MODELOS PREDITIVOS DE RESISTÊNCIA AO INIBIDOR DA HIV-PROTEASE LOPINAVIR

L. M. Raposo*, M. B. Arruda**, R. M. Brindeiro**, F. F. Nobre*

*Programa de Engenharia Biomédica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
** Laboratório de Virologia, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
e-mail: raposo@peb.ufrj.br

Resumo: O HIV é o agente etiológico responsável por uma das mais importantes doenças conhecidas, a Aids. Com o intuito de controlar os índices de morbidade e mortalidade, foram desenvolvidas as terapias antirretrovirais. Apesar de elas promoverem reduções nestes índices, alguns pacientes não têm apresentado um benefício clínico durável devido ao problema da resistência. Nesse estudo, o objetivo foi o de desenvolver modelos de predição de resistência ao inibidor da HIV-protease Lopinavir para auxiliar os médicos na avaliação da terapia e na escolha do antirretroviral mais adequado. Foram utilizadas duas técnicas de modelagem: a regressão logística e a rede neural probabilística. Os modelos apresentaram bons desempenhos com valores de acurácia variando de 89,6% a 93,6%, sensibilidades de 94,3% a 100% e especificidades de 87,8% a 92,2%.

Palavras-chave: Classificadores, Aprendizagem de Máquinas, Resistência.

Abstract: *HIV is the etiological agent responsible for one of the major known diseases, AIDS. Aiming to control the morbidity and mortality, antiretroviral therapies have been developed. Although they promote reductions in these indices, some patients have not presented a lasting clinical benefit due to the resistance problem.. In this study the aim was to develop models for predicting resistance to the HIV-protease inhibitor Lopinavir to assist clinicians in evaluating the therapy and choose the most adequate antiretroviral. Two modeling techniques were used: logistic regression and probabilistic neural network. The models showed good performance with accuracy values ranging from 89.6% to 93.6%, sensitivity between 94.3% and 100% and specificity of 87.8% to 92.2%.*

Keywords: *Classifiers, Machine Learning, Resistance.*

Introdução

O vírus da imunodeficiência humana (HIV) é o agente etiológico responsável por uma das principais doenças no mundo, a Aids. Com o intuito de reduzir os índices de morbidade e mortalidade, assim como um aumentar a qualidade de vida dos pacientes, foram desenvolvidos os antirretrovirais [1]. Em alguns pacientes, esta terapia não tem apresentado um benefício clínico durável, decorrente dos problemas de resistência aos medicamentos [2,3].

A resistência pode ser identificada pela genotipagem e/ou a fenotipagem. O primeiro teste determina mutações genéticas no HIV associadas à resistência aos antirretrovirais, sendo um teste rápido, menos custoso e mais acessível [4]. A fenotipagem fornece uma medida quantitativa direta da suscetibilidade do HIV aos antirretrovirais. É um teste muito caro, mais complexo e com uma demanda de tempo maior para gerar resultados [4]. Dessa forma, a utilização do genótipo no desenvolvimento de modelos capazes de predizer a resposta de um paciente a um medicamento pode ser uma boa alternativa na determinação de resistência.

Estudos vêm sendo desenvolvidos visando a determinação de um bom modelo de predição fazendo uso de técnicas de modelagem, tais como regressão linear [5], redes neurais [6,7], máquina de vetores de suporte [8] e árvores de decisão [9].

O objetivo deste trabalho foi desenvolver modelos preditivos de resistência ao inibidor da HIV-protease Lopinavir (LPV), usando as técnicas de regressão logística (RL) e a rede neural probabilística (PNN).

Materiais e métodos

O banco de dados utilizado é constituído por sequências de aminoácidos da enzima protease do gene pol (polimerase) do HIV-1, subtipo B, de 625 pacientes, provenientes das regiões sul, sudeste, centro-oeste e nordeste, infectados por este vírus. Os dados foram cedidos pelo Laboratório de Virologia Molecular do Centro de Ciências da Saúde da Universidade Federal do Rio de Janeiro (CCS - UFRJ/Brasil), integrante da rede de laboratórios de genotipagem do Ministério da Saúde (RENAGENO).

Foi considerado como variável resposta a resistência ao inibidor LPV. Para os pacientes que, no último regime terapêutico, não apresentaram resistência a essa droga ou o nível foi considerado intermediário, a variável resposta foi codificada com o valor 0, enquanto que aqueles que apresentaram resistência ao antirretroviral, a variável foi codificada com o valor 1. Essas classificações foram fornecidas pelo Algoritmo Brasileiro, um software de interpretação de genotipagem com a função de localizar as mutações genéticas no HIV e, por meio de um conjunto de regras pré-estabelecidas, indicar a quais antirretrovirais o vírus apresenta resistência.

As variáveis explicativas selecionadas inicialmente para o desenvolvimento dos modelos foram as posições com mutações mais frequentes encontradas no gene da HIV-protease associadas à resistência ao LPV, com base na lista da Sociedade Internacional de AIDS (IAS) [10]. A HIV-protease é composta por 99 aminoácidos e as posições são representadas por uma notação que consiste na letra correspondente ao nome do aminoácido da sequência consenso, seguido pelo número referente à posição do aminoácido. As posições selecionadas foram: L10, K20, L24, V32, L33, M46, I47, I50, F53, I54, L63, A71, G73, L76, V82, I84 e L90.

A codificação dos aminoácidos foi realizada usando uma representação binária. As posições que apresentaram alguma mutação foram codificadas com valor 1 enquanto que as posições que possuíam o aminoácido na forma selvagem receberam valor 0.

Das 625 amostras selecionadas para o estudo, 500 constituíram o conjunto de treinamento, utilizado no desenvolvimento dos modelos e na seleção das variáveis explicativas e 125 foram alocadas ao conjunto de teste. No grupo de treinamento, 400 pacientes não apresentavam resistência ao LPV, enquanto que 100 eram resistentes. No grupo de teste, 30 pacientes eram resistentes e 95 não apresentavam resistência a esse medicamento. A seleção das variáveis foi realizada usando as técnicas de regressão logística, validação cruzada do tipo *k-fold* e *bootstrap*.

Das 100 amostras resistentes do conjunto de treino, foram obtidas 100 amostras por *bootstrap* que foram combinadas com 100 amostras da classe de não resistentes do conjunto de treino selecionadas aleatoriamente com reposição, resultando em um conjunto equilibrado de 200 amostras. Esse procedimento foi repetido 1000 vezes, resultando em 1000 subconjuntos. Para cada um desses subconjuntos, um modelo logístico foi desenvolvido e as variáveis de cada um desses modelos foram selecionadas pelo método *stepwise*, usando o critério de informação de Akaike (AIC). Variáveis presentes em mais de 50% dos modelos foram selecionadas para compor os modelos finais.

Para a seleção do fator de alisamento das redes PNN, foram obtidos 100 subconjuntos balanceados de 200 amostras da mesma maneira descrita anteriormente. Para cada um desses subconjuntos, foram implementadas redes PNN com fatores de alisamento (variância) variando de 0,1 a 1 com passos de 0,1. A área sob a curva ROC (AUC) média das redes para cada fator de alisamento foi calculada através da técnica de validação cruzada do tipo *k-fold* com k igual a 10. Este procedimento foi repetido para cada uma das 100 amostras *bootstrap*, e o fator de alisamento com maior AUC média foi selecionado para cada amostra. O fator de alisamento final foi definido como a média dos 100 melhores fatores de alisamento obtidos.

Definidas as variáveis e o fator de alisamento da rede, foram obtidos quatro modelos PNN a partir do conjunto de treino, sendo cada um deles gerado a partir de 100 indivíduos não resistentes selecionados aleatoriamente de um total de 400 e 100 resistentes.

Os modelos foram avaliado usando a acurácia, sensibilidade, especificidade, AUC e índice *Kappa*. Para obter os valores de acurácia, sensibilidade e especificidade, o ponto de corte estipulado foi de 0,5.

Os modelos obtidos foram comparados com os algoritmos de interpretação HIVdb (versão 7.0) [11], Rega (versão 9.1.0) [12] e ANRS (*Agence Nationale de Recherche sur le Sida*) (versão 2013.09) [13], cujos desempenhos foram avaliados usando o mesmo conjunto de teste utilizado .

Softwares – Os dados foram compilados em planilha eletrônica do programa Microsoft Excel® e para as análises e desenvolvimento dos modelos foram utilizados os softwares R versão 3.0.1 e MATLAB versão R2009b.

Resultados

Quatro modelos preditivos de resistência ao LPV foram desenvolvidos com base nas metodologias de RL e redes PNN, fazendo uso das técnicas de *bootstrap* e a validação cruzada.

Através da RL foi possível selecionar dez posições significativas de mutação relacionadas à resistência ao LPV. Foram elas: A71 (64,2%), I54 (95,0%), I84 (89,7%), K20 (88,6%), L10 (92,7%), L24 (76,6%), L33 (50,1%), L90 (93,9%), M46 (85,6%) e V82 (94,3%). O fator de alisamento obtido pela média dos 100 melhores fatores das redes PNN foi igual a 0,63.

Os quatro modelos obtidos foram avaliados utilizando o conjunto de teste. A avaliação se baseou na AUC, acurácia, sensibilidade, especificidade e índice Kappa. A Tabela 1 mostra esses indicadores para cada modelo e na Figura 1 estão representadas as curvas ROC.

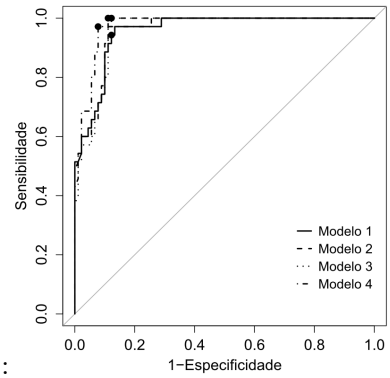

Figura 1: Curvas ROC dos quatro modelos obtidos. Os círculos correspondem aos pontos de corte de 0,5.

Tabela 1: Desempenho dos modelos obtidos pela rede neural probabilística.

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**
AUC	0,96	0,96	0,96	0,97
Acurácia (%)	89,6	92,0	91,2	93,6
Sensibilidade (%)	94,3	100	100	97,1
Especificidade (%)	87,8	88,9	87,8	92,2
Índice Kappa	0,76	0,82	0,80	0,85

Os algoritmos HIVdb, Rega e ANRS classificaram os dados em três níveis de resistência. Os pacientes classificados em suscetível e intermediário foram agrupados no grupo dos não resistentes enquanto que aqueles que foram classificados como resistentes permaneceram nesta classe. Esse agrupamento foi feito dessa maneira para acompanhar a classificação utilizada pelo Algoritmo Brasileiro, utilizado como referência neste estudo.

A Tabela 2 apresenta o desempenho desses algoritmos. Pode-se observar que o sistema HIVdb e o ANRS apresentaram desempenhos próximos.

Tabela 2: Desempenho dos algoritmos HIVdb, Rega e ANRS.

	HIVdb	**Rega**	**ANRS**
AUC	0,91	0,74	0,94
Acurácia (%)	0,93	0,86	0,97
Sensibilidade (%)	0,86	0,49	0,89
Especificidade (%)	0,96	100	100

Indice Kappa	0,82	0,58	0,92

Discussão

Neste estudo, foram desenvolvidos quatro grupos de classificadores de resistência aos inibidores da HIV-protease LPV baseando-se em duas metodologias (RL e PNN).

A seleção das variáveis significativas foi realizada por meio das técnicas de *bootstrap* e *stepwise* na RL. A combinação dessas duas técnicas foi aplicada com o objetivo de escolher as variáveis mais significantes, possibilitando uma melhor seleção e consequentemente, uma diminuição do erro de classificação.

Outra característica deste estudo foi a utilização de dados balanceados em relação ao desfecho (resistentes e não resistentes), cujo objetivo foi o de evitar o surgimento de uma maior tendência dos modelos em responder bem para as classes majoritárias em detrimento das minoritárias. Geralmente, a construção de modelos com dados desbalanceados apresenta o risco de se obter soluções com um bom desempenho apenas em áreas com maior abundância de observação nos dados de entrada, prejudicando características importantes do modelo, como a robustez e capacidade de generalização [14].

No estudo de Rhee et al. (2006) [15], a rede neural do tipo *feed-forward* foi um dos métodos utilizados na construção dos modelos, fazendo uso de um conjunto completo com 70 posições da HIV-protease e um conjunto de mutações selecionadas pela lista da IAS. Para o LPV, a acurácia foi igual a 0,76 para o conjunto completo de posições e 0,73 para o conjunto lista da IAS, valores inferiores aos encontrados neste estudo.

Pasomsub et al. (2010) [16] também desenvolveu redes neurais artificiais para predizer o fenótipo a partir das sequências genotípicas. Para o LPV, a AUC foi igual a 0,92 (0,88 - 0,95). Neste presente estudo, em termos de AUC, os modelos apresentaram valores variando de 0,96 a 0,97.

A maioria dos estudos não apresenta outras métricas além da acurácia, tais como a sensibilidade e especificidade. Isso limita a interpretação do desempenho do modelo e a comparação com os resultados deste estudo.

Utilizando o mesmo conjunto de teste deste trabalho, três algoritmos de interpretação foram avaliados: HIVdb, Rega e ANRS. Os modelos do presente estudo apresentaram desempenhos superiores ao do algoritmo Rega, e com relação aos demais, os desempenhos foram muito próximos.

Nossos resultados têm apresentado valores superiores e/ou próximos a de estudos anteriores, o que demonstra que a metodologia pode ser uma boa alternativa na classificação de pacientes em termos de resistência ao LPV.

Conclusão

Apesar das limitações, os modelos propostos neste trabalho representam uma ferramenta auxiliar inicial na classificação de novos indivíduos em relação ao desenvolvimento de resistência aos inibidores da HIV-protease, podendo se tornar úteis na escolha da melhor prática terapêutica para cada indivíduo HIV+.

Agradecimentos

Os autores agradecem às agências financiadoras Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (Faperj).

Referências

[1] Bushman F, L, au N, Emini E. New developments in the biology and treatment of HIV. Proceedings of the National Academy of Sciences. 1998; 95(19):11041-11042.

[2] Fätkenheuer G, Theisen A, Rockstroh J, Grabow T, Wicke C, Becker K et al. Virological treatment failure of protease inhibitor therapy in an unselected cohort of HIV-infected patients. Aids. 1997; 11(14):113-116.

[3] Zolopa A, Shafer R, Warford A, Montoya J, Hsu P, Katzenstein D et al. HIV-1 genotypic resistance patterns predict response to saquinavir-ritonavir therapy in patients in whom previous protease inhibitor therapy had failed. Annals of internal medicine. 1999; 131(11):813-821.

[4] Wang D, Larder B. Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. Journal of Infectious Diseases. 2003; 188(5):653-660.

[5] Van der Borght K, Verheyen A, Feyaerts M, Van Wesenbeeck L, Verlinden Y, Van Craenenbroeck E et al. Quantitative prediction of integrase inhibitor resistance from genotype through consensus linear regression modeling. Virology journal. 2013; 10(1):8.

[6] Draghici S, Potter R. Predicting HIV drug resistance with neural networks. Bioinformatics. 2003; 19(1):98-107.

[7] Pasosub E, Sukasem C, Sungkanuparph S, Kijsirikul B, Chantratita W. The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems. Japanese journal of infectious diseases. 2010; 63(2):87-94.

[8] Beerenwinkel N, Däumer M, Oette M, Korn K, Hoffmann D, Kaiser R et al. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic acids research. 2003; 31(13):3850-3855.

[9] Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proceedings of the National Academy of Sciences. 2002; 99(12):8271-8276.

[10] Johnson V, Calvez V, Gunthard H, Paredes R, Pillay D, Shafer R et al. Update of the drug resistance mutations in HIV-1: March 2013. Top Antivir Med. 2013; 21(1):6-14.

[11] Liu T, Shafer R. Web resources for HIV type 1 genotypic-resistance test interpretation. Clinical infectious diseases. 2006; 42(11):1608-1618.

[12] Rega.kuleuven.be. Laboratory for Clinical and Epidemiological Virology – Rega Institute KU Leuven [Internet]. 2014 [1 May 2014]. Available from: http://rega.kuleuven.be/cev/

[13] Medpocket.com. MedPocket Genotype - HIV genotypic drug resistance interpretation - Update September 2012 [Internet]. 2014 [1 May 2014]. Available from: http://www.medpocket.com/

[14] He H, Garcia E. Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on. 2009; 21(9):1263-1284.

[15] Rhee S, Taylor J, Wadhera G, Ben-Hur A, Brutlag D, Shafer R. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proceedings of the National Academy of Sciences. 2006; 103(46):17355-17360.

[16] Pasomsub E, Sukasem C, Sungkanuparph S, Kijsirikul B, Chantratita W. The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems. Japanese journal of infectious diseases. 2010; 63(2):87-94.