Desempenho de ferramentas genotípicas e *stacking* na predição de tropismo do subtipo C do HIV-1

Rita de Cássia Menezes Soares[1], Letícia Martins Raposo[1]

[1] Centro de Ciências Exatas e Tecnologia (CCET) – Universidade Federal do Estado do Rio de Janeiro (UNIRIO) – RJ – Brasil

{rita.soares, leticia.raposo}@uniriotec.br

Abstract. *Several tools developed to classify the tropism of HIV-1 have been designed based on strains of subtype B and may not perform satisfactorily for other subtypes. The present study evaluated the performance of genotypic algorithms in predicting the tropism of HIV-1 subtype C and applied the stacking technique to seek a model with better performance. Raymond's Rule showed better overall performance, but Geno2Pheno 0.20 had greater sensitivity. The proposed model had performance equal to Geno2Pheno 0.10, with sensitivity and specificity greater than 90%. The stacking technique can be useful to improve the prediction of tropism without new tests.*

Resumo. *Diversas ferramentas desenvolvidas para classificar o tropismo do HIV-1 foram projetadas com base em cepas do subtipo B e, portanto, podem não apresentar desempenhos satisfatórios para outros subtipos. O presente estudo avaliou o desempenho de algoritmos genotípicos na predição do tropismo do HIV-1 subtipo C e aplicou a técnica de stacking a fim de buscar um modelo com melhor desempenho. A Regra de Raymond apresentou melhor desempenho geral, porém o Geno2Pheno 0,20 teve maior sensibilidade. O modelo proposto apresentou desempenho igual ao Geno2Pheno 0,10, com sensibilidade e especificidade maiores que 90%. A técnica de stacking pode ser útil para melhorar a predição do tropismo sem novos testes.*

1. Introdução

A AIDS, síndrome da imunodeficiência humana, causada pelo vírus HIV, afetava 900 mil pessoas no Brasil em 2018 [GHO 2018]. Para infectar a célula humana, o HIV-1 depende de dois receptores celulares: o CD4, receptor primário, e um correceptor, CCR5 ou CXCR4, que interage diretamente com a terceira região hipervariável (alça V3) da proteína gp120 do envelope viral. Denominam-se R5 os vírus que utilizam o correceptor CCR5; X4 os que apresentam tropismo pelo CXCR4, e de R5X4, ou dual-trópicos, os que utilizam qualquer um deles [Cabral 2014].

Se um paciente apresentar mais de 2% de vírus com tropismo a CXCR4 em sua população viral, o uso de medicamentos inibidores de ligação ao CCR5, como o Maraviroque, será impossibilitado devido à seleção positiva dessas primeiras variantes, associadas à progressão mais rápida da doença e pior prognóstico [Swenson et al. 2012].

Para identificar o tropismo do HIV-1 existem métodos fenotípicos e genotípicos. Os genotípicos são capazes de prever o uso de correceptores pelas sequências de aminoácidos da alça V3, sendo uma alternativa barata e rápida à fenotipagem, que,

apesar de ser considerada melhor, tem aplicação clínica limitada [Cabral 2014]. A maioria dos algoritmos genotípicos foi desenvolvida com base no subtipo B do HIV-1, dos mais comuns no mundo [Gräf e Pinto 2013], e pode não ser ideal para outros.

O presente estudo avalia o desempenho das principais ferramentas genotípicas no subtipo C, que tem alta prevalência no Brasil [Gräf e Pinto 2013], e propõe o desenvolvimento de um metaclassificador com *stacking* desses algoritmos.

2. Materiais e métodos

2.1. Dados

Sequências de DNA da região V3 da gp120 foram obtidas com informações fenotípicas de tropismo [Los Alamos 2019] no formato FASTA. Os dados foram lidos no *software* R, versão 3.6.0 [R Core Team 2019]. Após tradução dos códons do DNA em aminoácidos com escala padrão NCBI, sequências V3 que não iniciavam e terminavam com cisteína (C) foram excluídas da análise [Chiou et al. 1992]. Cada sequência de aminoácidos foi alinhada ao consenso ancestral da região V3 do subtipo C [Los Alamos 2019] com algoritmo Needleman-Wunsch, parâmetros padrões. Após este passo, apenas cadeias com 28 a 35 aminoácidos foram utilizadas.

Duplicatas foram removidas, totalizando 562 sequências únicas de V3 (507 R5, 18 X4 e 37 R5X4). Para padronização com outros métodos genotípicos, as variantes X4 e R5X4 foram agrupadas na classe NR5 (não-R5). O pré-processamento das sequências foi realizado com as bibliotecas seqinr [Charif e Lobry 2007], Biostrings [Pagès et al. 2019] e stringr [Wickham 2019] do R.

Os dados foram divididos em conjuntos mutuamente exclusivos: 70% para treinamento e 30% para teste, retendo a proporção das classes originais (cerca de 10% de NR5). Foi aplicado SMOTE [Chawla et al. 2002] no conjunto de treinamento para lidar com o desbalanceamento das classes, obtendo-se cerca de 57% e 43% de vírus R5-trópicos e NR5-trópicos, respectivamente. O SMOTE, executado com o pacote DMwR do R [Torgo 2010], realiza *undersampling* da classe majoritária e *oversampling* da minoritária (NR5), criando observações sintéticas.

Para obter reprodutibilidade, trabalhou-se com uma semente específica: 9992. O conjunto de dados original foi disponibilizado para livre uso [Menezes e Raposo 2020].

2.2. Algoritmos

Seis algoritmos publicamente disponíveis foram aplicados em cada uma das sequências: Geno2Pheno (G2P) [Lengauer et al. 2007], Web PSSM [Jensen et al. 2006], PhenoSeq [Cashin et al. 2015], T-CUP 2.0 [Heider et al. 2014], AUTO-MUTE (todas as abordagens disponíveis: *Random Forest* (RF), *Support Vector Machine* (SVM), *Boosted Decision Tree* (BDT) e *Neural Network* (NN)) [Masso e Vaisman 2010] e Regra de Raymond [Raymond et al. 2008]. Todos os algoritmos foram executados com valores padrões ou específicos para o subtipo C quando possível.

G2P e Web PSSM utilizavam como entrada o DNA, permitindo gaps. PhenoSeq utilizava DNA sem gaps, e estes foram removidos. Nos três algoritmos, se houvesse mais de uma sequência de aminoácidos prevista para o mesmo DNA e uma delas fosse

X4, essa classificação era utilizada. Para o G2P, duas taxas de falsos positivos foram utilizadas: 20% e 10%. Valores abaixo do ponto de corte foram considerados NR5.

T-CUP 2.0, AUTO-MUTE e Regra de Raymond necessitavam da sequência de aminoácidos como valor de entrada. Caso a sequência tivesse gaps, esses foram substituídos pelo valor do consenso nas posições. T-CUP 2.0 foi executado a partir do pacote TCUP2 do software R com ponto de corte de 20% para a taxa de falso positivo. Para o cálculo da regra de Raymond, as cargas dos aminoácidos foram obtidas do banco de dados AAindex [Kawashima et al. 1999], índice KLEP840101, e somadas.

2.3. *Stacking*

O *stacking* é uma técnica de *ensemble learning* na qual é criado um metaclassificador que aprende com o conjunto de decisões tomadas por classificadores anteriores [Oza e Tumer 2008]. Os classificadores base usados foram todos os algoritmos genotípicos e a técnica aplicada no aprendizado foi árvore de decisão binária do tipo CART (*Classification and Regression Tree*) [Kuhn et al. 2019]. No CART, em cada nó é feita uma divisão dos dados de modo a otimizar algum critério. O metaclassificador foi gerado com o pacote Caret [Kuhn et al. 2019] e o método de treinamento foi "rpart2".

2.4. Avaliação dos algoritmos

Os algoritmos foram comparados com as seguintes medidas, calculadas pelo Caret: acurácia (Ac), sensibilidade (Sen), especificidade (Esp) e coeficiente de correlação de Matthews (MCC). O MCC é uma medida análoga ao coeficiente de correlação de Pearson e varia entre -1 e 1. Não sofre influência de dados desbalanceados e é uma das melhores métricas para avaliar o desempenho de modelos [Powers 2011].

3. Resultados e discussão

Conforme Tabela 1, todos os métodos atingiram uma especificidade superior a 82%, com a Regra de Raymond mostrando o melhor valor (100%). A maior sensibilidade foi alcançada pelo G2P 0,20 (100%). A acurácia das abordagens variou de 83,93% a 97,62%. A Regra de Raymond demonstrou o melhor desempenho geral, comparável aos resultados de Riemenschneider et al (2016). A sensibilidade foi a métrica de maior variabilidade, com valores de 37,50% a 100%. Todos os algoritmos propostos pelo AUTO-MUTE apresentaram valores baixos de sensibilidade, não superando 57%. A baixa sensibilidade proporcionará muitos falsos negativos (NR5 classificados como R5), levando a uso incorreto de medicamentos, o que pode piorar o quadro clínico.

Tabela 1. Desempenho de algoritmos no conjunto teste (maiores valores em negrito).

Teste	Ac (%)	Esp (%)	Sen (%)	MCC
Web PSSM (C Sinsi)	83,93	82,89	93,75	0,5238
PhenoSeq C	85,71	85,53	87,50	0,5224
T-CUP 2.0	96,43	97,37	87,50	0,8055
G2P (ponto de corte 0,20)	86,31	84,87	**100,00**	0,5901
G2P (ponto de corte 0,10)	94,05	94,08	93,75	0,7368
AUTO-MUTE BDT	94,05	99,34	43,75	0,5940

AUTO-MUTE RF	92,86	98,68	37,50	0,4988
AUTO-MUTE NN	88,10	93,42	37,50	0,3092
AUTO-MUTE SVM	93,45	97,37	56,25	0,5891
Regra de Raymond	**97,62**	**100,00**	75,00	**0,8549**
Metaclassificador	94,05	94,08	93,75	0,7368

Na Figura 1 está representada a árvore de decisão do metaclassificador. Apenas três classificadores base compuseram a árvore: T-CUP 2.0, G2P 0,10 e a Regra de Raymond, os de melhor MCC. O metaclassificador se igualou ao G2P 0,10, o único a apresentar sensibilidade e especificidade acima de 90%. Outros métodos de *stacking* podem ter ainda desempenho superior aos dos algoritmos genotípicos.

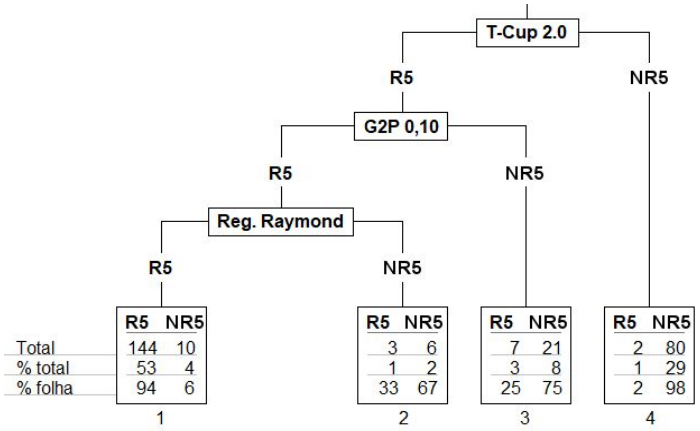

Figura 1. Árvore de decisão gerada por *stacking* dos algoritmos genotípicos.

Como limitação, o SMOTE pode ter gerado observações indesejadas nos dados de treino. Também não foi feita nenhuma otimização dos hiperparâmetros do metaclassificador, utilizando-se apenas os valores padrões do algoritmo no Caret.

4. Conclusão

O presente estudo mostrou que, apesar dos algoritmos genotípicos apresentarem alta especificidade, alguns não apresentam sensibilidade razoável na predição do tropismo no subtipo C. As maiores sensibilidades foram do Geno2Pheno 0,20 e 0,10 e Web PSSM. A árvore de decisão proposta teve desempenho igual ao Geno2Pheno 0,10, com sensibilidade e especificidade maiores que 90%. O *stacking* pode ser útil para melhorar a predição sem criação de novos testes. Futuramente, outras abordagens para lidar com o desbalanceamento serão avaliadas e outro algoritmos serão utilizados no *stacking*.

5. Reconhecimento

As autoras agradecem à Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ – pelo apoio financeiro ao trabalho.

6. Contribuição das autoras

Desenho e concepção do estudo: RM e LR. Aquisição, interpretação e análise dos

dados: RM e LR. Escrita e revisão do artigo: RM e LR. Supervisão: LR.

Referências

CABRAL, G. B. Avaliação da resposta à terapia antirretroviral de resgate contendo antagonista do correceptor CCR5 em pessoas vivendo com HIV/AIDS. São Paulo, 2014. Disponível em: <http://ses.sp.bvs.br/lildbi/docsonline/get.php?id=6117>. Acesso em: 13 dez. 2019.

CASHIN, K.; GRAY, L. R.; HARVEY, K. L.; et al. Reliable Genotypic Tropism Tests for the Major HIV-1 Subtypes. Scientific Reports, v. 5, n. 1, p. 1–8, 2015.

CHARIF, D.; LOBRY, J. R. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: BASTOLLA, U.; PORTO, M.; ROMAN, H. E.; et al (Orgs.). Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag, 2007, p. 207–232. (Biological and Medical Physics, Biomedical Engineering).

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, v. 16, p. 321–357, 2002.

CHIOU, S. H.; FREED, E. O.; PANGANIBAN, A. T.; et al. Studies on the role of the V3 loop in human immunodeficiency virus type 1 envelope glycoprotein function. AIDS research and human retroviruses, v. 8, n. 9, p. 1611–1618, 1992.

GHO - Global Health Observatory | By category | Number of people (all ages) living with HIV - Estimates by country. Disponível em: <http://apps.who.int/gho/data/view.main.22100?lang=en>. Acesso em: 25 out. 2019.

GRÄF, T.; PINTO, A. R. The increasing prevalence of HIV-1 subtype C in Southern Brazil and its dispersion through the continent. Virology, v. 435, n. 1, p. 170–178, 2013.

HEIDER, D.; DYBOWSKI, J. N.; WILMS, C.; et al. A simple structure-based model for the prediction of HIV-1 co-receptor tropism. BioData Mining, v. 7, p. 14, 2014.

JENSEN, M. A.; COETZER, M.; VAN'T WOUT, A. B.; et al. A Reliable Phenotype Predictor for Human Immunodeficiency Virus Type 1 Subtype C Based on Envelope V3 Sequences. Journal of Virology, v. 80, n. 10, p. 4698–4704, 2006.

KAWASHIMA, S.; OGATA, H.; KANEHISA, M. AAindex: Amino Acid Index Database. Nucleic Acids Research, v. 27, n. 1, p. 368–369, 1999.

KUHN, M.; WING, J.; WESTON, S.; et al. caret: Classification and Regression Training. [s.l.: s.n.], 2019. Disponível em: <https://CRAN.R-project.org/package=caret>. Acesso em: 23 jun. 2020.

LENGAUER, T.; SANDER, O.; SIERRA, S.; et al. Bioinformatics prediction of HIV coreceptor usage. Nature Biotechnology, v. 25, n. 12, p. 1407–1410, 2007.

Los Alamos. HIV Databases. Disponível em: <https://www.hiv.lanl.gov/content/index>. Acesso em: 14 dez. 2019.

MASSO, M.; VAISMAN, I. I. Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage. BMC Bioinformatics, v. 11, p. 494, 2010.

MENEZES, R.; RAPOSO, L. HIV Tropism Ensemble Methods. 2020. Disponível em: <https://doi.org/10.5281/zenodo.3905343>. Acesso em: 23 jun. 2020.

OZA, N. C.; TUMER, K. Classifier ensembles: Select real-world applications. Information Fusion, v. 9, n. 1, p. 4–20, 2008.

PAGÈS, H.; ABOYOUN, P.; GENTLEMAN, R.; et al. Biostrings: Efficient manipulation of biological strings. [s.l.: s.n.], 2019.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies, v. 2, n. 1, p. 37–63, 2011.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2019. Disponível em: <https://www.R-project.org/>. Acesso em: 23 jun. 2020.

RAYMOND, S.; DELOBEL, P.; MAVIGNER, M.; et al. Correlation between genotypic predictions based on V3 sequences and phenotypic determination of HIV-1 tropism. AIDS (London, England), v. 22, n. 14, p. F11-16, 2008.

RIEMENSCHNEIDER, M.; CASHIN, K. Y.; BUDEUS, B.; et al. Genotypic Prediction of Co-receptor Tropism of HIV-1 Subtypes A and C. Scientific Reports, v. 6, n. 1, p. 1–9, 2016.

SWENSON, L. C.; DÄUMER, M.; PAREDES, R. Next-generation sequencing to assess HIV tropism. Current opinion in HIV and AIDS, v. 7, n. 5, p. 478–485, 2012.

TORGO, L. Data Mining with R, learning with case studies. [s.l.]: Chapman and Hall/CRC, 2010. Disponível em: <http://www.dcc.fc.up.pt/ltorgo/DataMiningWithR>. Acesso em: 23 jun. 2020.

WICKHAM, H. stringr: Simple, Consistent Wrappers for Common String Operations. [s.l.: s.n.], 2019. Disponível em: <https://CRAN.R-project.org/package=stringr>. Acesso em: 23 jun. 2020.