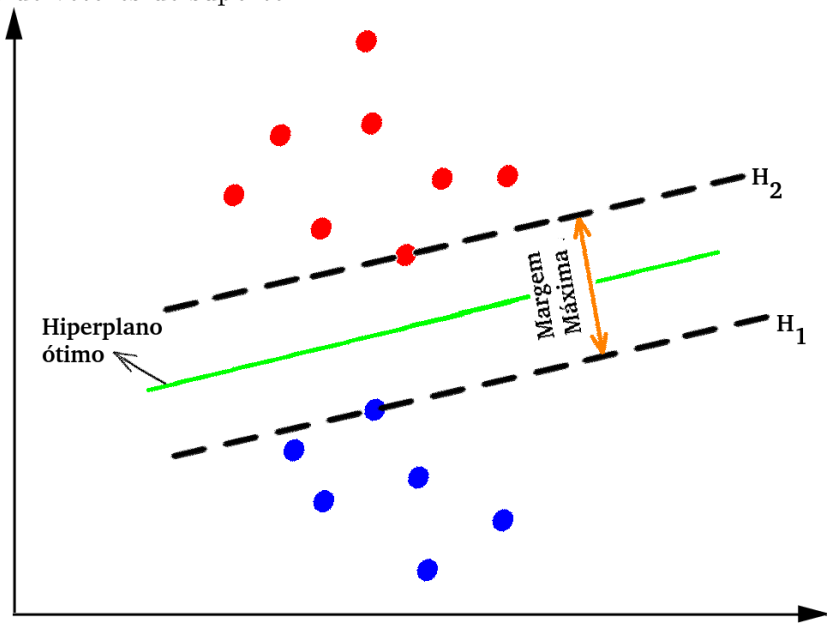
No caso de se analisar um conjunto de dados linearmente separável, constituído somente por dois atributos, classifica-se os dados utilizando um classificador linear. Complementarmente, no caso da existência de três atributos o classificador separa os dados por meio de um plano e por fim, quando existe mais do que três o classificador separa os dados por meio de um hiperplano. Portanto, no decorrer do texto será utilizado o termo hiperplano de forma genérica para os três possíveis casos citados. Dessa forma, um hiperplano pode ser definido como:

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b = 0, \tag{8.19}$$

onde $\vec{w} \cdot \vec{x}$ é o produto escalar entre os vetores $\vec{w}$ e $\vec{x}$, $\vec{w} \in X$ é o vetor normal ao hiperplano e $\frac{b}{\|\vec{w}\|}$ é a distância do hiperplano em relação à origem com $b \in \mathbb{R}$.

Na Figura 8.2 é apresentado um conjunto de dados linearmente separável, representado por círculos azuis associados a uma dada classe e círculos vermelhos associados a outra classe. Uma MVS é um classificador que define um hiperplano ótimo a partir da maximização da distância entre dois hiperplanos H_1 e H_2. A distância entre os hiperplanos é definida como margem. Desde que existem diversos hiperplanos

Figura 8.2: Representação de dados linearmente separáveis, margem máxima e hiperplano ótimo. H_1 e H_2 representam dois hiperplanos, onde a região entre eles representa a margem máxima e em verde o hiperplano ótimo. Os pontos onde os hiperplanos H_1 e H_2 os cortam são chamados de vetores de suporte.



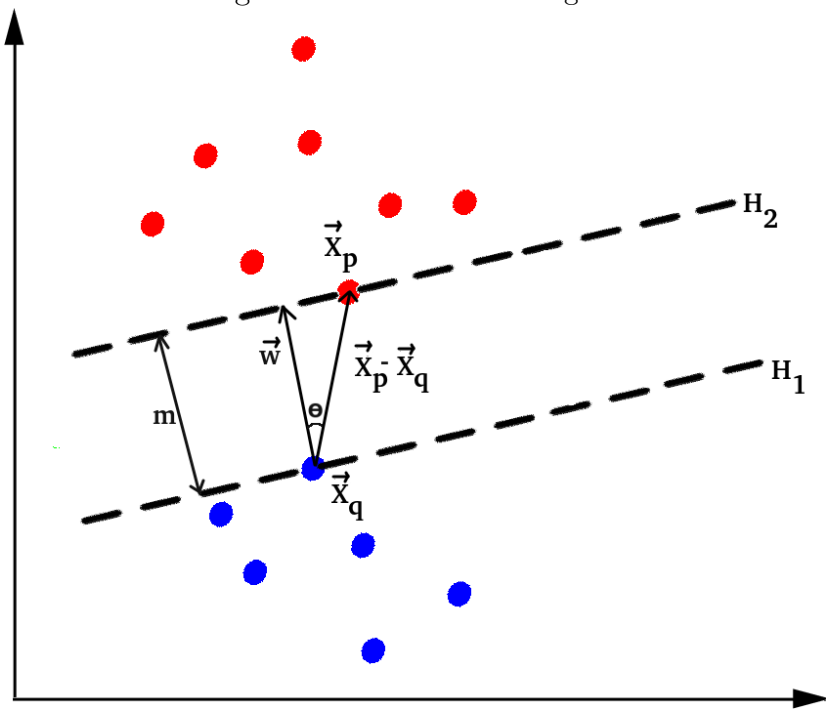
Fonte: elaborada pelo autor.

que dividem as classes, a definição de um hiperplano dessa forma abarca as possíveis variações das classes, e assim, minimiza o risco esperado (algumas vezes nomeado como erro de generalização) [60]. Analisando ainda a Figura 8.2, vê-se que não existe nenhum dado entre a margem e este caso mais simples será utilizado para iniciar a dissertação, ou seja, dados linearmente separáveis e sem dados entre a margem.

A partir do que foi exposto até aqui, vê-se que hiperplano separador ótimo é obtido a partir da restrição de não haver dados entre a margem. O primeiro passo para calcular este hiperplano será o de encontrar uma expressão que represente esta restrição. Considere que, na Figura 8.2, o hiperplano H_0 (em verde) separa o conjunto de dados e satisfaz $\vec{w}\cdot\vec{x_i}+b=0$. Complementarmente, considere ainda que o hiperplano H_1 satisfaz $\vec{w}\cdot\vec{x_i}+b=+1$ e H_2 satisfaz $\vec{w}\cdot\vec{x_i}+b=-1$. A expressão associada à restrição pode ser obtida a partir do escalonamento de $\vec{w}$ e b de forma que os dados mais próximos ao hiperplano separador satisfaçam a equação $|\vec{w}\cdot\vec{x_i}+b|=1$. Esses últimos passos implicam nas inequações:

$$\begin{cases} \vec{w}\cdot\vec{x_i}+b\geq+1, se\ y_i=+1, \\ \vec{w}\cdot\vec{x_i}+b\leq-1, se\ y_i=-1. \end{cases}$$

Figura 8.3: Cálculo da margem.



Fonte: adaptada da Ref. [61].

Por fim, pode-se aglutinar essas duas últimas inequações na seguinte inequação:

$$y_i(\vec{w} \cdot \vec{x_i} + b) - 1 \geq 0, \forall(\vec{x_i}, y_i) \in T. \tag{8.20}$$

Tem-se, então, a inequação associada a restrição de não ser permitido a existência de pontos entre as margens.

Para calcular a margem considere a Figura 8.3. A partir da análise desta última figura é possível calcular a margem m:

$$m = \|\vec{x}_p - \vec{x}_q\| cos(\theta) \tag{8.21}$$

$$m = \vec{w} \cdot \frac{(\vec{x}_p - \vec{x}_q)}{\|\vec{w}\|} \tag{8.22}$$

$$m = \frac{\vec{w} \cdot \vec{x}_p - \vec{w} \cdot \vec{x}_q}{\|\vec{w}\|} \tag{8.23}$$

$$m = \frac{b + 1 - b + 1}{\|\vec{w}\|} \tag{8.24}$$

$$m = \frac{2}{\|\vec{w}\|} \tag{8.25}$$

Dado o exposto até aqui, o hiperplano separador ótimo, então, pode ser obtido a partir da maximização de $\frac{2}{\|\vec{w}\|}$ respeitando as restrições dadas pela equação 8.20.

Problemas deste tipo podem ser resolvidos utilizando uma Função Lagrangiana, a qual relaciona a minimização de $\frac{1}{2}\|\vec{w}\|^2$ por meio de parâmetros denominados multiplicadores de Lagrange (α). A conversão de um problema de maximização em minimização e o quadramento de $\|\vec{w}\|$ não alteram a natureza do problema e tais ações são realizadas para que a solução tenha um mínimo global único e o problema se torne mais adequado às ferramentas de resolução [61, 62]. A função Lagrangiana pode ser escrita como

$$L(\vec{w}, b, \alpha) = \frac{\|\vec{w}\|^2}{2} - \sum_{i=1}^{n}(\alpha_i(y_i(\vec{w} \cdot \vec{x_i} + b) - 1)), \tag{8.26}$$

a qual deve ser minimizada, implicando em maximizar as variáveis α_i e minimizar $\vec{w}$ e b devido a solução deste tipo de problema ser um ponto de sela e assim pode-se escrever [63]:

$$\frac{\partial L}{\partial b} = 0, \tag{8.27}$$

$$\frac{\partial L}{\partial \vec{w}} = 0, \tag{8.28}$$

Calculando as Equações 8.27 e 8.28, tem-se:

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0, \tag{8.29}$$

$$\vec{w} = \sum_{i=1}^{n} \alpha_i y_i \vec{x_i}. \tag{8.30}$$

A formulação constituída pela minimização de $\vec{w}$ em $\frac{1}{2}\|\vec{w^2}\|$, respeitando as restrições dadas pela Equação 8.20, é denominada forma primal. Quando substitui-se as Equações 8.29 e 8.30 na Equação 8.26, o problema de otimização é então denominado dual, cuja formulação, pode ser vista nas expressões abaixo:

$$\underset{\vec{\alpha}}{\text{maximizar}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i \alpha_j y_i y_j (\vec{x_i} \cdot \vec{x_j})), \tag{8.31a}$$

$$\text{sujeito a} \qquad \alpha_i \geq 0, \forall i = 1, ..., n, , \tag{8.31b}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{8.31c}$$

A formulação dada dessa forma é interessante pois somente utiliza os dados de treinamento. O problema posto desta forma é um problema de otimização quadrático e um programa solucionador quadrático (no inglês, "quadratic programming solver") fornece os valores de α_i [64].

Assim, considere existir $\vec{\alpha}^*$ como sendo solução para o problema de otimização dual. Considere $\vec{w^*}$ como solução do problema primal, o qual pode ser determinado pela Equação 8.30. Por fim, o cálculo do parâmetro b^*, solução do problema primal, depende de $\vec{\alpha}^*$ e de uma das condições de Kuhn-Tucker (Equação 8.32). Tais condições são provenientes da teoria de otimização com restrições e que devem ser satisfeitas no ponto ótimo. Para o problema dual elaborado, tem-se:

$$\alpha_i^*(y_i(\vec{w^*} \cdot \vec{x_i} + b) - 1) = 0, \forall i = 1, ..., n. \tag{8.32}$$

Analisando a Equação 8.32, é possível constatar que $\alpha_i^* > 0$ apenas para os dados que se encontram sobre os hiperplanos H_1 e H_2, exatamente sobre as margens. Tais dados são denominados vetores de suporte (VS) e somente eles são utilizado no cálculo de w_i^* (Equação 8.30). A partir das últimas constatações associadas à Equação 8.32, ou seja, que somente os VS participam dos cálculos pertinentes, o valor de b^* é obtido a partir da Equação 8.32 e 8.33:

$$y_i(\vec{w^*} \cdot \vec{x_i} + b^*) - 1 = 0$$

$$\vec{w^*} \cdot \vec{x_i} + b = \frac{1}{y_i}, \tag{8.33}$$

onde, então, o valor de b^* é obtido a partir das médias dos VS:

$$b^* = \frac{1}{n_{VS}} \sum_{\vec{x_j} \in VS} (\frac{1}{y_j} - \vec{w^*} \cdot \vec{x_j}), \tag{8.34}$$

ou, ainda, substituindo w^* da Equação 8.30, tem-se

$$b^* = \frac{1}{n_{VS}} \sum_{\vec{x_j} \in VS} (\frac{1}{y_i} - \sum_{\vec{x_i} \in VS} (\alpha_i^* y_i \vec{x_i} \cdot \vec{x_j})). \tag{8.35}$$

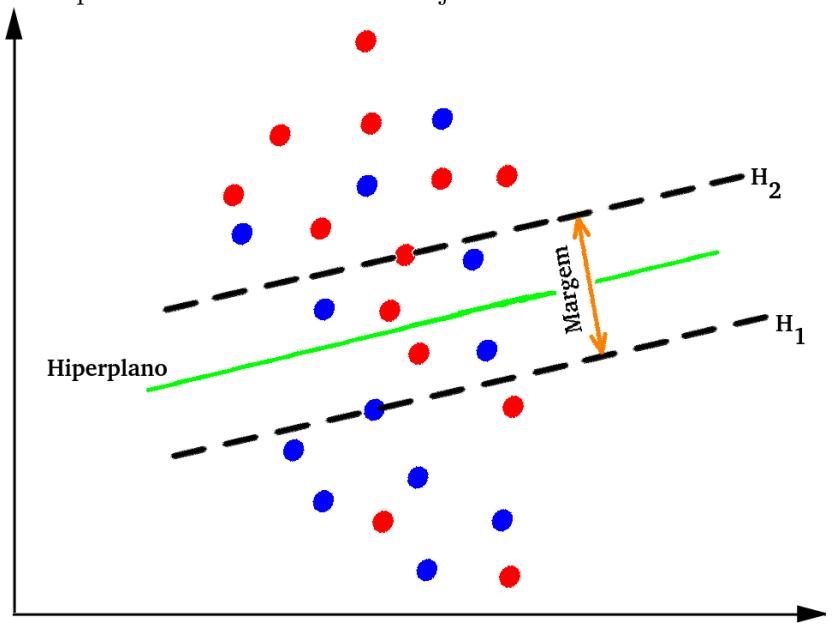
O objetivo da explanação até aqui foi encontrar os parâmetros $\vec{w}$ e b que gerassem um hiperplano plano associado à maximização da margem. No sentido de finalizar esta etapa, utiliza-se uma função sinal de forma a classificar os dados como $+1$ se $\vec{w} \cdot x + b > 0$ e como -1 se $\vec{w} \cdot x + b < 0$. Dessa forma, tem-se, então, o classificador $g(\vec{x})$:

$$g(\vec{x}) = \text{sgn}(\sum_{\vec{x_i} \in VS} (y_i \alpha_i^* \vec{x_i} \cdot \vec{x} + b^*)), \tag{8.36}$$

onde *sgn* representa a função sinal, w^* é calculado pela Equação 8.30 e b^* pela Equação 8.35.

Até agora tratou-se os dados linearmente separáveis. Nesta nova etapa será considerado dados que possuem "outliers" e ruídos (Figura 8.4). Estas últimas carac-

Figura 8.4: Representação de um conjunto de dados hipotético contendo ruídos e "outliers". O retrato também pode ser associado a um conjunto de dados não linear.



Fonte: elaborada pelo autor.

terísticas são abarcadas nas MVS com margens suaves. Dessa forma, então, admite-se que alguns dados possam violar a restrição posta pela equação 8.20. Tal relaxamento é feito com a inserção de variáveis de folga ξ_i, o qual torna as restrições impostas ao problema de otimização primal:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1, ..., n. \tag{8.37}$$

É possível visualizar nesta equação que um erro nos dados de treinamento é apontado por um valor de $\xi_i > 1$. Quando as MVSs de margens rígidas foram analisadas, o problema de otimização primal era a minimização de $\vec{w}$ em $\frac{1}{2}\|\vec{w^2}\|$. No caso das MVS com margens suaves o problema é ligeiramente diferente. O problema de otimização pode ser exposto na equação 8.38:

$$\min_{\vec{w}, b, \vec{\xi}} \quad \frac{1}{2}\|\vec{w}\|^2 + C(\sum_{i=1}^{n} \xi_i) \tag{8.38}$$

onde $\sum_{i=1}^{n} \xi_i$ caracteriza o limite no número de erros de treinamento e o termo de regularização C representa o peso atribuído à minimização dos erros nos dados de treinamento em relação à minimização da complexidade do modelo. A Equação 8.38 possui, então, um termo associado à maximização das margens pela minimização de $\vec{w}$, e um segundo termo associado à minimização dos erros sobre o conjunto de treinamento.

A obtenção do problema de otimização dual passa pelos passos expostos anteriormente e difere somente pela restrição, a qual agora dependerá de C:

$$\underset{\vec{\alpha}}{\text{maximizar}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\vec{x_i} \cdot \vec{x_j}) \tag{8.39a}$$

$$\text{sujeito a} \quad 0 \leq \alpha_i \leq C, \forall i = 1, ..., n, , \tag{8.39b}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{8.39c}$$

Assim como exposto anteriormente, parte-se do pressuposto que $\vec{\alpha^*}$ é solução do problema dual e $\vec{w^*}$, b^* e ξ^* são as soluções do problema primal. Calcula-se $\vec{w^*}$ por meio da Equação 8.30 e ξ_i é obtido a partir da definição da equação [65]:

$$\xi_i^* = \max\{0, 1 \ - y_i \sum_{j=1}^{n} (y_i \alpha_j^* \vec{x_j} \cdot \vec{x_i} + b^*)\}. \tag{8.40}$$

De forma simplificada, o cálculo desta equação fornece uma medida de quanto um ponto foi erroneamente classificado. Se um ponto foi corretamente classificado ele tem o valor associado a este cálculo de 0.

Novamente, a partir das condições de Kühn-Tucker, é possível delinear os valores de b^* e os vetores suporte, onde, então, elas são:

$$\alpha_i^* (y_i (\vec{w^*} \cdot \vec{x}_i + b^*) - 1 + \xi_i^*) = 0, \tag{8.41}$$

$$(C - \alpha_i^*) \xi_i^* = 0. \tag{8.42}$$

Assim como no caso estudado anteriormente, os dados x_i para os quais $\alpha_i^* > 0$ são designados como vetores de suporte. Contudo, existem casos a serem analisados. Quando $\alpha_i^* > 0$, se $\alpha_i^* < C$, então pela Equação 8.42 tem-se, $\xi_i^* = 0$, implicando em um ponto corretamente classificado sobre as margens. Quando $\alpha_i^* = C$, a partir da Equação 8.42, $\xi_i^* \geqslant 0$, situação que gera três possíveis desdobramentos. Primeiro, um ponto incorretamente classificado, $\xi_i^* > 1$. Segundo, pontos corretamente classificados entre as margens, $0 < \xi_i^* \leq 1$. Terceiro, caso raro, $\xi_i^* = 0$, pontos sobre as margens. O cálculo de b^* é feito utilizando a Equação 8.29, porém os VS utilizados serão aqueles que possuírem $\alpha_i^* < C$. Finalizando este caso, a classificação pode ser obtida pela Equação 8.36 associada à Equação 8.39a.

Até agora tratou-se de casos que envolviam dados linearmente separáveis, porém, dados reais, via de regra, não são bem comportados como os utilizados como exemplos até agora. Para lidar com estes tipo de ocorrência, é necessário lançar mão

de um artifício matemático juntamente com a utilização de *MVS* com margens suaves. Tal artifício é caracterizado por mapear o conjunto de dados de treinamento, denominado de entradas $\mho$, em um novo espaço de maior dimensão, denominado espaço de características $\Im$. A situação descrita pode ser definida pela operação apresentada na Equação 8.43:

$$\mathbf{\Phi} = \mho \rightarrow \Im \tag{8.43}$$

O mencionado método é garantido ter sucesso baseado no teorema de Cover [66], onde sucintamente, afirma-se que se a transformação não for linear e o espaço de características for suficientemente alto, tem-se alta probabilidade de a transformação gerar dados linearmente separáveis. Para conduzir esta estratégia, inicialmente aplica-se a transformação na Equação 8.39a, passo este que resulta na Equação 8.44a:

$$\underset{\vec{\alpha}}{\text{maximizar}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{\Phi}(\vec{x_i}) \cdot \mathbf{\Phi}(\vec{x_j})) \tag{8.44a}$$

De forma análoga ao exposto sobre as *MVS* de margens suaves, pode-se adaptar a solução de b^* da seguinte forma:

$$b^* = \frac{1}{n_{VS:\alpha^*<C}} \sum_{\vec{x}_j \in VS:\alpha_j^*<C} (\frac{1}{y_i} - \sum_{\vec{x}_i \in VS} \alpha_i y_i \mathbf{\Phi}(\vec{x}_i) \cdot \mathbf{\Phi}(\vec{x}_j)), \tag{8.45}$$

e o classificador obtido é:

$$g(\vec{x}) = \text{sgn}(\sum_{\vec{x_i} \in VS} (y_i \alpha_i^* \mathbf{\Phi}(\vec{x}_i) \cdot \mathbf{\Phi}(\vec{x}) + b^*). \tag{8.46}$$

A forma como as *MVS* foram mostradas até agora torna possível a aplicação do que é conhecido como truque do "Kernel". Tal artifício está baseado na constatação de que nas Equações 8.44a, 8.45, 8.46 não é necessário saber a transformação $\mathbf{\Phi}$ para realizar os cálculos, mas sim como realizar o produto escalar $\mathbf{\Phi}(\vec{x_i}) \cdot \mathbf{\Phi}(\vec{x_j})$ no espaço de características. Dessa forma, então, um "Kernel" K pode ser definido da seguinte forma:

$$K(\vec{x_i}, \vec{x_j}) = \mathbf{\Phi}(\vec{x_i}) \cdot \mathbf{\Phi}(\vec{x_j}) \tag{8.47}$$

Alguns dos "Kernels" mais utilizados podem ser visualizados na Tabela 8.4. Para que a utilização de "Kernel" fique adequada às MVS, de forma simplificada, é necessário que ele gere matrizes positivas semi-definidas $\mathbf{K}$, onde K_{ij} é definido por $K(ij) = K(\vec{x}_i, \vec{x}_j)$, $\forall i = 1, ..., n$ [68]. Os "Kernels" listados na Tabela 8.4 possuem parâmetros que precisam ser determinados pelo usuário e deve-se tomar cuidado ao utilizar o "Kernel" sigmoidal, pois este possui apenas alguns valores que satisfazem a condição exposta na frase

Tabela 8.4: "Kernels" mais utilizados. RBF é a abreviação de "Radial-Basis Function", porém ele também é conhecido como "Kernel" Gaussiano.

Kernel	**Função** $K(\vec{x}_i, \vec{x}_j)$	**Parâmetros**
RBF	$exp(-\zeta\|\|\vec{x}_i - \vec{x}_j\|\|^2)$	ζ
Polinomial	$(\delta(\vec{x}_i \cdot \vec{x}_j) + \kappa)^d$	δ, κ e d
Sigmoidal	$tanh(\delta((\vec{x}_i \cdot \vec{x}_j)) + \kappa)$	δ e κ

Fonte: adaptada da Ref. [67]

anterior.

As bases gerais sobre *MVS* foram expostas. Contudo, as soluções dos problemas duais não foram explicitadas devido a estas serem extensas e este não ser o foco do estudo. Existem numerosos pacotes capazes de solucionar os problemas aqui expostos e eles já estão implementados em uma série de soluções computacionais, como no caso da biblioteca *scikit-learn* associada à linguagem *Python*, a qual será utilizada neste estudo. Via de regra, a solução envolve adaptar às soluções produzidas para dados não volumosos, tornando-as adequadas à aplicação em dados volumosos, em geral, por meio da decomposição em subproblemas menores [56].

Capítulo 9

Resultados e discussão

A metodologia proposta foi aplicada e analisada em um conjunto de dados *benchmark* fornecido em um estudo que gerou o preditor *LBTOPE* [48]. Para gerar o referido estudo, os autores por meio do *IEDB*, recuperaram sequências imunogênicas lineares associadas às células B e não imunogênicas. Após a aquisição das sequências, os autores produziram 5 conjuntos de dados nomeados *Lbtope_Fixed*, *Lbtope_Fixed_non_redundant*, *Lbtope_Variable*, *Lbtope_Variable_non_redundant* e *Lbtope_Confirm*. A coleção de dados *Lbtope_Variable* é constituída por sequências de tamanho variável, contendo 14876 epítopos lineares de células B e 23321 sequências de não-epítopos. O conjunto de dados *Lbtope_fixed* é constituído por sequências de tamanho fixo de 20 aminoácidos, contendo 12063 epítopos lineares de células B e 20589 sequências de não-epítopos. Para alcançar o tamanho fixo de 20 resíduos, os autores adotaram as seguintes ações:

- Sequências menores que 5 resíduos foram removidas;
- Sequências maiores do que 20 resíduos foram aparadas, em ambas extremidades, e posteriormente os 20 resíduos do meio foram utilizados;
- Sequências menores que 20 resíduos foram estendidas adicionando número igual de resíduos em ambas extremidades (a sequência foi mapeada no antígeno de origem e estendida);
- Sequências idênticas foram removidas.

A coleção de dados *Lbtope_Confirm* contém 1042 epítopos lineares de células B e 1795 sequências de não-epítopos validados por mais de um experimento. Os conjuntos de dados *Lbtope_Variable* e *Lbtope_Fixed* possuem sequências redundantes e a utilização deles pode gerar vieses indesejáveis, falsa alta performance de predição e

Tabela 9.1: Performance de predição do *LBTOPE* utilizando o conjunto de dados *Lbtope_Variable_non_redudant*. *Limiar* se refere a um limiar encontrado pelos autores que maximiza o Coeficiente de Correlação de Matthews (CCM), *Sensit.* se refere a sensitividade, *Espec.* se refere a especificidade, *Acur.* se refere a acurácia e *AAC* se refere a área abaixo da curva *ROC*.

Caracteristica	**Limiar.**	**Sensit.**	**Espec.**	**Acur**	**CCM**	**AAC**
Comp. de aminoácidos	-0,3	56,41	59,9	58,39	0,16	0,60
Comp. e transição	-0,6	58.41	66.86	63,19	0,25	0,66
Pares de aminoácidos	-0,3	63,51	62,95	63,19	0,26	0,68
Comp. de dipeptídeos	-0,1	66	67,24	66,7	0,33	0,73

Fonte: obtida da Ref. [48].

demasiado tempo de treinamento [69–72]. Para evitar este problema, os autores produziram mais dois conjuntos de dados nomeados de *Lbtope_Variable_non_redundant* e *Lbtope_Fixed_non_redundant*. A coleção de dados *Lbtope_Variable_non_redundant* foi produzida a partir da *Lbtope_Variable* após uma de duas sequências que possuíssem mais do que 80% de redundância ser removida, gerando 8011 sequências de epítopos lineares de células B e 10868 sequências de não-epítopos. O conjunto de dados *Lbtope_Fixed_non_redundant* foi produzido a partir do *Lbtope_Fixed* por meio do mesmo critério de redundância e contém 7824 sequências de epítopos lineares de células B e 7853 sequências de não-epítopos.

Para gerar o vetor de características, os autores utilizaram separadamente: perfil binário, perfil físico-químico, composição de dipeptídeos, propriedades de par de aminoácidos de Chen e composição de transição e distribuição. A predição das sequências foi realizada utilizando *SVM* e *k-vizinhos mais próximos*. O treinamento foi realizado utilizando 90% das sequências por meio de validação cruzada com cinco partes e os 10% restantes foram utilizados para teste e análise de performance. Considerando-se que as performances de predição em dados redundantes na maioria das vez são superestimadas, neste estudo maior atenção será dada aos conjuntos de dados não-redundantes. Assim sendo, quando os autores utilizaram o conjunto de dados *Lbtope_Variable_non_redundant* e produziram as predições por meio de *SVM* (*k-vizinhos mais próximos* não foi analisado neste conjunto de dados) obtiveram os resultados expostos na Tabela 9.1. Quando analisaram o *Lbtope_Fixed_non_redundant* por meio de *SVM* (novamente *k-vizinhos mais próximos* não foi analisado neste conjunto de dados) eles obtiveram os resultados expostos na Tabela 9.2.

A partir da análise das Tabelas 9.1 e 9.2, é possível concluir que a composição de dipeptídeos está associada às melhores performances. Dado que o objetivo desta tese é analisar a melhora de predição de epítopos lineares de células B, pos-

Tabela 9.2: Performance de predição do *LBTOPE* utilizando o conjunto de dados *Lbtope_Fixed_non_redudant*. *Limiar* se refere a um limiar encontrado pelos autores que maximiza o Coeficiente de Correlação de Matthews (CCM), *Sensit.* se refere a sensitividade, *Espec.* se refere a especificidade, *Acur.* se refere a acurácia e *AAC* se refere a área abaixo da curva *ROC*.

Caracteristica	**Limiar.**	**Sensit.**	**Espec.**	**Acur**	**CCM**	**AAC**
Comp. de aminoácidos	0	59,35	57,31	58,33	0,17	0,61
Comp. e transição	0	54,38	57,31	55,85	0,12	0,57
Pares de aminoácidos	0	65,88	60,57	63,26	0,26	0,66
Comp. de dipeptídeos	0	65,78	63,97	64,86	0,30	0,69

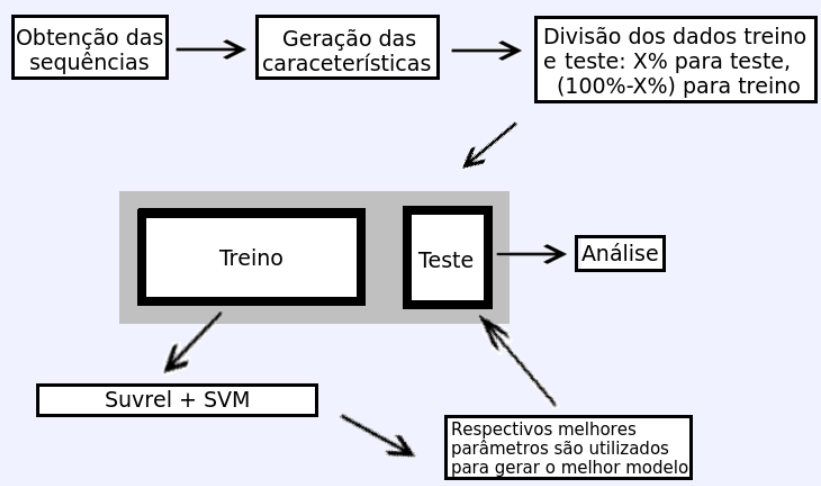
Fonte: obtida da Ref. [48].

sivelmente gerando um preditor público, e como a composição de dipeptídeos gerou as melhores performances, decidiu-se utilizar este gerador de característica. Pode-se arguir que as decisões foram demasiadamente arbitrárias. Contudo, deve-se levar em conta que o tempo de processamento quando utiliza-se MVS é alto e não se tinha tempo hábil para analisar uma grande variedade de casos.

Os passos envolvidos na análise da possibilidade de melhora de predição por meio do Suvrel podem ser vistos na Figura 9.1. A caixa *Obtenção das sequências* representa a recuperação das sequências utilizadas no estudo do *LBTOPE*. Como exposto anteriormente, dentre os conjuntos disponíveis, maior atenção será dada ao *Lbtope_Variable_non_redundant* e ao o *Lbtope_Fixed_non_redundant*. No segundo passo, denominado *Geração das características*, as sequências de aminoácidos são caracterizadas por meio da composição de dipeptídeos. No terceiro passo, o conjunto de dados é dividido em 90% para treinamento e 10% para teste. Utilizando o conjunto treinamento, obtém-se o tensor métrico $g_{\mu\nu}$ por meio da Equação 8.11. O cálculo desta última equação depende do valor de γ. Para encontrar o valor de γ que possivelmente gere a melhor performance, propõe-se a estratégia descrita a seguir. A partir da relação entre as distâncias nomeadas de efetivas relativas e dos valores de γ, procura-se o primeiro γ que gere autovalores positivos associados ao tensor métrico e gere a maior distância efetiva relativa, onde, este é nomeado $\gamma_{ótimo}$. Para discorrer sobre o cálculo das distância efetivas relativas, considere que um experimento i é descrito pelo vetor $\vec{x_i} = (x_{i1}, x_{i2}, ..., x_{iF})$ no espaço de dimensões $\mathbb{R}^F$ e cada caraterística μ representa uma dimensão, $\mu = 1, 2, ..., F$. A partir destas últimas definições é possível calcular o valor médio da caraterística μ associada a cada condição (*imunogênica* ou *não-imunogênica*):

$$\overline{x}_{\mu}^{Cond} = \frac{1}{n^{Cond}} \sum_{i \in Cond}^{n} x_{i\mu}, \tag{9.1}$$

Figura 9.1: Visão geral da abordagem empregada para análise de possível melhora de performance da predição de epítopos lineares de células B por meio do *Suvrel*. A caixa *Obtenção das sequências* representa a recuperação das sequências utilizadas no estudo do *LBTOPE*. A caixa *Geração das características* representa o passo da geração do vetor de características. A próxima caixa representa a divisão dos dados em treino e teste. A seta saindo da caixa *Treino* representa o conjunto de treino sendo dividido em cinco partes e utilizado para treinar o *SVM* associado ao *Suvrel*. Após a busca pelos melhores parâmetros associados ao *SVM* e *Suvrel*, estes são utilizados para treinar o modelo novamente com o conjunto total de treinamento. O modelo então é testado no conjunto teste e a análise de performance é realizada.



Fonte: elaborada pelo autor.

onde $Cond$ é a condição imunogênica ou não-imunogênica, n^{Cond} é o número de sequências associado a $Cond$ e n é número total de sequências. Com o auxílio desta última equação, a distância entre os pontos médios associados à condição imunogênica e não-imunogênica pode ser calculada:

$$\Delta x = \sqrt{\sum_{u}^{F} (\overline{x}_{\mu}^{Imun} - \overline{x}_{\mu}^{Nao-Imun})^2}. \tag{9.2}$$

Para prosseguir com a explanação, considere que o raio de um experimento i associado a condição imunogênica é:

$$r_i^{Imun} = \sqrt{\sum_{\mu}^{F} (x_{i\mu}^{Imun})^2}, \tag{9.3}$$

e o raio médio associado à condição imunogênica é:

$$\overline{r}^{Imun} = \frac{1}{n^{Imun}} \sum_{i \in Imun}^{n^{Imun}} r_i^{Imun}. \tag{9.4}$$

A dispersão dos raios associados à condição imunogênica pode então ser calculado:

$$\sigma^{Imun} = \frac{1}{n^{Imun}} \sum_{i \in Imun}^{n} \sqrt{(r_i^{Imun} - \overline{r}^{Imun})^2} \tag{9.5}$$

e de modo análogo pode-se executar estes cálculos para a condição não-imunogênica. Por fim a distância média entre as duas condições é dada por:

$$D = \frac{\Delta x}{\sqrt{\sigma^{Imun} \sigma^{Nao-imun}}} \tag{9.6}$$

A distância calculada após a aplicação do Suvrel é denominada D_{Suvrel} e sem Suvrel é $D_{SemSuvrel}$, e assim, a distância efetiva relativa ao cálculo sem o Suvrel é:

$$D_{efetiva} = \frac{D_{Suvrel}}{D_{SemSuvrel}}. \tag{9.7}$$

Após a finalização destes últimos passos, em seguida, transforma-se os dados de treino e teste por meio da operação $\sqrt{g_{\mu\nu}}$ com o $\gamma_{ótimo}$. Para treinar o SVM, decidiu-se seguir o aconselhamento presente na Ref. [73]. Os passos são os seguintes:

- Normalização dos dados; nesta tese utilizou-se média 0 e variância 1;
- Utilização do *kernel* radial;

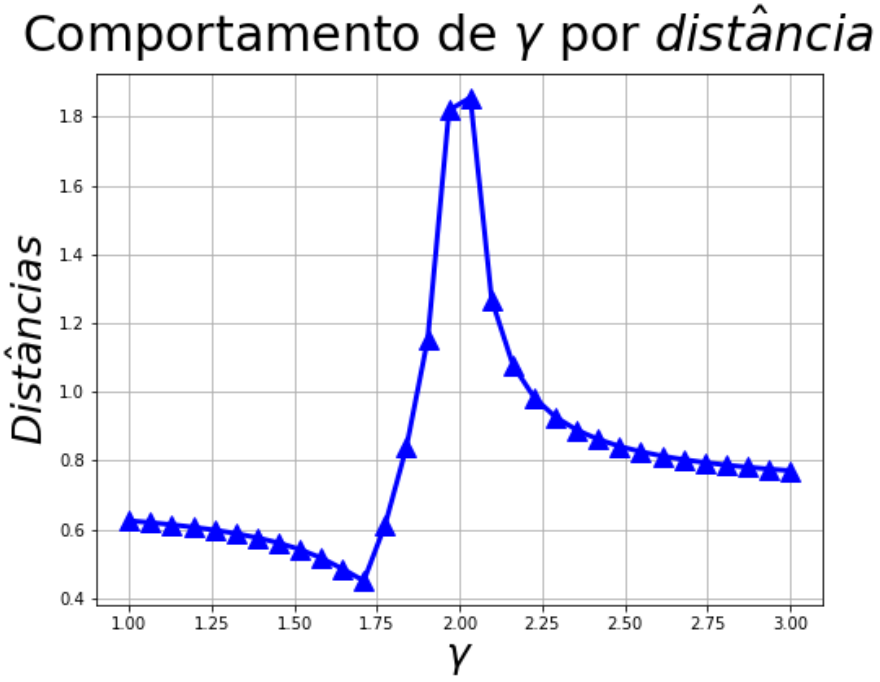
- Utilização de validação cruzada (por motivos de comparação, serão utilizadas cinco partes como no estudo do *LBTOPE*) para encontrar os melhores parâmetros. Como decidiu-se pelo "kernel" radial, seguiu-se as indicações de utilizar os valores de ζ associado ao "kernel" radial da seguinte maneira: $\zeta = 2^{-15}, 2^{-13}, ..., 2^{3}$ associando em pares com $C = 2^{-5}, 2^{-3}, ..., 2^{15}$. Informações sobre o parâmetro ζ podem ser obtidas na Seção 8.5.1 e C representa o peso atribuído à minimização dos erros nos dados de treinamento em relação à minimização da complexidade do modelo e mais informações sobre este termo também podem ser obtidas em na Seção 8.5.1.

- Finalmente, a utilização dos melhores parâmetros para treinamento utilizando todo o conjunto de dados de treinamento.

A validação cruzada consiste em dividir um conjunto de treinamento em k partes. Dado um par de C e ζ, o modelo é treinado em $k-1$ partes e testado na parte restante dos dados que não foi utilizada para o treinamento. Em seguida, define-se $k-1$ partes de modo que o teste deverá ser feito em uma parte que ainda não foi utilizada para teste. O modelo tido como ótimo é aquele que teve uma dada medida estatística média maior associada às k partes, onde, no caso deste trabalho foi a acurácia, analisando todos os pares C e ζ. Por fim, o melhor modelo é treinado utilizando todo o conjunto de treinamento. Assim, a fase de treinamento está finalizada. O passo seguinte é aplicar o melhor modelo nos 10% dos dados que não foram utilizados para o treinamento, ou seja, o conjunto teste. A última etapa é a procura de um limiar que maximize o Coeficiente de Correlação de Matthews (CCM) para que sensitividade e a especificidade fiquem próximas no conjunto teste. Para realizar esta tarefa, obtém-se a probabilidade p de cada classificação do conjunto teste, transformando-a por meio de $2.0(0.5-p)$. A partir destes últimos valores procura-se um limiar entre -1 e 1 que maximize o CCM. Por fim, tem-se um modelo preditivo. É importante ressaltar que os passos associados a maximização do CCM foram expostos no artigo do preditor *LBTOPE* [48] e foram aqui seguidos para se ter base comparativa adequada. Após estes passos, é possível analisar a performance do modelo através do cálculo dos valores de sensitividade, especificidade, acurácia, CCM e área abaixo da curva ROC (AAC).

O primeiro caso analisado se deu por meio do conjunto de dados *Lbtope_Fixed_non_redundant*. Como dito anteriormente, o vetor de características foi produzido utilizando a composição de dipeptídeos. Após dividir o referido conjunto, utilizou-se o conjunto de treinamento para analisar a relação do γ presente no cálculo do *Suvrel* e as distâncias efetivas relativas (Figura 9.2). Para se ter uma ideia geral do comportamento utilizou-se o intervalo de γ de 1 a 3. Tendo conhecimento deste comportamento

Figura 9.2: Distâncias efetivas relativas do conjunto de dados *Lbtope_Fixed_non_redundant* utilizando intervalo de γ de 1 a 3.



Fonte: elaborada pelo autor.

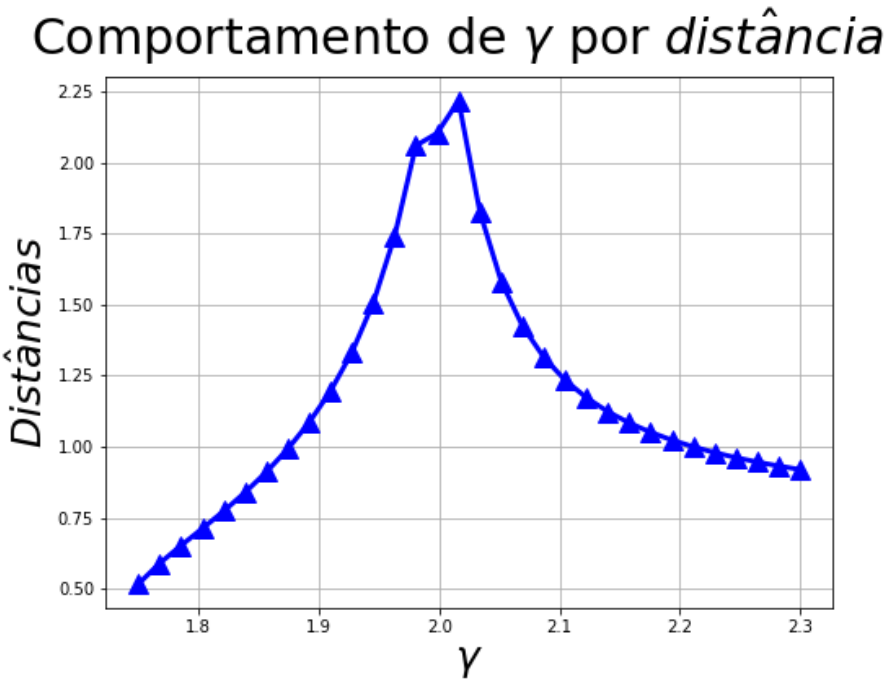
Tabela 9.3: Comparação dos resultado entre o *LBTOPE* e a metodologia proposta utilizando o conjunto de dados *Lbtope_Fixed_non_redundant*.

	Limiar	**Sensit.**	**Especif.**	**Acurác.**	**CCM**	**AAC**
Suv. $\gamma = 2,02$	-0,06	71,48	61,32	66,39	0,33	0,71
LBTOPE	0	65,75	63,97	64,86	0,3	0,69

geral, esperava-se que as maiores distâncias efetivas estivessem por volta de $\gamma = 2,0$ e este fato é constatado. Ainda assim, julgou-se necessário analisar a região por volta de $\gamma = 2,0$ com mais detalhe e isto é feito na Figura 9.3. Para definir o valor de γ associado ao Suvrel procedeu-se da seguinte maneira: ranqueou-se os valores de γ de forma crescente e selecionou-se aquele que não gerasse autovalores negativos associados ao tensor métrico $g_{\mu\nu}$ e que gerasse a maior distância efetiva relativa. Neste caso encontrou-se o valor de $\gamma = 2.02$. A Tabela 9.3 mostra a performance deste modelo juntamente com a análise dos resultados no mesmo conjunto de dados pelo *LBTOPE*.

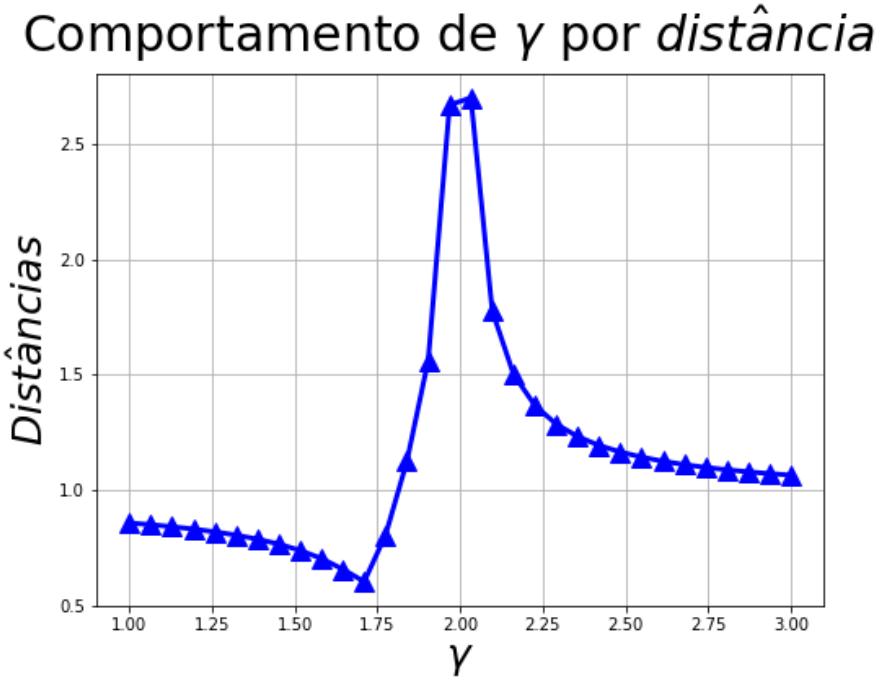
A análise do conjunto de dados *Lbtope_Variable_non_redundant* seguiu os mesmos passos descritos anteriormente. O comportamento das distâncias efetivas dependente de γ é mostrado nas figuras 9.4 e 9.5. A Tabela 9.4 mostra a performance do melhor modelo juntamente com a análise dos resultados no mesmo conjunto de dados

Figura 9.3: Distâncias efetivas relativas do conjunto de dados *Lbtope_Fixed_non_redundant* utilizando intervalo de γ de $1,75$ a $2,3$.



Fonte: elaborada pelo autor.

Figura 9.4: Distâncias efetivas relativas do conjunto de dados *Lbtope_Variable_non_redundant* utilizando intervalo de γ de 1 a 3.



Fonte: elaborada pelo autor.

Figura 9.5: Distâncias efetivas relativas do conjunto de dados *Lbtope_Variable_non_redundant* utilizando intervalo de γ 1, 75 a 2, 3.

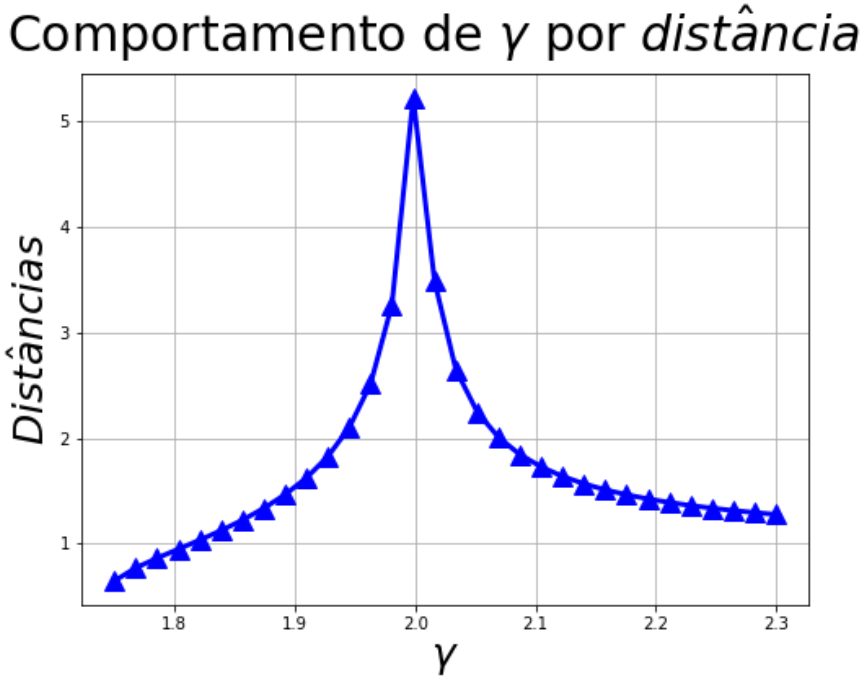


Tabela 9.4: Comparação dos resultado entre o *LBTOPE* e a metodologia proposta utilizando o conjunto de dados *Lbtope_Variable_non_redundant.*

	Limiar	**Sensit.**	**Especif.**	**Acurác.**	**CCM**	**AAC**
Suv. $\gamma = 2,02$	-0,34	82,87	52,52	65,39	0,36	0,74
LBTOPE	-0,1	66	67,24	66,7	0,33	0,73

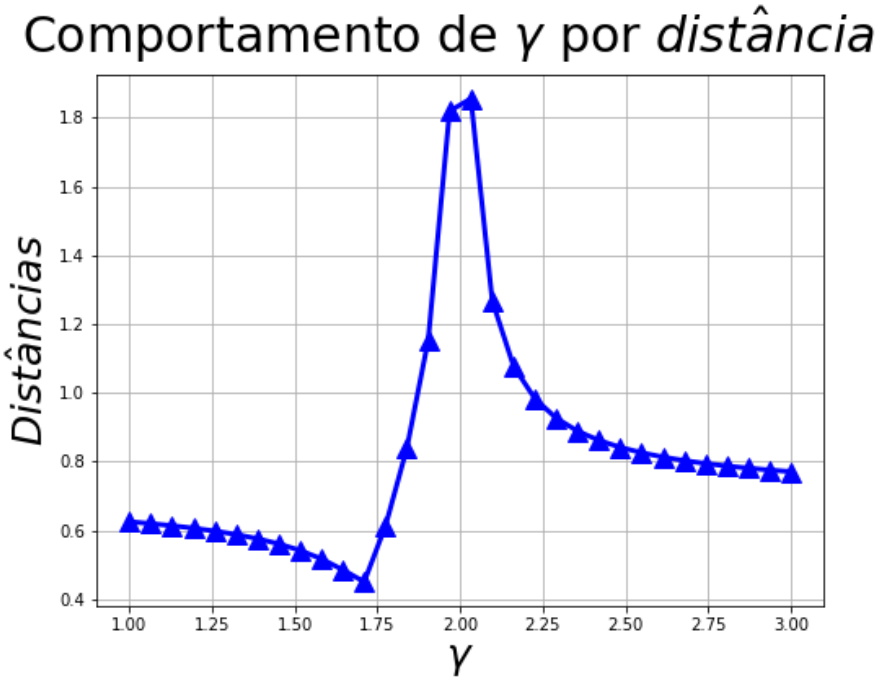
Fonte: elaborada pelo autor.

pelo *LBTOPE.*

A análise do conjunto de dados *LBConfirm* seguiu os mesmos passos descritos anteriormente. O comportamento das distâncias efetivas dependentes de γ é mostrado nas figuras 9.6 e 9.7. A Tabela 9.5 mostra a performance do melhor modelo juntamente com a análise dos resultados no mesmo conjunto de dados pelo *LBTOPE.*

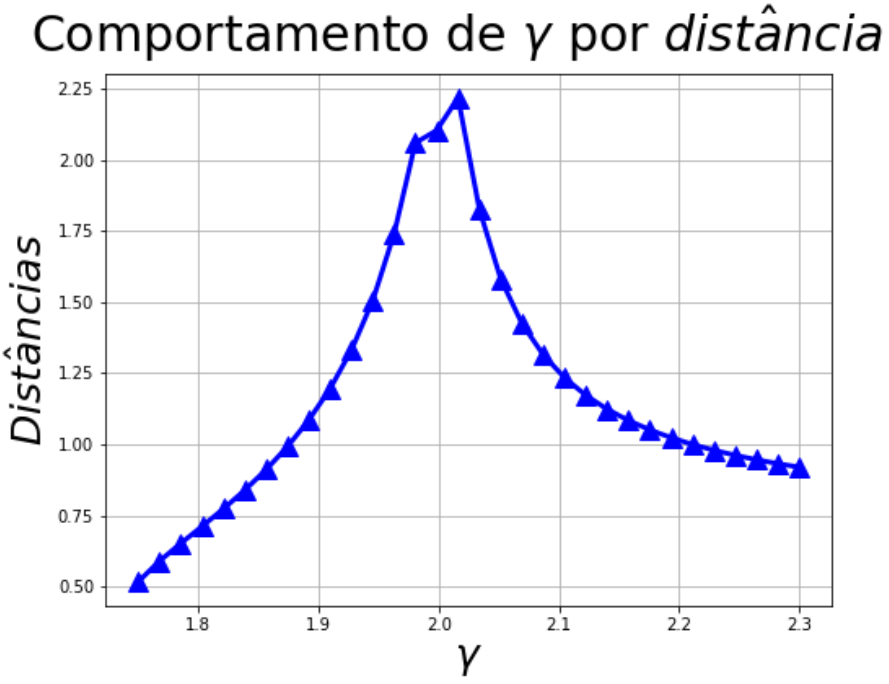
No estudo presente no artigo do *LBTOPE* os autores avaliaram mais 3 conjuntos de dados associados aos preditores *ABCpred*, *BCPred* e modelo de *Chen*. Breves informações sobre os preditores *ABCPred* e *BCpred* podem ser obtidas na Seção 8.5 e em [36, 74], respectivamente. O modelo de Chen se trata da solução proposta por Chen et al. [75] baseada na criação de um modelo de escala de antigenicidade baseado em par de aminoácido e *MVS.* O primeiro caso analisado se deu por meio do conjunto de dados associado ao *ABCpred.* Dado que as sequências nesses conjunto de dados

Figura 9.6: Distâncias efetivas relativas do conjunto de dados *LBConfirm* utilizando intervalo de γ 1 a 3.



Fonte: elaborada pelo autor.

Figura 9.7: Distâncias efetivas relativas do conjunto de dados *LBConfirm* utilizando intervalo de γ $1,75$ a $2,3$.



Fonte: elaborada pelo autor.

Tabela 9.5: Comparação dos resultado entre o *LBTOPE* e a metodologia proposta utilizando o conjunto de dados *LBConfirm*.

	Limiar	**Sensit.**	**Especif.**	**Acurác.**	**CCM**	**AAC**
Suv. $\gamma = 2,02$	-0,1	83,65	88,33	86,62	0,71	0,91
LBTOPE	-0,3	84,62	81,01	82,33	0,64	0,91

Fonte: elaborada pelo autor.

Tabela 9.6: Comparação entre o *LBTOPE*, ABCPred e a metodologia proposta.

	Sensit.	**Especif.**	**Acurác.**	**CCM**
Suvrel Fixed no ABCPred	79,62	43	61	0,24
LBTOPE Fixed no ABCPred	70	33,71	57,9	0,04
ABCpred nos próprios dados	57,14	71,54	64,36	0,29

Fonte: elaborada pelo autor com parte das informações vindas da Ref. [48].

tem comprimento fixo de 20 aminoácidos, o autor do *LBTOPE* optou por utilizar o modelo treinado nos dados *Lbtope_Fixed*. Contudo, dado que julgou-se que o treinamento em dados redundantes não fosse prudente resolveu-se analisar a performance da metodologia proposta treinada nos dados *Lbtope_Fixed_non_redundant*. A Tabela 9.6 mostra os resultados. O segundo caso analisado foi realizado por meio do conjunto de dados associado ao modelo de *Chen*. As mesmas considerações associadas aos dados escolhidos pelos autores do *LBTOPE* e o tamanho das sequências também são levantados neste caso. A Tabela 9.7 mostra os resultados. O terceiro caso está associado aos dados do *BCPred*. O autor do *LBTOPE* comparou seu modelo treinando no conjunto de dados *LBtope_Fixed*, *LBtope_Variable* e *LBtope_Confirm*. Dado que as sequências associadas ao *BCPred* tem tamanho fixo de 20 aminoácidos, julgou-se coerente comparar o nosso modelo treinado no *LBtope_Fixed_non_redundant* com o *LBTOPE* treinado no *LBtope_Fixed*. A Tabela 9.8 mostra os resultados.

Tabela 9.7: Comparação entre o *LBTOPE*, modelo de *Chen* e a metodologia proposta.

	Sensit.	**Especif.**	**Acurác.**	**CCM**
Suvrel Fixed no Chen	75	41,97	58,71	0,18
LBTOPE Fixed no Chen	70,76	35,89	53,33	0,07

Fonte: elaborada pelo autor com parte das informações vindas da Ref. [48].

Tabela 9.8: Comparação entre o *LBTOPE*, preditor *BCPred* e a metodologia proposta.

	Sensit.	**Especif.**	**Acurác.**	**CCM**
Suvrel Fixed no BCpred	73,75	39,85	56,82	0,14
LBTOPE Fixed no BCpred	69,33	33,81	51,57	0,03
BCPred no LBTOPE_Fixed	58,87	50,47	53,57	0,09

Fonte: elaborada pelo autor com parte das informações vindas da Ref. [48].

A predição de epítopos lineares de células B comparada com a predição de epítopos conformacionais de células B e T é a mais desafiadora. A possível razão para isso se deve a natureza dos epítopos lineares de células B, ou seja, seu tamanho variando de $2 - 85$ resíduos [48]. Nas análises aqui produzidas tal situação também foi verificada. Contudo, ainda é possível redigir alguns pontos importantes sobre o estudo aqui realizado. Os objetivos desta tese são analisar a possível melhora de predição de epítopos lineares de células B se baseando no *LBTOPE*. É possível concluir que a metodologia aqui proposta baseada na utilização de *MVS* e *Suvrel* melhorou a performance de predição comparada ao *LBTOPE*. Também é possível concluir que a metodologia proposta também teve performance superior comparada ao *LBTOPE* em outros conjuntos de dados. De uma forma geral, esperava-se que o método *Suvrel* melhorasse de forma mais acentuada as performances ligadas aos objetivos. Contudo, essa foi uma abordagem inicial e alguns pontos podem ser melhor explorados, como por exemplo, estudo de outros vetores de características e a utilização de outros métodos de seleção de características juntamente com o Suvrel. Porém, é possível analisar os resultados de outra forma. A maioria dos epítopos de células B são conformacionais, por volta de 90%, e os preditores baseados em estrutura possuem performance superior aos baseados em sequência [76, 77]. Embasado nestas informações pode-se arguir que a melhor alternativa é dedicar mais esforço a predição conformacional [77]. É possível sugerir por meio dos resultados aqui expostos que por mais que a predição de epítopos lineares melhore ainda não mostra resultados que possam trazer conforto na sua utilização e a melhor alternativa seja despender esforços na predição conformacional. Como dito anteriormente, este estudo é inicial e não optou-se pela predição conformacional devido a esta ser mais desafiadora do que a predição sequencial e os dados e métricas utilizados para este tipo de predição serem indiretos e mais complexos. Dessa forma, em uma perspectiva futura, é possível analisar os impactos do *Suvrel* associados à predição conformacional.

Com o propósito de aprofundar o estudo decidiu-se analisar a possível me-

Tabela 9.9: Comparação dos resultado entre o *LBTOPE* e a metodologia proposta utilizando o conjunto de dados *Lbtope_Fixed.*

	Limiar	**Sensit.**	**Especif.**	**Acurác.**	**CCM**	**AAC**
Suv. $\gamma = 2,02$	-0,15	73,88	84,4	80,51	0,58	0,86
LBTOPE	-0,3	80,5	81,67	81,24	0,61	0,88

Tabela 9.10: Comparação dos resultado entre o *LBTOPE* e a metodologia proposta utilizando o conjunto de dados *Lbtope_Variable.*

	Limiar	**Sensit.**	**Especif.**	**Acurác.**	**CCM**	**AAC**
Suv. $\gamma = 2,02$	-0,14	72,74	79,2	76,68	0,51	0,83
LBTOPE	-0,2	75,18	76,34	75,89	0,51	0,83

Fonte: elaborada pelo autor.

lhora associada a predição do *LBTOPE* utilizando a metodologia apresentada treinado-a no conjunto de dados *Lbtope_Fixed* e *Lbtope_Variable.*

Novamente, o primeiro caso analisado se deu por meio dos dados *Lbtope_Fixed* e os resultados são mostrados na Tabela 9.9. A medida de sensitividade, a qual mede a porcentagem de sequências imunogênicas corretamente classificadas diminuiu consideravelmente, e essa situação não é vantajosa. As outras medidas, acurácia, CCM e *AAC* também diminuíram e somente a especificidade aumentou. Dada esta análise conclui-se que a utilização do *Suvrel* não trouxe vantagens ao preditor *LBTOPE.*

O segundo caso analisado se deu por meio do conjunto de dados *Lbtope_Variable* e os resultados são mostrados na Tabela 9.10. Embora algumas medidas melhoraram ligeiramente, diminuir a sensitividade não é um ponto positivo.

Seguindo os mesmos passos, analisou-se o conjunto de dados associado ao preditor *ABCPred*, *BCPred* e modelo de Chen, por meio dos mesmos procedimentos utilizados quando os conjuntos de dados *Lbtope_Fixed* e *Lbtope_Variable* foram analisados. A Tabela 9.11 mostra os resultados utilizando os dados *ABCPred.* O segundo caso analisado se deu por meio do conjunto de dados associado ao modelo de *Chen.* A Tabela 9.12 mostra os resultados. O terceiro caso está associado aos dados do *BCPred.* A Tabela 9.13 mostra os resultados. Analisando estes últimos resultados conclui-se que a abordagem proposta melhorou ligeiramente a performance associada ao preditor *LBTOPE.* Contudo a melhora foi menor do que a apresentada na primeira parte quando treinou-se utilizando dados não-redundantes. Dessa forma, é sugerido que o *Suvrel* melhora a performance de predição mas a sua utilização deve vir seguida de tratamentos corretos relacionados aos dados. Complementarmente ainda tem-se aspectos relacionados ao tempo computacional. O tempo necessário para treinar os modelos utilizando

Tabela 9.11: Comparação entre o *LBTOPE*, ABCPred e a metodologia proposta.

	Sensit.	**Especif.**	**Acurác.**	**CCM**
Suvrel Fixed_non_redundant no ABCPred	79,62	43	61	0,24
Suvrel Fixed_redund. no ABCPred	71,5	48	59,55	0,2
LBTOPE Fixed_redund. no ABCPred	70	33,71	57,9	0,04

Fonte: elaborada pelo autor com parte das informações vindas da Ref. [48].

Tabela 9.12: Comparação entre o *LBTOPE*, modelo de *Chen* e a metodologia proposta.

	Sensit.	**Especif.**	**Acurác.**	**CCM**
Suvrel Fixed_non_redundant no Chen	75	41,97	58,48	0,18
Suvrel Fixed_redund. no Chen	65,25	51,26	58,26	0,17
LBTOPE Fixed_redund. no Chen	70,76	35,89	53,33	0,07

Fonte: elaborada pelo autor com parte das informações vindas da Ref. [48].

Tabela 9.13: Comparação entre o *LBTOPE*, preditor *BCPred* e a metodologia proposta.

	Sensit.	**Especif.**	**Acurác.**	**CCM**
Suvrel Fixed_non_redundant no BCPred	73,75	39,85	56,82	0,14
Suvrel Fixed_redund. no BCpred	63,62	44	53,82	0,08
LBTOPE Fixed_redund. no BCPred	69,33	33,81	51,57	0,03

Fonte: elaborada pelo autor com parte das informações vindas da Ref. [48].

dados não-redundantes foi consideravelmente menor do que utilizando redundantes.

Capítulo 10

Conclusões

De forma sucinta, o objetivo desta parte da tese é analisar a possível melhora dos resultados de predição de epítopos lineares de células B associados ao preditor *LBTOPE* [48] por meio de Máquinas de Vetores de Suporte e o método *Suvrel*. De forma mais detalhada, o primeiro objetivo é a análise da possível obtenção de resultados de performance melhores do que o preditor utilizando os dados que os autores do preditor geraram. O segundo objetivo é a investigação da possível obtenção de resultados de performance superiores em dados gerados por outros autores.

A análise detalhada foi apresentada na Seção 9, contudo, no sentido de tornar a conclusão objetiva resolveu-se analisar as acurácias e coeficientes de correlação de Matthews (CCM) nos casos analisados. A acurácia foi escolhida pois mede o quanto um preditor acerta e CCM foi escolhido por se tratar de uma medida que relaciona os positivos, negativos, verdadeiros negativos e positivos e falsos positivos e negativos de forma balanceada. Como exposto anteriormente, maior atenção foi dada aos conjuntos de dados não-redundantes. Dessa forma, no primeiro caso analisado (caso A) treinou-se e testou-se a abordagem proposta (SMVS, *Suvrel* associado a Máquinas de Vetores de Suporte) com o conjunto de dados *Lbtope_Fixed_non_redundant* e comparou-se os resultados quando o preditor *LBTOPE* também foi treinado e testado com este conjunto de dados. Os resultados dos cálculos de acurácia e CCM são mostrados nas figuras 10.1 e 10.2, respectivamente, na letra "A" dos eixos x. O segundo caso (caso "B") é similar ao primeiro com a exceção de que o conjunto de dados foi o *Lbtope_Variable_non_redundant*. O terceiro caso (caso "C") mostra os resultados quando as duas abordagens foram treinadas e testadas no conjunto de dados *LBConfirm*. Do quarto ao sexto caso (casos "D", "E" e "F") a abordagem proposta foi treinada no conjunto de dados *Lbtope_Fixed_non_redundant* e o *LBTOPE* foi treinado no conjunto de dados *Lbtope_Fixed_redundant* e ambas foram testadas em conjuntos de dados utiliza-

dos em outros estudos nomeados de *ABCPred*, *Chen* e *BCPred*. Informações resumidas dos casos analisados podem ser obtidas na Tabela 10. Quando analisa-se estes casos por meio das acurácias (Figura 10.1) vê-se que somente no caso B a abordagem proposta teve resultado inferior. Porém, quando os coeficientes de correlação de Matthews são analisados nos casos citados (Figura 10.2), em nenhum caso a SMVS tem resultados inferiores.

No sentido de realizar uma análise mais aprofundada, treinou-se a SMVS utilizando o conjunto de dados redundantes *Lbtope_Fixed_redundant* e comparou-se com o *LBTOPE* treinado neste mesmo conjunto de dados (caso "G"). Similarmente, tal análise também foi produzida com o conjunto de dados *Lbtope_Variable_redundant* (caso "H"). Nos três últimos casos ("I", "J", "L") a SMVS foi treinada no conjunto de dados de dados *Lbtope_Fixed_redundant* assim como o *LBTOPE* e testadas nos conjuntos de *ABCPred*, *Chen* e *BCPred*. Analisando a acurácia (Figura 10.1) vê-se que a SMVS teve resultado inferior no caso "G" e CCM (Figura 10.2) somente no caso "G" também.

Com exceção de três casos (dois casos "G" e um "B") todos os outros 19 casos a abordagem proposta teve resultado superior e em alguns casos até com ordem de grandeza superior. Quando compara-se os resultados obtidos pela abordagem proposta vê-se que os casos utilizando dados não-redundantes (caso "A", "B", "D", "E", "F") comparado com os redundantes (caso "G", "H", "I", "J", "L") de forma geral tiveram resultados relativos superiores. Dado isso, sugere-se que utilização do método *Suvrel* com dados não-redundantes tem performance superior devido as informações redundantes não trazerem benefícios e perturbarem a obtenção de informações relevantes. Os dados não-redundantes além de gerar performance de predição superior também reduzem consideravelmente o tempo de treinamento. Dado o exposto, e sabendo que o preditor *LBTOPE* é público, e os resultados aqui apresentados mostram que a SMVS é superior, futuramente pode-se hospedá-lo em uma página *web* e torná-la pública também.

Dado os resultados expostos, ainda assim é possível analisar se é possível melhorar ainda mais a performance de predição. Onde, então, tal análise pode ser realizada em projetos futuros por meio da utilização de outros vetores de características, ou por meio do uso de outros métodos de seleção de características juntamente com o *Suvrel*.

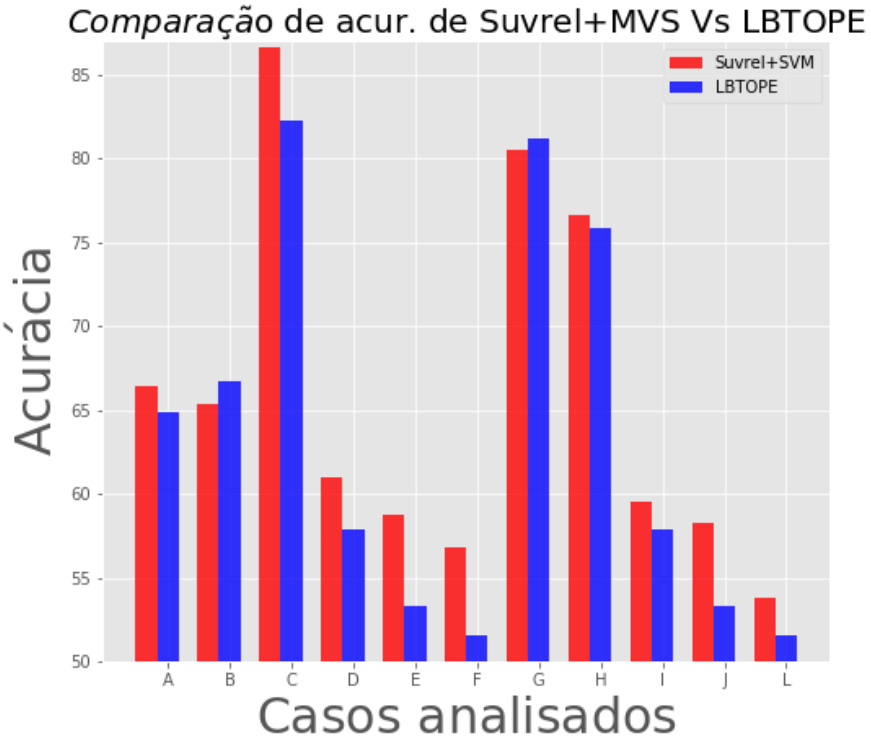
A maioria dos epítopos de células B são conformacionais, e os preditores baseados em estrutura possuem performance superior aos baseados em sequência [76, 77]. A partir do conhecimento destas informações pode-se argumentar que a melhor

Tabela 10.1: Informações sobre os conjuntos de dados utilizados em cada caso analisado nesta tese para treinamento e teste da abordagem proposta (SMVS) e o preditor *LBTOPE*.

Caso	Base de dados utilizada
A	SMVS e *LBTOPE Lbtope_Fixed_non_redundant*
B	SMVS e *LBTOPE Lbtope_Variable_non_redundant*
C	SMVS e *LBTOPE LBConfirm*
D	SMVS *Lb_Fix_non_redun*, *LBTOPE Lb_Fix_redun*, testados no ABCPred
E	SMVS *Lb_Fix_non_redun*, *LBTOPE Lb_Fix_redun*, testados no Chen
F	SMVS *Lb_Fix_non_redun*, *LBTOPE Lb_Fix_redun*, testados no BCPred
G	SMVS e *LBTOPE Lbtope_Fixed_redundant*
H	SMVS e *LBTOPE Lbtope_Variable_redundant*
I	SMVS *Lbtope_Fixed_redun*, *LBTOPE Lb_Fix_redun*, testados no ABCPred
J	SMVS *Lbtope_Fixed_redun*, *LBTOPE Lb_Fix_redun*, testados no Chen
L	SMVS *Lbtope_Fixed_redun*, *LBTOPE Lb_Fix_redun*, testados no BCPred

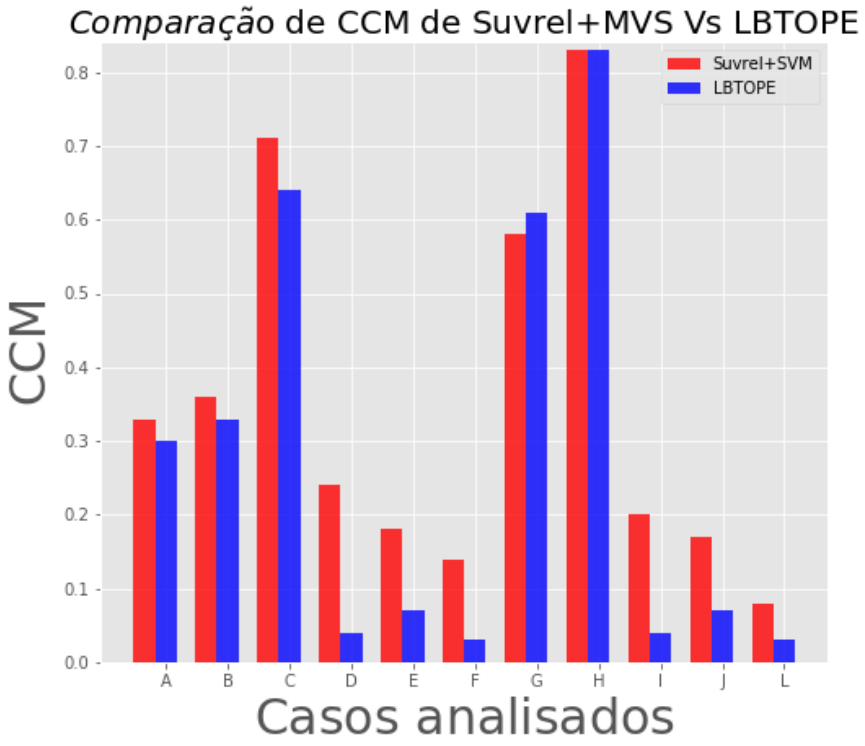
Fonte: elaborada pelo autor.

Figura 10.1: Comparação de acurácias entre a abordagem proposta (SMVS) nesta tese e o preditor *LBTOPE*. acur é a abreviação de acurácia. A, B, C, D, E, F, G, H, I, J, L representam os casos listados na Tabela 10.



Fonte: elaborada pelo autor.

Figura 10.2: Comparação de coeficientes de correlação de Matthews (CCM) entre a abordagem proposta (SMVS) nesta tese e o preditor *LBTOPE*. A, B, C, D, E, F, G, H, I, J, L representam os casos listados na Tabela 10.



Fonte: elaborada pelo autor.

alternativa é dedicar mais esforço à predição conformacional [77]. É possível sugerir por meio dos resultados aqui expostos que por mais que a predição de epítopos lineares melhore ainda não mostra resultados que possam trazer conforto na sua utilização e a melhor alternativa seja despender esforços na predição conformacional. Dessa forma, no futuro, pode-se analisar os impactos do *Suvrel* associados à predição conformacional.

REFERÊNCIAS

1 FRASER, C. M. et al. The value of complete microbial genome sequencing (you get what you pay for). *Journal of Bacteriology*, v. 184, n. 23, p. 6403–6405, 2002.

2 Cathy Yarbrough and April Thompson. *International Human Genome Sequencing Consortium Announces "Working Draft" of Human Genome*. 2000. https://www.genome.gov/10001457/2000-release-working-draft-of-human-genome-sequence/. Acessado em: 22 de nov. de 2014.

3 Sergio Danilo Pena. *Dez anos de genoma humano*. 2010. http://www.cienciahoje.org.br/noticia/v/ler/id/4322/n/dez_anos_de_genoma_humano. Acessado em: de 12 nov. 2014.

4 HUSMEIER, D.; DYBOWSKI, R.; ROBERTS, S. *Probabilistic Modelling in Bioinformatics and Medical Informatics*. [S.l.]: Springer-Verlag New York Inc, 2004. ISBN 9781852337780.

5 OTTO, T. et al. ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes. *Bioinformatics*, v. 26, n. 5, p. 705–707, 2010.

6 GOECKS, J. et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, v. 11, n. 8, p. R86, 2010.

7 OSHLACK, A.; ROBINSON, M.; YOUNG, M. From RNA-seq reads to differential expression results. *Genome Biology*, v. 11, n. 12, p. 220, 2010.

8 MARTIN, J.; WANG, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.*, v. 12, n. 10, p. 671–682, 2011.

9 OZSOLAK, F.; MILOS, P. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, v. 12, n. 2, p. 87–98, 2011. ISSN 1471-0056.

10 NAGALAKSHMI, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, v. 320, n. 5881, p. 1344–1349, 2008.

11 WANG, E. et al. Cancer systems biology in the genome sequencing era: part 1, dissecting and modeling of tumor clones and their networks. In: ELSEVIER. *Seminars in cancer biology*. [S.l.], 2013. v. 23, n. 4, p. 279–285.

12 WANG, E.; LENFERINK, A.; O'CONNOR-MCCOURT, M. Cancer systems biology: exploring cancer-associated genes on cellular networks. *arXiv preprint arXiv:0712.3753*, 2007.

13 PAVLOPOULOS, G. A. et al. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *Gigascience*, BioMed Central, v. 4, n. 1, p. 38, 2015.

14 WANICHTHANARAK, K.; FAHRMANN, J. F.; GRAPOV, D. Genomic, proteomic, and metabolomic data integration strategies. *Biomarker insights*, SAGE Publications Sage UK: London, England, v. 10, p. BMI–S29511, 2015.

15 COPLEY, T. R. et al. An integrated rnaseq-1 h nmr metabolomics approach to understand soybean primary metabolism regulation in response to rhizoctonia foliar blight disease. *BMC plant biology*, BioMed Central, v. 17, n. 1, p. 84, 2017.

16 ROBINSON, M. D.; McCarthy, D. J.; SMYTH, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, v. 26, n. 1, jan. 2010. ISSN 1367-4811. PMID: 19910308.

17 TAMBONIS, T.; BOARETO, M.; LEITE, V. B. Differential expression analysis in rna-seq data using a geometric approach. *Journal of Computational Biology*, v. 25, n. 11, p. 1257–1265, 2018.

18 Boareto, M. et al. Supervised variational relevance learning, an analytic geometric feature selection with applications to omic datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 12, n. 3, p. 705–711, May 2015.

19 HATEM, A. et al. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, v. 14, n. 1, p. 184, 2013. ISSN 1471-2105. Disponível em: http://www.biomedcentral.com/1471-2105/14/184.

20 OZSOLAK, F.; MILOS, P. M. Single-molecule direct rna sequencing without cdna synthesis. *Wiley Interdisciplinary Reviews: RNA*, John Wiley & Sons, Inc., v. 2, n. 4, p. 565–570, 2011. ISSN 1757-7012. Disponível em: http://dx.doi.org/10.1002/wrna.84.

21 TAMBONIS, T. Análise do método suvrel na expressão diferencial a partir da matriz de contagens gerada com dados de rna-seq. Universidade Estadual Paulista (UNESP), 2014.

22 FINOTELLO, F.; CAMILLO, B. D. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 2014. Disponível em: http://bfgp.oxfordjournals.org/content/early/2014/09/18/bfgp.elu035.abstract.

23 LI, H.; HOMER, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics*, v. 11, n. 5, p. 473–483, Sep 2010.

24 PICKRELL, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, v. 464, n. 7289, p. 768–772, Apr 2010.

25 RAPAPORT, F. et al. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biology*, v. 14, n. 9, p. R95, 2013. ISSN 1465-6906. Disponível em: http://genomebiology.com/2013/14/9/R95.

26 HARDCASTLE, T.; KELLY, K. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, v. 11, n. 1, p. 422, 2010. ISSN 1471-2105. Disponível em: http://www.biomedcentral.com/1471-2105/11/422.

27 ROBINSON, M.; OSHLACK, A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, v. 11, n. 3, p. R25, 2010. ISSN 1465-6906. Disponível em: http://genomebiology.com/2010/11/3/R25.

28 ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. *Genome Biology*, v. 11, n. 10, p. R106, 2010. ISSN 1465-6906. Disponível em: http://genomebiology.com/2010/11/10/R106.

29 LI, J. et al. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 2011. Disponível em: http://biostatistics.oxfordjournals.org/content/early/2011/10/13/biostatistics.kxr031.abstract.

30 SMYTH, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, v. 3, n. 1, p. 1–25, 2004.

31 LAW, C. W. et al. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, v. 15, n. 2, p. R29, 2014.

32 CASELLA, G. *Statistical inference*. Australia Pacific Grove, CA: Thomson Learning, 2002. ISBN 0-534-24312-6.

33 TARAZONA, S. et al. Differential expression in rna-seq: A matter of depth. *Genome Research*, 2011. Disponível em: http://genome.cshlp.org/content/early/2011/09/07/gr.124321.111.abstract.

34 B, M. E. Z. e Louzada Neto F. e P. B. A curva roc para testes diagnósticos. *Caderno de Saude Coletiva, Rio de Janeiro*, v. 11, n. 1, p. 7–31, 2003.

35 FAWCETT, T. An introduction to roc analysis. *Pattern Recogn. Lett.*, Elsevier Science Inc., New York, NY, USA, v. 27, n. 8, p. 861–874, jun. 2006. ISSN 0167-8655. Disponível em: http://dx.doi.org/10.1016/j.patrec.2005.10.010.

36 SAHA, S.; RAGHAVA, G. Prediction of continuous b-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function, and Bioinformatics*, Wiley Online Library, v. 65, n. 1, p. 40–48, 2006.

37 GORODKIN, J. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, Elsevier, v. 28, n. 5-6, p. 367–374, 2004.

38 POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Bioinfo Publications, 2011.

39 ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. *Imunologia celular e molecular*. [S.l.]: Elsevier Brasil, 2015.

40 LARSEN, J. E.; LUND, O.; NIELSEN, M. Improved method for predicting linear B-cell epitopes. *Immunome Res*, v. 2, p. 2, 2006.

41 GROOT, A. S. D. Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discovery Today*, v. 11, n. 5–6, p. 203 – 209, 2006. ISSN 1359-6446. Disponível em: http://www.sciencedirect.com/science/article/pii/S1359644605037207.

42 BAMBINI, S.; RAPPUOLI, R. The use of genomics in microbial vaccine development. *Drug Discovery Today*, v. 14, n. 5–6, p. 252 – 260, 2009. ISSN 1359-6446. Disponível em: http://www.sciencedirect.com/science/article/pii/S1359644608004340.

43 BATORI, V. et al. An in silico method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context. *Journal of Molecular Recognition*, John Wiley & Sons, Ltd., v. 19, n. 1, p. 21–29, 2006. ISSN 1099-1352. Disponível em: http://dx.doi.org/10.1002/jmr.752.

44 DAVIES, M. N.; FLOWER, D. R. Harnessing bioinformatics to discover new vaccines. *Drug Discovery Today*, v. 12, n. 9–10, p. 389 – 395, 2007. ISSN 1359-6446. Disponível em: http://www.sciencedirect.com/science/article/pii/S1359644607001353.

45 RESENDE, D. et al. An assessment on epitope prediction methods for protozoa genomes. *BMC Bioinformatics*, v. 13, n. 1, p. 309, 2012. ISSN 1471-2105. Disponível em: http://www.biomedcentral.com/1471-2105/13/309.

46 EL-MANZALAWY, Y.; HONAVAR, V. Recent advances in B-cell epitope prediction methods. *Immunome Res*, v. 6 Suppl 2, p. S2, 2010.

47 POTOCNAKOVA, L.; BHIDE, M.; PULZOVA, L. B. An introduction to b-cell epitope mapping and in silico epitope prediction. *Journal of immunology research*, Hindawi, v. 2016, 2016.

48 SINGH, H.; ANSARI, H. R.; RAGHAVA, G. P. Improved method for linear b-cell epitope prediction using antigen's primary sequence. *PloS one*, Public Library of Science, v. 8, n. 5, p. e62216, 2013.

49 BUNSON, M. *Encyclopedia of the Roman Empire*. [S.l.]: Infobase Publishing, 2014.

50 PARHAM, P. *The immune system*. [S.l.]: Garland Science, 2014.

51 ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. *Imunologia celular e molecular*. [S.l.]: Elsevier Brasil, 2012.

52 VITA, R. et al. The immune epitope database (IEDB) 3.0". *Nucleic Acids Res.*, v. 43, n. Database issue, p. D405–412, Jan 2015.

53 LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, Oxford University Press, v. 22, n. 13, p. 1658–1659, 2006.

54 FU, L. et al. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, Oxford University Press, v. 28, n. 23, p. 3150–3152, 2012.

55 BHASIN, M.; RAGHAVA, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *Journal of Biological Chemistry*, ASBMB, v. 279, n. 22, p. 23262–23266, 2004.

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

57 CHERVONENKIS, A.; VAPNIK, V. Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data(average risk minimization based on empirical data, showing relationship of problem to uniform convergence of averages toward expectation value). *Automation and Remote Control*, v. 32, p. 207–217, 1971.

58 VLADIMIR, V. N.; VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer Heidelberg, 1995.

59 LIMA, C. A. d. M. et al. Comitê de máquinas: uma abordagem unificada empregando máquinas de vetores-suporte. [sn], 2004.

60 GUNN, S. R. et al. Support vector machines for classification and regression. *ISIS technical report*, v. 14, n. 1, p. 5–16, 1998.

61 FERREIRA, R. H. d. S. Uma metodologia para a detecção de mudanças em imagens multitemporais de sensoriamento remoto empregando support vector machines. 2014.

62 PREMALATHA, M.; LAKSHMI, C. V. Svm trade-off between maximize the margin and minimize the variables used for regression. *International Journal of Pure and Applied Mathematics*, Academic Publications, Ltd., v. 87, n. 6, p. 741–750, 2013.

63 LIMA, A. R. G. Máquinas de vetores suporte na classificaçao de impressoes digitais. *Universidade Federal do Ceará, Departamento de Computação, Fortaleza-Ceará*, 2002.

64 FLETCHER, T. Support vector machines explained. *Tutorial paper*, 2009.

65 CRISTIANINI, N.; SHAWE-TAYLOR, J. et al. *An introduction to support vector machines and other kernel-based learning methods*. [S.l.]: Cambridge university press, 2000.

66 HAYKIN, S. *Neural networks: a comprehensive foundation*. [S.l.]: Prentice Hall PTR, 1994.

67 CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Acm, v. 2, n. 3, p. 27, 2011.

68 HERBRICH, R. *Learning kernel classifiers: theory and algorithms*. [S.l.]: MIT press, 2001.

69 DING, H. et al. Prediction of golgi-resident protein types by using feature selection technique. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 124, p. 9–13, 2013.

70 CHOU, K.-C.; SHEN, H.-B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic k-nearest neighbor classifiers. *Journal of Proteome Research*, ACS Publications, v. 5, n. 8, p. 1888–1897, 2006.

71 CHOU, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, Elsevier, v. 273, n. 1, p. 236–247, 2011.

72 SIKIC, K.; CARUGO, O. Protein sequence redundancy reduction: comparison of various method. *Bioinformation*, Biomedical Informatics Publishing Group, v. 5, n. 6, p. 234, 2010.

73 HSU, C.-W. et al. A practical guide to support vector classification. Taipei, 2003.

74 EL-MANZALAWY, Y.; DOBBS, D.; HONAVAR, V. Predicting linear b-cell epitopes using string kernels. *Journal of Molecular Recognition: An Interdisciplinary Journal*, Wiley Online Library, v. 21, n. 4, p. 243–255, 2008.

75 CHEN, J. et al. Prediction of linear b-cell epitopes using amino acid pair antigenicity scale. *Amino acids*, Springer, v. 33, n. 3, p. 423–428, 2007.

76 KRINGELUM, J. V. et al. Reliable b cell epitope predictions: impacts of method development and improved benchmarking. *PLoS computational biology*, Public Library of Science, v. 8, n. 12, p. e1002829, 2012.

77 ANDERSEN, P. H.; NIELSEN, M.; LUND, O. Prediction of residues in discontinuous b-cell epitopes using protein 3d structures. *Protein Science*, Wiley Online Library, v. 15, n. 11, p. 2558–2567, 2006.