Inteligência Computacional Aplicada na Predição e Identificação de Novas Mutações de Resistência do HIV-1, aos Inibidores Antirretrovirais Nelfinavir e Lopinavir/Ritonavir.

Robson M. Silva[1], Mariane R. Rita[1], Flávio F. Nobre[2]

[1]ICE/DEMAT– Universidade Federal Rural do Rio de Janeiro (UFRRJ)
BR 465 km 7 - CEP 23890-000 – Seropédica, RJ – Brasil

[2]Grupo de Engenharia Genômica, Programa de Engenharia Biomédica – COPPE/UFRJ – Rio de Janeiro, RJ - Brasil.

{robsonms,mariane_rita}@ufrrj.br,scarnaval@gmail.com,
flavio@peb.ufrj.br

Abstract. *We propose a computational model (GA/SVM) based on the use of genetics algorithm (GA) and support vector machine (SVM) classifier in order to identify possible new resistance mutations in the therapy targets genes of HIV-1.for the Nelfinavir(NFV) and Lopinavir/Ritonavir (LPV/r). The model was capable of selecting positions with I72 to the subtype B and specific mutation of subtype C with T74, simultaneously to NFV and LPV/r. The sensitivity obtained with the incorporation of positions frequently selected by GA in SVM classifier, was higher than 99.49% for subtype B and 96.70% for subtype C. The results were promising on the use of model for identification of new mutations from patients with clinical therapeutic failure to protease inhibitors NFV and LPV/r, and in predict the resistance of HIV-1.*

Resumo. *Propomos um modelo computacional (AG/SVM) baseado na utilização do algoritmo genético (AG) e no classificador de máquina de vetor de suporte (SVM), a fim de identificar novas mutações de resistência para o Nelfinavir (NFV) e Lopinavir/Ritonavir(LPV/r). O modelo foi capaz de selecionar as posições I72 para o subtipo B e a mutação T74 específica do subtipo C, simultaneamente para NFV e LPV/r. A sensibilidade obtida com a incorporação das posições mais frequentemente selecionadas pelo AG no classificador SVM, foi superior a 99,49% para o subtipo B e 96,70% para o subtipo C. Os resultados obtidos mostram-se promissores o uso do modelo na identificação de novas mutações em pacientes com falha terapêutica aos inibidores da protease NFV e LPV/r, bem como na predição da resistência do HIV-1.*

1. Introdução

De acordo com o Boletim Epidemiológico da UNAIDS [2011], estima-se que cerca de 33,4 milhões de pessoas no mundo estejam vivendo com o HIV-1/AIDS e que destas, cerca 31,3 milhões sejam adultos e 2,1 milhões sejam crianças com menos de 15 anos. O número estimado de novas infecções foi de 2,7 milhões, sendo que mais da metade correspondem a pessoas entre 15 e 24 anos. No Brasil, segundo o Ministério da Saúde [2011], estima-se que 630 mil pessoas estejam infectadas pelo HIV-1, o que corresponde a um terço da população infectada pelo vírus na América Latina, com taxa

de prevalência estimada de 0,5 (0,3 – 1,6)% na população adulta (15 a 49 anos). Deste total, apenas 235 mil têm conhecimento de sua sorologia e 180 mil encontram-se em tratamento.

Durante a última década [David, 2010], as terapias antirretrovirais (TARV) reduziram a mortalidade em pacientes portadores do HIV-1, porém não conseguem impedir totalmente o surgimento de novas formas virais resistentes, causadas principalmente pela elevada taxa mutacional do HIV-1. O desenvolvimento de resistência do HIV-1 aos antirretrovirais é um fator limitante para o sucesso da TARV. Pois além de não responderem adequadamente ao tratamento, os portadores de vírus resistentes podem transmitir esses vírus mutantes, representando um grave problema de saúde pública. O acúmulo de mutações de resistência e o surgimento de novas mutações promotoras de resistência aos medicamentos antirretrovirais representa um desafio importante na melhoria do tratamento de pacientes com vírus multirresistentes.

Nos últimos anos, diversas metodologias foram desenvolvidas visando avaliar fenotipicamente a resistência do HIV-1 aos medicamentos antirretrovirais, bem como avaliar genotipicamente o perfil mutacional do vírus. Dentre essas metodologias, pode ser citado o trabalho de Deforce *et al*. [2007] que propõe a aplicação de redes neurais bayesianas , de sorte a visualizar as relações entre o tratamento, mutações de resistência e a presença de polimorfismo para os inibidores de protease Indinavir (IDV), Saquinavir (SQV) e Nelfinavir (NFV). Os resultados obtidos permitiram identificar as posições de mutações 30N, 88S e 90M para o NFV, 90 M para o SQV e 82 A/T para IDV, como as principais mutações de resistência.

Silva, [2009] propõe um modelo computacional híbrido baseado na utilização de algoritmo genético (AG) e no classificador *Kernel* Discriminante de Fisher (KDF). Tal modelo utiliza a codificação dos resíduos de aminoácidos pela escala de hidrofobicidade e busca identificar possíveis novas mutações de resistência no gene da protease do genoma do HIV-1. Além disso, visa predizer a resistência em pacientes em falha terapêutica no Brasil, para os inibidores de protease: Saquinavir, Nelfinavir e Lopinavir.

Tais experimentos foram realizados de modo independente para cada droga e a acurácia obtida foi de respectivamente 88%, 81,25% e 84,93%. Em relação à identificação de possíveis novas mutações de resistência, o modelo se mostrou muito promissor, pois foi capaz de selecionar as principais posições de mutações de resistência para os inibidores em estudo.

Sing e Beerenwinkel [2006], propuseram um modelo de mistura com base na teoria de kernel de Fisher e árvore (MTreeMix Fisher Kernel) para a previsão de resistência do HIV-1 aos inibidores da transcriptase reversa (NNRTIs e NRTIs) e protease (IP). Os resultados obtidos apresentaram coeficientes de determinação superiores para todas as classes de inibidores, quando comparado com o método da regressão linear, indicando que as mutações selecionadas podem possibilitar previsões mais confiáveis no que tange a resistência a TARV. Entretanto muitas dessas metodologias têm limitações na interpretação de novas mutações.

O presente trabalho propõe um modelo computacional híbrido baseado em algoritmo genético (AG) e o classificador de máquina de vetor de suporte (SVM) para a predição de resistência em pacientes portadores do HIV-1, dos subtipos B e C, com falha terapêutica aos inibidores da protease Nelfinavir (NFV) e Lopinavir co-formulado com o inibidor Ritonavir (LPV/r).

Este artigo é estruturado da seguinte forma: seção 1 (Introdução), a justificativa e o objetivo deste trabalho são descritos nesta seção; seção 2 (Materiais e Métodos), esta seção descreve os principais métodos e materiais utilizados ao longo do desenvolvimento deste trabalho; a seção 3 (Resultados e Discussão) mostra os resultados obtidos com a aplicação do modelo computacional em comparação com o modelo de Stanford; seção 4 (Conclusão) que além de enfatizar as realizações que o trabalho conseguiu alcançar também cita uma proposta de inserir um novo parâmetro no modelo computacional.

2. Materiais e Métodos

O conjunto de dados constitui-se de 754 sequências do gene da protease provenientes de isolados séricos, além de informações clínicas referentes à contagem de linfócitos T-CD4$^+$ e carga viral e descrição dos regimes terapêuticos utilizados pelos pacientes portadores do HIV-1, resistentes à terapia antirretroviral. Os dados foram obtidos junto ao Laboratório de Virologia Molecular da Universidade Federal do Rio de Janeiro (UFRJ/Brasil), integrante da rede de laboratórios de genotipagem do Ministério da Saúde (RENAGENO). Um resumo das características do conjunto de dados pode ser vista na tabela 1.

Tabela 1. Distribuição das características do conjunto de dados.

	Subtipo	
	B	C
Frequência subtipo (%)	81,94	18,06
[nº de indivíduos]	617	137
Contagem de CD4+(células/mm^3) [média ± desvio padrão]	292,18 ± 6,58	413,55 ± 9,28
Carga viral (log10 cópias/ml) [média ± desvio padrão]	4,56 ± 0,67	4,71 ± 0,57

O método proposto consiste em um modelo computacional (AG/SVM), combinando Algoritmo Genético (AG) com Classificador de Máquina de Vetor de Suporte (SVM), na seleção e predição de mutações de resistência no gene da protease.

O Algoritmo Genético (AG) é uma técnica de otimização, que fornece um mecanismo de busca paralela e adaptativa baseado no princípio Darwiniano da seleção natural [Goldberg, 1989], utilizado na seleção de variáveis (posições de mutações da protease). Sua sistemática consiste primeiramente na geração aleatória de uma população de possíveis soluções (indivíduos) para o problema considerado, que evolui de acordo com operadores probabilísticos concebidos a partir de metáforas aos processos biológicos (reprodução, cruzamento genético e mutação), de sorte que há uma tendência de que os indivíduos mais aptos representem soluções cada vez melhores à medida que o processo evolutivo continua.

Máquina de Vetor de Suporte (SVM) é um método fundamentado na Teoria da Aprendizagem Estatística, baseada no princípio da minimização do risco estrutural (SRM), desenvolvida por Vapnik [1995], com o intuito de solucionar problemas de classificações de padrões. Consiste na seleção de um hiperplano que minimize o risco estrutural, a partir da resolução de um problema convexo quadrático. Este método de

aprendizado quando associada à função núcleo permite a construção de classificadores não-lineares, através do mapeamento dos dados iniciais em um espaço de dimensão superior ao original. Neste caso, o classificador linear em um espaço de dimensão superior corresponderá a um classificador não-linear no espaço original (Wasserman, 2004), garantindo ao classificador SVM uma boa generalização, ou seja, uma boa capacidade de predizer corretamente dados não relacionados na amostra de treinamento.

O fluxograma da metodologia utilizada esta representado na Figura 1, e suas etapas são descritas a seguir.

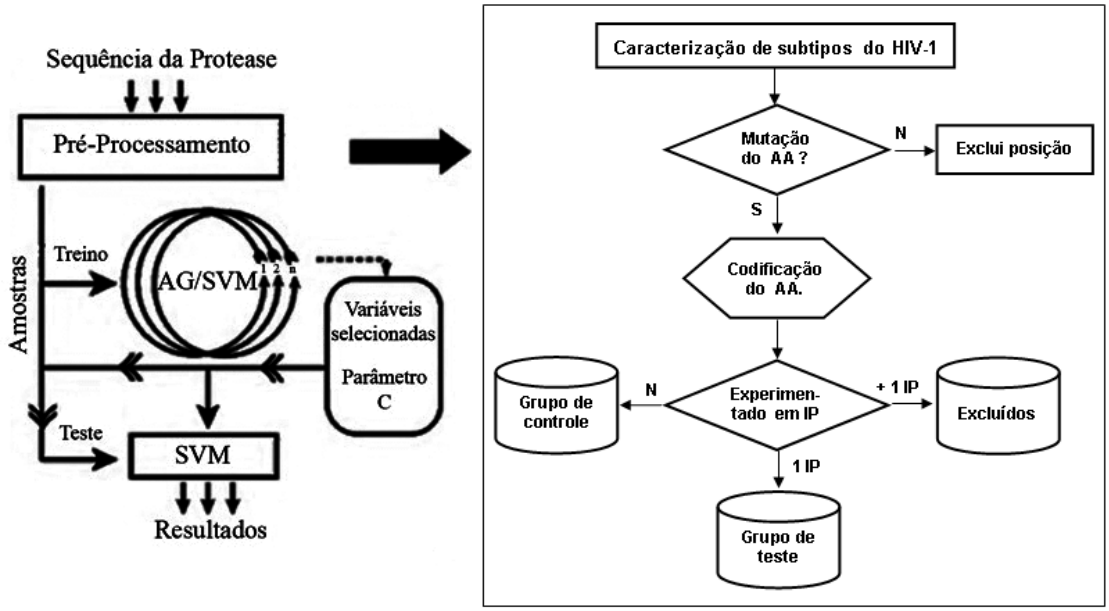

Figura 1. Fluxograma do modelo AG/SVM.

Inicialmente foram eliminadas as posições do gene da protease que não apresentaram mutações quando comparadas com a sequência de consenso HXB-2 para os vírus selvagens de subtipo B e subtipo C. Os resíduos de aminoácidos restantes foram codificados utilizando o fator de hidrofobicidade (hidrofóbicos ou hidrofílicos) de Kyte e Doolittle [1982], ponderados pelos respectivos pesos moleculares e normalizado no intervalo entre -1 e +1, através da equação:

$$HKD[i][j] = \frac{abs(hidro\ AA(i) * PM - hidro\ AA(j) * PM)}{1000}$$

Os dados foram divididos em dois grupos: o conjunto de treinamento, composto por 70% dos pacientes experimentados ou não aos inibidores NFV e LPV/r e o conjunto de teste independente, com os 30% pacientes restantes.

O processo evolutivo visa minimizar a soma ponderada do erro de treinamento do classificador SVM e a razão dos resíduos de aminoácidos:

$$F(x, c, d) = w_1 * Err + w_2 * \left(\frac{N_{res}}{NT_{res}}\right)$$

Onde w_1 é o peso do erro de treinamento do classificador SVM, w_2 o peso da razão entre o número de resíduos de aminoácidos selecionados pelo algoritmo genético (N_{res}) e o número total de resíduos mutável (NT_{res}).

O modelo computacional AG/SVM evolui enquanto não é satisfeito um dos seguintes critérios de parada, (1) Valor mínimo da função objetivo estabelecido, ou, (2) Número máximo de gerações. A relação dos parâmetros utilizados no modelo AG/SVM está resumida na Tabela 2.

Tabela 2. Parâmetros utilizados pelo modelo AG/SVM na seleção de mutações e predição de resistência para os inibidores NFV e LPV/r.

Parâmetros	Valor
Nº de simulações	30
População (n)	50
Gerações	100
Razão de *crossover*	0,8
Razão de mutação	0,09
Razão de inserção	0,8
Formato da variável	Binário [0, 1]
Método de seleção	Amostrador Estocático Universal
Função de *kernel*	Gaussiana
Parâmetro de regularização (C)	[0,1; 1,0; 10; 100]
Variância da função *kernel*	[0,5; 0,25; 0,125; 0,0625]

Ao final de cada uma das 30 simulações são selecionados os aminoácidos das posições de mutação mais frequentemente selecionadas ou seja, aquelas com maior frequência de seleção média superior ao primeiro quartil da distribuição de seleção das posições de mutação entre os 50 indivíduos na geração final do AG. Esses aminoácidos são então empregados no treinamento do classificador (SVM) e na predição de resistência da amostra independente de teste. Para avaliar o desempenho do classificador SVM na predição de resistência nas amostras no conjunto de treino e teste, foram utilizados os índices de sensibilidade (verdadeiro positivo), especificidade (verdadeiro negativo) e o erro de validação cruzada *leave one out*.

O modelo computacional AG/SVM proposto foi implementado utilizando o pacote GEATbx [Pohlheim, 2007], na implementação do algoritmo genético e o *software* MATLAB® (The Math Works, Inc; https*://www.mathworks.com*) na otimização do classificador SVM.

3. Resultados e Discussão

O acúmulo de resistência às drogas antirretrovirais e as consequentes falhas terapêuticas são um problema mundial para o sucesso da terapia da AIDS. A replicação residual sob pressão seletiva resulta no aparecimento de mutações no genoma do HIV-1 que diminui a suscetibilidade aos fármacos, reduzindo progressivamente a potência dos componentes do esquema terapêutico. A seleção de novas mutações de resistência na sequência do gene *pol* da polimerase do HIV-1, que promovem ou colaboram na resistência a TARV possibilita o aumento da sensibilidade de classificação de resistência, favorecendo assim o desenvolvimento de modelos mais eficientes na pesquisa de resistência antirretroviral.

Neste trabalho foi desenvolvido um modelo computacional de seleção de resíduos de aminoácidos (posições de mutações), na sequência da protease do HIV-1 por algoritmo genético (AG), associado ao classificador de máquina de vetor de suporte (SVM) na categorização de resistência. A codificação dos resíduos de aminoácidos pela escala de hidrofobicidade e ponderada pelo peso molecular, possibilitou ao modelo levar em consideração as propriedades físico-químicas dos aminoácidos, propriedades estas independentes ao subtipo do HIV-1.

No conjunto de 30 simulações realizadas com sequências do subtipo B experimentado no último regime terapêutico ao NFV (132 amostras), o desempenho médio na validação cruzada *leave-one-out* no conjunto de teste foi de 81,30%, sensibilidade de 99,50% e especificidade de 79,71% na caracterização de resistência, selecionando as mutações: L10, I15, L19, K20, D30, M36, R41, M46, I62, L63, I64, A71, I72, V77, I84, N88, L90 e I93. Para as sequências experimentadas no último regime terapêutico ao LPV/r (92 amostras), o desempenho médio verificado com a aplicação do modelo AG/SVM, no conjunto de simulações foi de 80,71% de validação cruzada *leave-one-out*, 99,90% de sensibilidade e 71,79% de especificidade. Tendo selecionado as mutações: L10, I15, G17, K20, E35, M36, S37, R41, M46, I50, I54, D60, I62, L63, K70, A71, I72, V77, V82, L90 e I93.

Para as sequências do subtipo C experimentado no último regime terapêutico ao NFV (34 amostras), o desempenho médio obtido no conjunto de 30 simulações foi de 80,59 % para validação cruzada *leave-one-out*, 96,75% para sensibilidade e 75,39% de especificidade na caracterização de resistência selecionando as mutações: L10, K14, K20, D30, M36, K45, M46, L63, A71, T74, V82, N88, L89 e L90. Para as sequências experimentadas ao inibidor LPV/r (17 amostras), o resultado obtido pelo modelo computacional foi de 76,61% de validação cruzada *leave-one-out*, 99,00% de sensibilidade e 72,31% de especificidade selecionando as mutações: L10, T12, G16, l19, K20, L33, M36, S37, M46, I54, Q61, I62, L63, A71, T74, V82, L89 e L90.

Nas tabelas 3 e 4 apresentamos a comparação entre os resultados obtidos pelo modelo computacional AG/SVM, em função da média (± desvio padrão) para a sensibilidade (S) e especificidade (E), com a utilização das mutações mais frequentes selecionadas pelo algoritmo genético (modelo) e as mutações descritas por Jonhson *et al*. [2011] para os inibidores NFV e LPV/r para os subtipos B e C, no conjunto de teste nas 30 simulações.

Tabela 3. Comparação entre os resultados obtidos (média ± desvio padrão) para o subtipo B, com a aplicação do classificador SVM, na caracterização de resistência, utilizando as mutações mais frequentemente selecionadas pelo AG e as mutações clássicas descritas na literatura como promotoras de resistência (Padrão – International AIDS Society/IAS 2011).

	Subtipo B			
	Modelo AG/SVM		Padrão IAS	
	NFV	LPV/r	NFV	LPV/r
Sensibilidade (S)	99,50± 1,03	99,90 ± 0,45	95,25 ± 8,40	94,37 ± 10,73
Especificidade (E)	79,23 ±14,83	73,12 ± 14,65	65,01 ± 17,94	54,33 ± 20,60

Tabela 4. Comparação entre os resultados obtidos (média ± desvio padrão) para o subtipo C, com a aplicação do classificador SVM, na caracterização de resistência, utilizando as mutações mais frequentemente selecionadas pelo AG e as mutações clássicas descritas na literatura como promotoras de resistência (Padrão – International AIDS Society/IAS 2011).

	Subtipo C			
	Modelo AG/SVM		Padrão IAS	
	NFV	LPV/r	NFV	LPV/r
Sensibilidade (S)	96,75± 7,61	99,00 ± 4,47	78,25 ± 11,81	80,00 ± 14,51
Especificidade (E)	75,38 ±13,64	72,31 ± 13,12	75,86 ± 12,44	74,46 ± 19,11

Os resultados obtidos em função das métricas sensibilidade (S) e especificidade (E) foram avaliados em termos da média e desvio padrão, visando verificar a existência de diferença estatística significante (p-valor < 0,05). Os resultados mostram a existência de diferença estatística na caracterização de resistência (sensibilidade) utilizando as mutações mais frequentemente selecionadas pelo AG e as mutações clássicas descritas na literatura para os inibidores NFV e LPV/r nos subtipos B e C.

4. Conclusão

A contribuição desse trabalho consistiu no desenvolvimento de um modelo computacional (AG/SVM) que combina a busca paralela dos algoritmos genéticos na seleção de mutações e o classificador SVM na predição de resistência em sequências da protease em pacientes portadores do HIV-1 de subtipos B e C em falha terapêutica no Brasil para aos inibidores NFV e LPV/r.

Os resultados obtidos pelo modelo AG/SVM mostraram-se promissores quanto à utilização de algoritmo genético na seleção de variáveis (mutações) associadas à resistência aos inibidores NFV e LPV/r, a partir de informações genotípicas do gene da protease do HIV-1, visto que o mesmo foi capaz de selecionar as principais mutações de resistência para os inibidores em estudo.

Além disso, o modelo foi capaz de selecionar mutações específicas do subtipo C, como as mutações nas posições G16 e T74, que foram selecionadas simultaneamente para o NFV e LPV/r. Cabe ressaltar que em a mutação na posição I74 já fora descrita como possível mutação de resistência para o inibidor NFV [Silva, 2009]. A mutação na posição I72 também foi selecionada pelo modelo em sequência do subtipo B para o inibidor LPV/r. No trabalho de King *et al.* [2007] encontramos a associação da referida mutação ao inibidor Lopinavir.

A incorporação no classificador SVM das posições de mutações mais frequentemente observadas possibilitou uma sensibilidade média no conjunto de teste superior aos valores médios obtidos pelos modelos que incorporaram somente as posições de mutações clássicas descritas na literatura. Esses achados indicam que no conjunto das mutações mais frequentemente selecionadas pelo modelo AG/SVM podem existir mutações indicativas de resistência aos inibidores NFV e LPV/r ainda não descritas na literatura. Apesar dos bons resultados obtidos com a aplicação do modelo na predição de resistência em sequências de subtipo C, há a necessidade de aumentar o número de amostras resistentes pertencentes a esse subtipo.

Como proposta para trabalhos futuros pretende-se aplicar novas técnicas de seleção de variáveis e a incorporação no classificador das variáveis referentes à contagem de CD4 e carga viral. Visando melhorar o modelo final em termo de predição, bem como, verificar a seleção de novas posições de mutações.

5. Referências

David, V. A.M.C. (2010) “HIV-1 drug-resistance patterns among patients on failing treatment in a large number of European countries” , Acta Dermatover, APA, vol. 19 p. 3-8.

Deforche, K. Silander, T., Camacho, R., et *al.* (2007) “Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance”. Bioinformatics, vol. 22, p. 2975-2979.

Goldberg, D. (1989) “Genetic Algorithms in Search, Optimization, and Machine Learning Reading”, MA, USA, Addison-Wesley.

Jonhson, A., Fraçoise, B., Bonaventura, C., *et al*. (2011) "Update of the Drug Resistance Mutations in HIV-1 :Spring 2011", Topics in HIV Medicine, vol. 16. p. 62-68.

Kyte, R., Doolittle, F. (1982). “A Simple Method for Displaying the Hidropathic Character of a Protein”, *Journal Molecular Biology*, p. 157-165.

King, M.S., R. Rode, I. Cohen-Codar, V. Calvez, A.G. Marcelin, G.J. Hanna, and D.J. Kempf. (2007). "Predictive genotypic algorithm for virologic response to lopinavir-ritonavir in protease inhibitor-experienced patients", Antimicrob Agents Chemother vol. 51: p. 3067-3074.

Ministério da Saúde DST/AIDS (2012) (https://www.aids.gov.br), acessado em 02/2012.

Pohlheim, H., (2007), "GEATbx : Genetic and Evolutionary Algorithm Toolbox for Use with Matlab", Disponível em http://www.geatbx.com/ . Acesso em 18 nov. 2007.

Silva R. M., (2009), "Algoritmo Genético e Kernel Discriminante de Fisher Aplicado a Identificação de Mutações De Resistência do Hiv-1 aos Inibidores Antirretrovirais da Protease". Tese de Doutorado, COPPE/Programa de Engenharia Biomédica - UFRJ.

Sing, T., e Beerenwinkel, N. (2007). "Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype", Advances in Neural Information Processing Systems 19, MA, MIT, USA, p. 1-9.

UNAIDS (2011). "AIDS epidemic update", p. 1-9. (https://www.unaids.org).

Vapnik V. N. (1998). "Statistical Learning Theory", John Wiley and Sons ltd., England.

Wasserman, L. (2004) "All of statistics: a concise course in statistical inference", Springer, New York.