RESSALVA

Atendendo solicitação do autor,
o texto completo desta tese será
disponibilizado somente a partir
de 24/10/2021

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 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*.

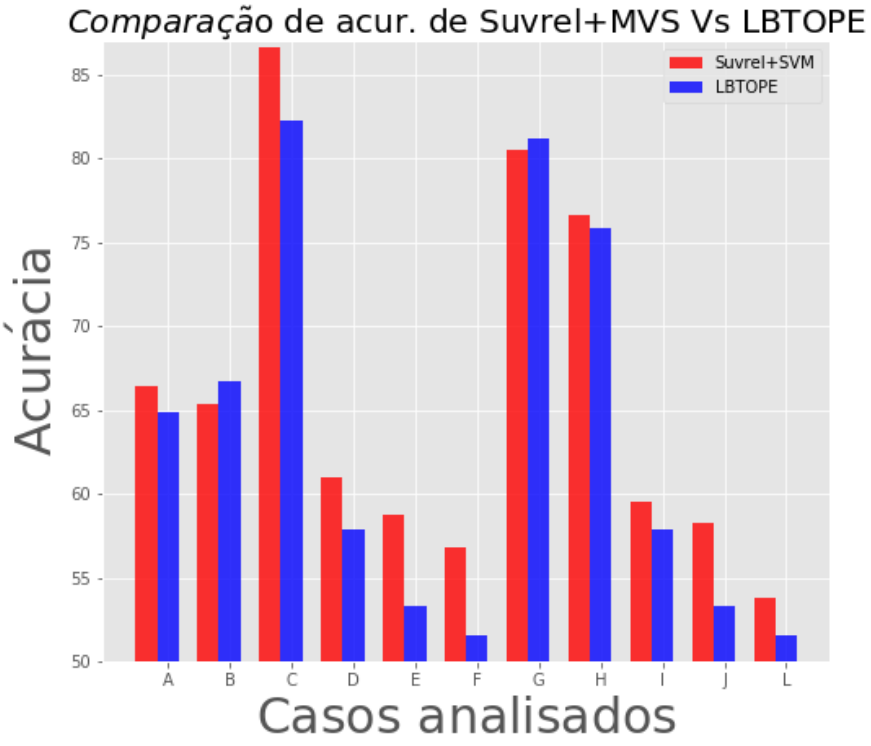
A maioria dos epítopos de células B são conformacionais, e os preditores baseados em estrutura possuem performance superior aos baseados em sequência [76, 77]. A partir do conhecimento destas informações pode-se argumentar que a melhor

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

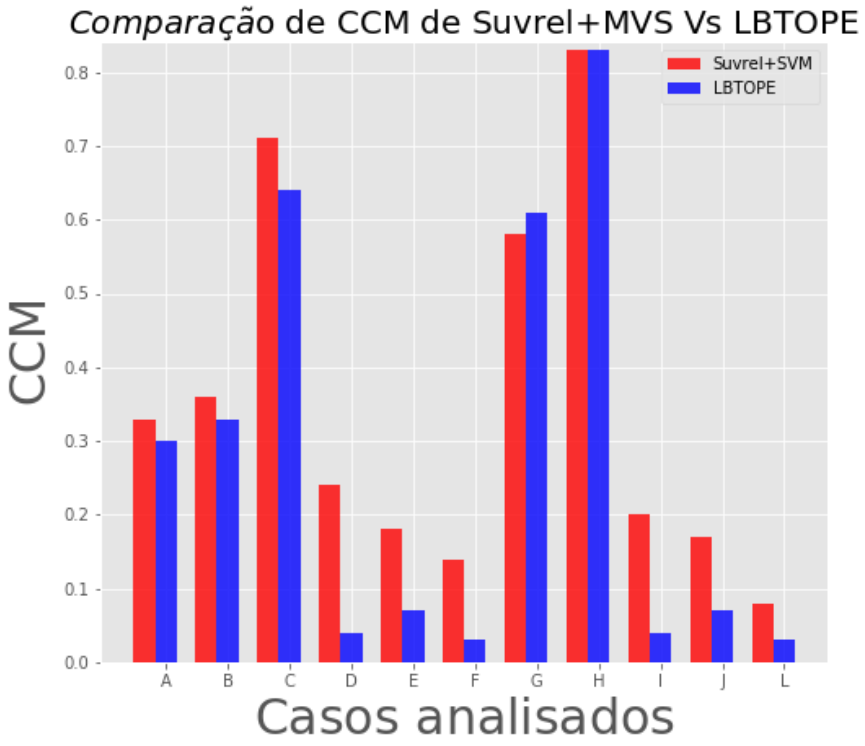
Fonte: elaborada pelo autor.

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.