Framework para classificação das mutações de vírus HIV

Mina Cintho

Dissertação apresentada
ao
Programa Interunidades em Bioinformática
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Programa Interunidades de Pós-Graduação em Bioinformática
Orientador: Prof. Dr. João Eduardo Ferreira
Coorientador: Prof. Dr. Roberto M.Cesar-Jr

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, 10 junho de 2014

Resumo

Um grande número de medicamentos utilizados no tratamento contra o HIV agem procurando inibir a ação das proteínas transcriptase reversa e protease. Mutações existentes nas sequências dessas proteínas podem estar relacionadas à resistência aos medicamentos e podem prejudicar o desempenho de um tratamento. O estudo do genótipo dos vírus pode ajudar na tomada de escolhas específicas em tratamentos para cada indivíduo, tornando maiores a chance de sucesso. Com a maior acessibilidade a exames de genotipagem, uma grande quantidade de sequências do vírus está disponível, contendo um grande volume de informação. Padrões de ocorrência de mutações são exemplos de informações contidas nessas sequências e são importantes por estarem relacionados à resistência aos medicamentos. Um dos caminhos que pode ser capaz de nos levar ao entendimento desses padrões de mutações é a aplicação de técnicas de agrupamento e *biclustering*. Essas técnicas visam a geração de grupos ou *biclusters* que possuam dados com propriedades em comum. São empregadas em casos em que não há grande quantidade de informação prévia e existem poucas hipóteses sobre os dados. Assim, pode-se encontrar os padrões de mutações que ocorrem nessas sequências e tentar relacioná-los com a resistência aos medicamentos, utilizando métodos de agrupamento e *bicluster* em sequências de protease e transcriptase reversa. Existem alguns sistemas que tentam predizer a resistência ou susceptibilidade das sequências, porém, devido à grande complexidade dessa relação, ainda é necessário esclarecer o vínculo entre combinações de mutações e níveis de resistência fenotípica. Desta forma, a principal contribuição deste trabalho é o desenvolvimento de um *framework* baseado na aplicação dos algoritmos $K-$Médias e Bimax às sequências de transcriptase reversa e protease de pacientes infectados com HIV, em uma codificação binária. O presente trabalho também introduz uma representação visual dos grupos e *biclusters* baseada em dados de microarranjos para casos em que se tem grandes volumes de dados, de forma a facilitar a visualização da informação extraída e a caracterização dos grupos e *biclusters* no domínio da doença.

Palavras-chave: HIV, K-médias, protease, transcriptase reversa.

Abstract

Drugs used in HIV treatment intend to inhibit protease and reverse transcriptase. Mutations in the sequences of these proteins can be related to drug resistance and can reduce treatment efficacy. Studying virus genotype may help choosing specific treatments for each patient, increasing success probability. As genotyping tests become available, a great amount of virus sequences, which comprehend lots of information, are more accessible. Patterns of mutation are examples of information comprised in the sequences and are important since are related to drug resistance. One way that can lead to the understanding of these mutation patterns is the use of clustering and biclustering techniques. These techniques search for clusters or biclusters comprising data with similar attributes. They are used when there is not a lot of previous information and there are few hypothesis about the data. Therefore, it may be possible to find patterns of mutations in the sequences and to relate them to drug resistance using clustering and biclustering techniques with protease and reverse transcriptase sequences. There are a few systems that predict drug resistance according to the sequence of the virus, however, due to the complexity of the relationship, it is still necessary to elucidate the connection between mutation combinations and the level of phenotypic resistance. Accordingly, this work main contribution is the development of a framework based on Kmeans and Bimax algorithms with protease and reverse transcriptase sequences from HIV patients in a binary form. This work also presents a visual representation of the clusters and biclusters based on microarray data suitable for large data volumes, helping the visualization of information extracted from data and cluster and bicluster characterization in the disease domain.

Sumário

Lista de Abreviaturas

ABC	abacavir
ATV	atazanavir
AZT	zidovudine
d4T	stavudine
ddI	didanosine
DNA	ácido desoxirribonucleico
DRV	darunavir
EFV	efavirenz
ETV	etravine
FPV	fosamprenavir
grupo B6.1	grupo número 1 do conjunto de sequências do subtipo B, com k igual a 6
HAART	*highly antiretroviral therapy*
HIV	vírus da imunodeficiência humana
IDV	indinavir
LPV	lopinavir
NNRTI	inibidores de transcriptase reversa não análogos de nucleosídeos
NRTI	análogos de nucleosídeo e nucleotídeo
NVP	nevirapine
PI	inibidores de protease
PR10	posição de aminoácido número 10 da proteína protease
RNA	ácido ribonucleico
RT215	posição de aminoácido número 215 da proteína transcriptase reversa
SISGENO	sistema de controle de exames de genotipagem
SQV	saquinavir
TAM	mutações análogas a timidina
TPV	tipranavir

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

No tratamento de pacientes infectados com o vírus da imunodeficiência humana (HIV), a maioria dos medicamentos ministrados aos pacientes visa a inibição da ação das proteínas transcriptase reversa, protease e integrase (Shafer *et al.*, 2000b), uma vez que são de grande importância na replicação do vírus. Assim, as sequências destas proteínas são alvo de estudo por conterem importantes informações capazes de influenciar o resultado do tratamento de pacientes.

Esses estudos sobre as sequências de proteínas do HIV indicam, por exemplo, que o vírus possui uma extensa variabilidade genética resultante da ausência de mecanismos de verificação [1] e das altas taxas de mutações e replicação do vírus (Mansky, 1998). Essa variabilidade genética é um importante fator no mecanismo de resistência aos medicamentos. Ao longo de seu tratamento, caso um paciente esteja infectado com vírus que contenha mutações que confiram vantagens seletivas, pode ocorrer seleção deste vírus. A seleção ocorre com a replicação do vírus que passa a se tornar dominante mediante a população viral no plasma sanguíneo, devido à vantagem adquirida com a ocorrência da mutação, ocasionando falha terapêutica.

Dessa forma, a ocorrência de mutações, principalmente na protease e na transcriptase reversa, levando à resistência aos medicamentos, torna a relação entre mutação e resistência importante de ser elucidada, bem como a relação entre o genótipo do vírus e a probabilidade de sucesso ou falha de um tratamento. Entendendo as mutações presentes nas sequências de cada paciente, a interação entre essas mutações, os padrões de ocorrência dessas mutações e a probabilidade de resistência aos medicamentos, será possível escolher medicamentos mais eficazes no tratamento individual de cada paciente específico.

A relação entre as sequências de aminoácido das proteínas dos vírus, sua variabilidade e a predição de resistência aos medicamentos é bastante complexa e ainda não foi completamente elucidada. A complexidade da interação é dada por fatores como a resistência cruzada, limitação na detecção de resistência do vírus a um medicamento, grande quantidade de mutações possíveis e os diferentes resultados que podem surgir da combinação e interação entre as mutações (Shafer *et al.*, 2000b).

A resistência cruzada acontece quando um medicamento ministrado a um paciente causa resistência a outro medicamento não ministrada. Assim, uma mutação que causa resistência a um medicamento pode causar concomitante resistência a outro medicamento, sendo mais comum entre medicamentos de mesma classe.

Outro fator importante para complexidade da relação é a limitação na detecção de resistência do vírus a um medicamento. A limitação surge em casos em que a população de uma variante resistente não é muito numerosa e não é possível detectar sua existência. Assim, a resistência não é detectada e ocorre falha terapêutica.

Também é fator de influência na complexidade da relação entre o genótipo do vírus e a susceptibilidade aos medicamentos a grande quantidade de mutações possíveis e os diferentes resultados que podem surgir da combinação e interação entre as mutações. Existe um extenso número de possíveis mutações que geram resistência e as interações e diferentes combinações entre essas mutações modificam a eficiência do vírus na sua proliferação. Uma única mutação pode ser responsável pela resistência a um medicamento, chamada então de mutação primária, que diretamente causa a diminuição da susceptibilidade do vírus ao tratamento. No entanto, essa mutação pode necessitar de uma mutação dita acessória, para melhorar a aptidão (*fitness*) do vírus, ou ainda, uma mutação pode depender da ausência de outras mutações para causar resistência.

Apesar desses obstáculos, trabalhos anteriores relacionam a ocorrência de mutações com a resistência aos medicamentos (Johnson *et al.*, 2010). A maioria desses trabalhos inserem mutações em um clone viral

[1] Em inglês: proofread

susceptível e comparam fenótipos ou observam o surgimento de mutações em laboratório ou em pacientes em tratamento conhecido. Assim, derivam relações diretas entre a ocorrência das mutações e a susceptibilidade aos medicamentos.

A partir desses estudos, sistemas baseados em regras (Lathrop *et al.*, 1998; Shafer *et al.*, 2000a), *look up tables* (algoritmoBrasileiro; Schinazi *et al.*, 2000) e bancos de dados (losAlamos) compilam o conhecimento produzido e aplicam ao tratamento de pacientes. Os sistemas baseados em regras, por exemplo, calculam valores para a probabilidade de susceptibilidade ou sucesso do tratamento para uma dada sequência por meio da geração de um sistema de penalidades para a ocorrência de mutações. *Look up tables* e bancos de dados relacionam a ocorrência de mutações com a resistência aos medicamentos. No entanto, essas aplicações dependem do conhecimento gerado pela literatura científica, sua qualidade, confiabilidade e aplicabilidade para realizar suas predições e no caso de novas mutações é necessário esperar por estudos que as relacionem com a aquisição de resistência e no caso do desenvolvimento de novos medicamentos, estudos que apontem as mutações importantes para o surgimento de resistência.

Assim como os sistemas baseados em regras e as *look up tables* realizam uma classificação das sequências de HIV, criando rótulos de predição de resistência aos medicamentos, algoritmos de reconhecimento de padrões extraem informações dos dados com o intuito de classificá-los. Algoritmos de reconhecimento de padrões como algoritmos de agrupamento, ou *clustering*, visam o agrupamento de dados de forma que cada grupo contenha elementos, chamados padrões, similares entre si (Kriegel *et al.*, 2009). Os elementos de um mesmo grupo devem ser mais semelhantes entre si do que quando comparados com elementos de um outro grupo. Um exemplo de algoritmo de agrupamento é o $K-$Médias (Hartigan e Wong, 1979; Lloyd, 1982; MacQueen, 1967), popular em aplicações em que se tem grandes volumes de dados. Por conseguinte, algoritmos de agrupamento podem representar uma boa abordagem a ser seguida na extração de informações de dados de sequências. Aplicando-se esses algoritmos em sequências de protease e transcriptase reversa pode-se obter uma classificação, formando grupos com sequências similares em seus padrões de mutações.

Outra abordagem mais recente de reconhecimento de padrões é o *biclustering* (Mechelen *et al.*, 2004a), que tem como objetivo encontrar submatrizes seguindo certos padrões em meio a matriz de dados. Esses padrões podem ser dados por: subconjuntos de linhas que sigam um critério de homogeneidade em um subconjunto de colunas, ou subconjuntos de colunas que sigam um critério de homogeneidade em um subconjunto de linhas (Kriegel *et al.*, 2009). Com essa abordagem pode ser possível encontrar posições de aminoácidos nas sequências de transcriptase reversa e protease, representadas na forma de subconjuntos de colunas, contendo mutações que ocorrem simultaneamente em sequências de vírus de pacientes, representadas na forma de subconjuntos de linhas.

Esses algoritmos de reconhecimento de padrões são frequentemente aplicados em dados genômicos. Métodos de *clustering* são aplicados, por exemplo, em dados de expressão gênica (Quackenbush, 2001; Slonim, 2002), análise de dados proteômicos e metabolômicos (Goodacre *et al.*, 1998), comparação e predição de estrutura de proteínas (Kaplan *et al.*, 2004; Krasnogor e Pelta., 2004) e geração de redes regulatórias de genes (Tavazoie *et al.*, 1999). Já métodos de biclustering são empregados na análise de dados de expressão gênica (Ideker *et al.*, 2001; Kluger *et al.*, 2003; Madeira e Oliveira, 2004), corregulação de genes (Ben-Dor *et al.*, 2003; Liu e Wang, 2003; Yang *et al.*, 2003), anotação automática de genes (Segal *et al.*, 2001, 2003; Tanay *et al.*, 2002) e classificação de amostra e tecido (Kluger *et al.*, 2003; Murali e Kasif, 2003; Sheng *et al.*, 2003). Assim, o reconhecimento de padrões tem sido aplicado a dados de sequências e deve ser adequado ao nosso propósito.

Desta forma, a utilização de algoritmos de reconhecimento de padrões em dados genômicos e proteômicos e a ampla disponibilidade de dados de testes genotípicos de HIV levantam duas importantes perguntas: 1) é possível classificar as sequências, baseando-se somente na ocorrência de mutações nas diferentes posições de aminoácido das proteínas? e 2)é possível alcançar uma classificação capaz de expressar o conhecimento atual sobre a relação entre mutações e resistência aos medicamentos? Nesse trabalho, temos como principal contribuição a aplicação de algoritmos de agrupamento e *biclustering* em sequências de protease e transcriptase reversa, uma vez que esses algoritmos podem ser capazes de responder essas perguntas, extraindo informação das sequências e gerando grupos e *biclusters* que podem estar relacionados à predição de resistência aos medicamentos. Os grupos e *biclusters* poderiam, então, ser usados na predição de resposta de tratamento de pacientes infectados com HIV, aumentando a probabilidade de sucesso do tratamento.

Os resultados desse trabalho foram apresentados no *8th IEEE International Conference on eScience 2012* em Chicago, no II Workshop e III Workshop do Programa Interunidades em Bioinformática. No II Workshop do Programa Interunidades em Bioinformática, o trabalho recebeu o terceiro lugar no Prêmio de Melhor Trabalho Apresentado e no III Workshop do Programa Interunidades em Bioinformática o primeiro lugar no Prêmio de Melhor Trabalho Apresentado.

1.2 Objetivo

O principal objetivo deste trabalho é o desenvolvimento de um *framework* capaz de gerar grupos e *biclusters* que representem os padrões de mutações mais frequentes em sequências de protease e transcriptase reversa. Esse *framework* é baseado na aplicação dos algoritmos $K-$Médias e Bimax nas sequências de aminoácido de proteínas, com intuito de explorar e melhor compreender os padrões de ocorrência de mutações, a interação entre mutações e a influência de uma mutação na ocorrência de outra mutação.

O *framework* descrito neste trabalho também introduz um esquema de visualização de grupos e *biclusters*, baseado em dados de microarranjo, adequado para grandes volumes de dados. Esse esquema de visualização auxilia na sumarização das informações contidas nos grupos e *biclusters*, bem como na sua caracterização no domínio do HIV.

Essas informações procedentes do agrupamento e do *biclustering* podem contribuir para a elucidação: dos mecanismos de interação entre mutações, dos padrões de ocorrência de mutações e da influência dessas mutações na alteração da probabilidade de sucesso no tratamento a pacientes. Essas informações podem ser aplicadas no estabelecimento de tratamentos específicos para cada paciente, de acordo com o genótipo do vírus de cada paciente.

1.3 Organização do Trabalho

No Capítulo 2 são expostos os fundamentos básicos para o estudo do problema: HIV, Resistência, Agrupamento e Biclustering. Também são retratados trabalhos relacionados que aplicam análises de chi-quadrado, teste de Fisher, teste de Benjamini-Hochberg, informação mútua, teste de permutação, análise de componentes principais, cadeia de Markov, plot poissoness e agrupamento para classificar sequências de HIV. No Capítulo 3 é descrito o framework proposto e sua implementação, no Capítulo 4 os resultados e a discussão. Finamente, no Capítulo 5 é apresentada a conclusão do trabalho.

Capítulo 2

Fundamentos e Trabalhos Relacionados

2.1 Fundamentos

Iniciamos com a Seção 2.1.1 apresentando os fundamentos biológicos: HIV e resistência aos medicamentos. Em seguida, na Seção 2.1.2 são descritos os fundamentos computacionais: reconhecimento de padrões, agrupamento, K-médias, *biclustering* e Bimax.

2.1.1 Fundamentos Biológicos

2.1.1.1 HIV e Resistência

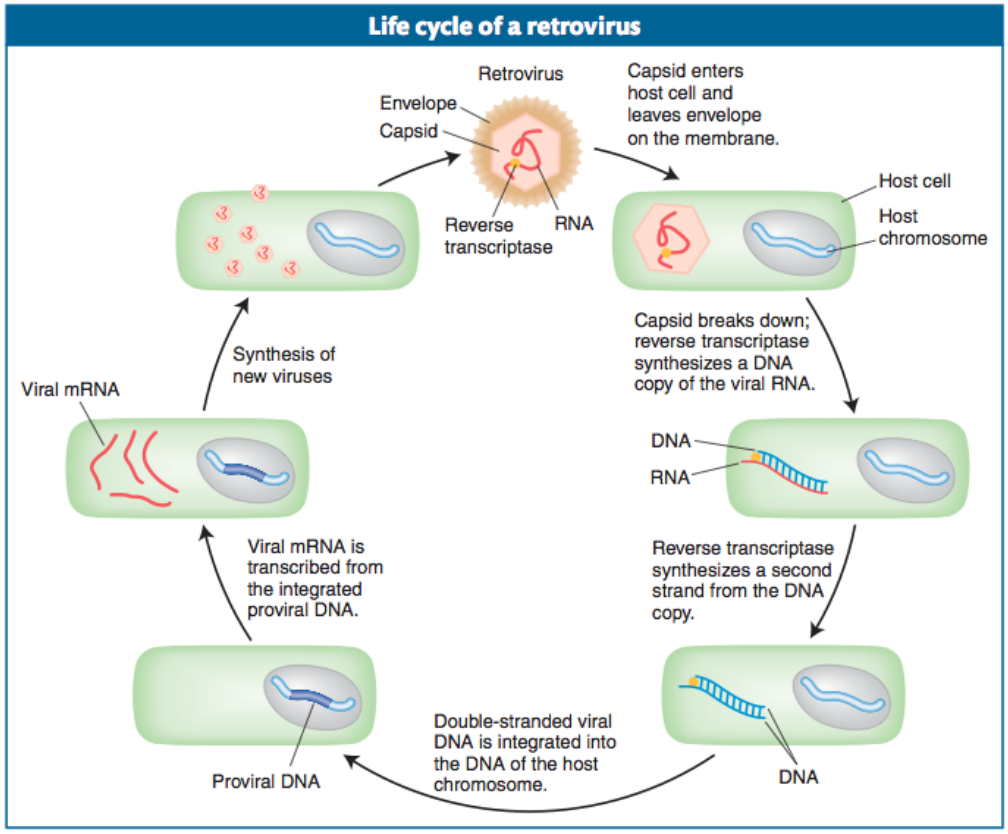

Figura 2.1: *Ciclo de vida dos retrovírus (extraída de (Suzuki et al., 2010))*

O vírus da imunodeficiência humana tem um ciclo de replicação típico de retrovírus, iniciando na infecção, seguido da ação das proteínas transcriptase reversa e integrase, expressão gênica viral, montagem do vírus e maturação com a ação da protease. Na infecção, o vírus fusiona sua célula com a célula do hospedeiro e o RNA, transcriptase reversa, integrase e outras proteínas virais entram na célula (Suzuki *et al.*, 2010). Então a transcriptase reversa produz DNA a partir do RNA do vírus através da transcrição reversa, o

DNA é transportado para dentro do núcleo e a integrase integra o DNA do vírus em locais aleatórios dos cromossomos do hospedeiro. Após a integração, o DNA está sujeito a regulação transcricional da célula bem como a seus próprios mecanismos de controle transcricional. Os genes são então transcritos e o novo RNA viral é utilizado para produção de proteínas virais, que se movem para a superfície da célula e são geradas novas formas de HIV imaturas. A protease entra em atividade e libera proteínas individuais do HIV e o vírus adquire a capacidade de realizar a infecção.

Entre todas essas etapas do ciclo de replicação do HIV, as mais utilizadas como alvo para terapia são as etapas de ação da trancriptase reversa e da protease (Shafer *et al.*, 2000b), já que ambas proteínas são importantes para o desenvolvimento da doença. A transcriptase reversa atua na produção de DNA a partir do RNA, além da formação de dupla fita do DNA e da remoção de RNA que não está sendo utilizado através de uma atividade intrínseca de RNaseH. Já a protease é responsável pela clivagem da poliproteína produzida pelos genes *gag-pol*. A inibição dessa proteína interfere na via de montagem do vírus, resultando em partículas não infecciosas.

A inibição da ação da transcrição reversa no tratamento de pacientes é dada pelo mecanismo de término da cadeia de DNA em produção (análogos de nucleosídeo e nucleotídeo, NRTI), como por exemplo aos medicamentos abacavir (ABC), zidovudine (AZT), stavudine (d4T) e didanosine (ddI) ou pela inibição da ação da proteína (inibidores de transcriptase reversa não análogos de nucleosídeos, NNRTI), como agem por exemplo efavirenz (EFV), etravine (ETV) e nevirapine (NVP) (Shafer *et al.*, 2000b). Já os medicamentos que têm como alvo a protease (inibidores de protease, PI) são compostos baseados em substrato e agem como inibidores competidores das reações proteolíticas (Boden e Markowitz, 1998; Huff e Kahn, 2001), como por exemplo aos medicamentos atazanavir (ATV), darunavir (DRV), fosamprenavir (FPV), indinavir (IDV), lopinavir (LPV), saquinavir (SQV) e tipranavir (TPV) (Shafer *et al.*, 2000b).

Como forma de otimizar a eficácia do tratamento, os medicamentos são ministrados em conjunto, combinando medicamentos com distintos mecanismos de ação, ou classes de medicamentos, na chamada *Highly Antiretroviral Therapy (HAART)*. A HAART reduz a probabilidade do vírus ser resistente a todos medicamentos e tem o propósito de inibir mais de um estágio do ciclo de vida do vírus de forma a prorrogar a obtenção de resistência do vírus ao tratamento (Deeks, 2003).

Mesmo com a aplicação da HAART, vírus contendo mutações que conferem vantagem seletiva em meio a terapia podem se proliferar, se tornando dominantes mediante a população viral, causando falha terapêutica. Assim, a existência de variantes resistentes aos medicamentos limita a efetividade do tratamento a longo prazo. No entanto, com o maior esclarecimento sobre a relação entre ocorrência de mutações na protease e na transcriptase reversa, a frequência de ocorrência das mutações, a interação entre as mutações e a resistência aos medicamentos seria possível desenvolver tratamentos personalizados e mais eficientes.

Consequentemente, é necessário que sejam identificadas as mutações que conferem resistência aos medicamentos para que se faça a melhor escolha de tratamento possível. Com esse intuito, por meio do sequenciamento e alinhamento de uma extensa quantidade de sequências foi identificado um grande número de mutações nos genes de protease e transcriptase reversa. Dois fatores que colaboram para o alto índice de ocorrência de mutações são a alta taxa de recombinação que ocorre nos casos em que uma célula é infectada com mais de uma variante viral (Hu e Temin, 1990; Levy *et al.*, 2004) e a reativação de variantes de vírus latentes em cromossomos de células infectadas. Além disso, são importantes fatores a elevada taxa de replicação do vírus e ausência de mecanismos de revisão (Mansky, 1998) na transcrição. Com a alta taxa de replicação e ausência de mecanismos de revisão, erros ocorrem e se acumulam, já que não são corrigidos, gerando variabilidade genética. A transcriptase reversa é uma das maiores responsáveis pela taxa de mutação ou variabilidade genética do HIV (Preston *et al.*, 1988). A alta taxa de erros na transcriptase reversa, 1 em 10.000 bases, e grande velocidade de replicação do vírus, 108-109 virions por dia, favorecem o acontecimento de mutações e a seleção de vírus resistentes.

Como essa extensa variabilidade genética do HIV está relacionada à resistência aos medicamentos, estudos também têm sido realizados no sentido de verificar possíveis relações existentes entre mutações nessas sequências e a capacidade de transmissão, patogenicidade e resposta a tratamentos (Baeten *et al.*, 2007; Johnson *et al.*, 2010; Kanki *et al.*, 1999; Laeyendecker *et al.*, 2006; Shafer *et al.*, 2000b). Esses estudos incluem a elucidação de correlações genótipo-fenótipo em isolados de HIV em laboratório, correlações genótipo-fenótipo em isolados clínicos de HIV, genótipo-histórico de tratamento e correlações genótipo-resultado clínico (Shafer *et al.*, 2000b). Estudos de correlação genótipo-fenótipo em isolados de HIV em laboratório permitem a identificação de mutações de resistência "canônicas"aos medicamentos, diferentemente de estudos correlação genótipo-fenótipo em isolados de HIV clínicos que permitem a identificação do efeito fenotípico de mutações *in vitro* no padrão em que ocorrem *in vivo*. Já correlações genótipo-tratamento podem mostrar mutações ocorrendo na presença de pressão para escape de medicamentos anti-retrovirais e correlações genótipo-resultado clínico mostram a influência das mutações na resposta virológica a um tratamento.

Embora dados sobre correlações genótipo-histórico de tratamento sejam os mais informativos para o

entendimento da relação entre genótipo e tratamento, esses dados são mais raros. Como exames de genotipagem são mais baratos, mais rápidos e superiores na detecção do processo de desenvolvimento de resistência que exames de fenotipagem(AntiretroviralGuidelines, 2012), grandes volumes de dados de sequenciamento estão se tornando disponíveis. Esses dados contêm grande quantidade de informação sobre o vírus, ainda a serem explorados em progresso do conhecimento atual.

A subdivisão do vírus em diversas categorias é um exemplo de informação contida em suas sequências. As diferentes categorias são os grupos, subtipos, sub-subtipos e formas recombinantes circulantes. Os chamados grupos representam as diferentes linhagens que podem ser M, N ou O, os subtipos representam os maiores clados (grupo monofilético, ou ancestral comum com todos seus descendentes (Dupuis, 1984)) do grupo M, os sub-subtipos representam linhagens bastante próximas de uma linhagem de subtipo, mas não distante suficiente para compor um subtipo diferente e as formas recombinates circulantes representam um linhagem recombinante (Robertson *et al.*, 2000). O grupo M (*Main* ou principal) é o grupo de maior predominância em circulação (Taylor *et al.*, 2008) e seu subtipo mais estudado é o subtipo B.

Apesar dessa subdivisão, o efeito da ocorrência de mutação na resistência aos medicamentos é igual para todas categorias (Tang e Shafer, 2012), ou seja, uma mutação que cause resistência a um medicamento em um subtipo, também causará resistência em outro subtipo. Adicionalmente, em experimentos nos Estados Unidos e Europa, os medicamentos se mostraram tão efetivas em subtipos B, como em outros subtipos (Bannister *et al.*, 2006; Geretti *et al.*, 2009; Scherrer *et al.*, 2011). No entanto, os subtipos diferem quanto à probabilidade de adquirir uma dada mutação (Tang e Shafer, 2012). A diferença nas probabilidades de ocorrência de mutações são dadas por diferenças sutis de aminoácidos nas sequências que geram diferenças nas estruturas, no contexto ao redor de um nucleotídeo e no uso de códons. Logo, se o estudo das mutações se concentrar nas populações de subtipo B, outras mutações importantes em outros subtipos podem não ser captadas.

Além das variações nas probabilidades de ocorrência de mutações dos subtipos, o esclarecimento da relação entre resposta a tratamentos e resistência aos medicamentos com o fenótipo e genótipo do vírus são dificultadas pela complexidade desta interação. A complexidade é proveniente da combinação de 3 fatores: a resistência cruzada, as limitações de se detectar resistências relevantes e os efeitos das mutações individualmente e em combinação.

O primeiro acontece quando há resistência a um medicamento a qual o vírus ainda não foi exposto e ocorre principalmente entre medicamentos de mesma classe, limitando as opções de tratamento. O segundo é dado pelos exames de resistência genotípicos padrões que não conseguem detectar variantes presentes em níveis menores que 20% na população de vírus no plasma e pelos exames fenotípicos que não conseguem detectar linhagens mutantes em meio a linhagens selvagens dependendo da proporção das duas populações (D'Aquila, 2000; Hirsch *et al.*, 2000; Laethem *et al.*, 1999; Tang e Shafer, 2012).

No terceiro, a resistência a medicamentos ministrados no tratamento contra o HIV pode ser resultante da ocorrência de uma única mutação, acúmulo de mutações ou a combinação de ausência e presença de mutações. Por exemplo, uma mutação selecionada por conferir resistência a um medicamento pode ser estruturalmente importante, por exemplo, se estiver situada próximo do sítio ativo da proteína ou interagir com o substrato. No entanto, por ser estruturalmente importante, pode acabar ocasionando perda de *fitness*, por exemplo, perda de estabilidade da estrutura ou de atividade da proteína. A presença de uma outra mutação, que sozinha teria pouco ou nenhum efeito na susceptibilidade a medicamentos, pode ser capaz de compensar essa perda de *fitness* e a combinação das duas mutações ser positivamente selecionada. Portanto, a interação entre as mutações representa um importante fator a ser considerado.

No caso da protease, existem ainda os casos em que mutações localizadas a maiores distâncias do sítio ativo e sem interação direta com o substrato reduzem a susceptibilidade (Rhee *et al.*, 2003). O mecanismo de ação dessas mutações não é completamente compreendido (Muzammil *et al.*, 2003) e tornam a elucidação da relação entre as mutações e a probabilidade de resistência aos medicamentos em sequências de protease ainda mais complexa.

Apesar das dificuldades no esclarecimento da relação entre mutações e resistência aos medicamentos, o conhecimento gerado por estudos é aplicado ao tratamento de pacientes por meio de *look-up tables*, bancos de dados, e sistemas baseados em regras. *Look-up tables* e bancos de dados, como a *look-up table* brasileira (algoritmoBrasileiro) e o *The Los Alamos resistance database* (losAlamos), partem da associação de mutações e resistência aos medicamentos, e criam regras para a correspondência entre a substituição de um aminoácido e a resistência a um medicamento (Beerenwinkel *et al.*, 2001). Já *Customized Treatment Strategies for HIV (CTSHIV)* (Lathrop *et al.*, 1998) e o *HIV mutation search engine for queries (HIV-SEQ)* (Shafer *et al.*, 2000b) são exemplos de sistemas baseados em regras que calculam valores para probabilidade de falha ou sucesso terapêutico. Essas *look-up tables*, bancos de dados e sistemas baseados em regras são gerados a partir da compilação de trabalhos científicos e do conhecimento atual das mutações e resistência aos medicamentos.

Existem cerca de 10 sistemas de interpretação de sequências de HIV, sendo aproximadamente metade

proprietária e metade pública (Rhee *et al.*, 2009). Contudo, *look-up tables*, bancos de dados e sistemas baseados em regras são baseados em trabalhos científicos, sua qualidade, confiabilidade e aplicabilidade. Adicionalmente, na ocorrência de novas mutações ou no desenvolvimento de novos medicamentos é necessário esperar pela realização desses estudos para que sejam incluídas as novas mutações e novos medicamentos no cálculo da predição de resistência aos medicamentos. Portanto, seria interessante alcançar uma classificação relacionada à predição de susceptibilidade aos medicamentos considerando-se apenas as sequências do vírus. Uma vez superado esse obstáculo, seria possível auxiliar a tomada de decisão sobre o tratamento a ser ministrado a cada paciente individualmente a partir do seu resultado de exame de genotipagem.

2.1.2 Fundamentos Computacionais

2.1.2.1 Reconhecimento de Padrões

Tecnologias como microarranjo e sequenciamento de nova geração permitiram a geração de enormes volumes de dados de sequência e expressão gênica. Essas tecnologias têm levado a busca por métodos sofisticados de processamento e análise de dados que sejam capazes de lidar com enormes conjuntos de dados eficientemente.

Métodos de análise de dados como algoritmos de agrupamento e *biclustering* têm sido bastante utilizados na análise de grandes volumes de dados. Algoritmos de clustering têm sido aplicados a dados de expressão gênica (Quackenbush, 2001; Slonim, 2002), análise de dados proteômicos e metabolômicos (Goodacre *et al.*, 1998), comparação de proteínas e predição de estrutura (Kaplan *et al.*, 2004; Krasnogor e Pelta., 2004) e na geração de redes regulatórias de genes (Bilu e Linial, 2002; Tavazoie *et al.*, 1999). Métodos de biclustering têm sido aplicados na análise de dados de expressão gênica (Ideker *et al.*, 2001; Kluger *et al.*, 2003; Madeira e Oliveira, 2004), corregulação de genes (Ben-Dor *et al.*, 2003; Liu e Wang, 2003; Yang *et al.*, 2003), anotação automática de genes (Segal *et al.*, 2001, 2003; Tanay *et al.*, 2002) e classificação de amostra e tecido (Kluger *et al.*, 2003; Murali e Kasif, 2003; Sheng *et al.*, 2003).

Assim como *look up tables* e sistemas baseados em regras classificam as sequências de protease e transcriptase reversa de acordo com a predição de resposta a tratamentos contra o HIV, técnicas de reconhecimento de padrões extraem informações contidas em conjuntos de dados para classificá-los. Portanto, a aplicação de métodos de reconhecimento de padrão pode ajudar a responder perguntas como: *é possível classificar as sequências, baseado na ocorrência de mutações nas diferentes posições de aminoácido das proteínas?* e *é possível alcançar uma classificação capaz de expressar o conhecimento atual sobre a relação entre mutações e resistência aos medicamentos?*

2.1.2.2 Agrupamento

Nos casos em que pouco se conhece sobre os dados e sobre sua distribuição e é necessário fazer o menor o número de suposições possível sobre os mesmos, o método de agrupamento ou *clustering* é apropriado para estudar o relacionamento entre os dados e sua estrutura (Jain *et al.*, 1999). Algoritmos de agrupamento organizam os dados, chamados de padrões, geralmente representando-os como vetores de medidas ou pontos no espaço multi-dimensional. Os vetores ou pontos são representados por atributos que caracterizam o dado e o agrupamento é realizado baseando-se em uma medida de similaridade desses atributos (Jain *et al.*, 1999). Desta forma, um dado pode ser representado por um vetor x :

$$x = (x_1, x_2, ..., x_n)$$

n dimensional, sendo n determinado pela quantidade de atributos que caracterizam o padrão x e, a partir dessa representação, pode-se tentar estabelecer similaridades entre os padrões.

Os grupos ou *clusters* resultantes do processo devem possuir padrões com propriedades em comum, ou seja, similares. Dessa forma, padrões pertencentes a um mesmo grupo devem ser mais similares entre si do que quando comparados com padrões de outro grupo. A construção dos grupos pode ajudar na sugestão de hipóteses sobre o relacionamento entre os dados e sobre a estrutura desses dados, ajudando na análise e extração de informações.

O agrupamento dos dados nesse método é baseado em medidas de distância ou similaridades. A definição de medidas de similaridades abrange uma ampla variedade de possibilidades e geralmente envolve subjetividade e escolhas como a natureza (discreta, contínua ou binária), escala (nominal, ordinal ou intervalar) e outras características (Johnson e Wichern, 1982). Essas escolhas influenciam na disposição dos dados e, consequentemente, podem influenciar nos formatos dos grupos.

A medida de similaridade pode ser calculada pela representação na forma de coeficientes de correlação, medidas de associação, como, por exemplo, frequências ou na forma de medidas de distâncias. A distância Euclidiana é uma das medidas mais utilizadas (Jain *et al.*, 1999).

2.1.2.3 Agrupamento Não-Hierárquico: $K-$Médias

Algoritmos de agrupamento podem ser hierárquicos e não-hierárquicos. Agrupamentos não-hierárquicos obtêm uma partição única dos dados, diferentemente de agrupamento hierárquicos nos quais ocorre uma partição aninhada dos padrões (Jain *et al.*, 1999). Agrupamentos hierárquicos são representados graficamente por dendrogramas. No entanto, dendrogramas não são representações adequadas para grandes volumes de dados.

Agrupamentos não-hierárquicos são mais adequados para casos nos quais é necessário analisar grandes volumes de dados. Isso porque os algoritmos de agrupamento não-hierárquicos não necessitam da criação, manutenção e armazenamento de uma matriz de distâncias entre os padrões, o que é custoso em termos de tempo e espaço. Além da matriz de distâncias, a grande quantidade de partições possíveis dos padrões é fator importante na eficiência dos algoritmos de agrupamento. A partição dos dados em dois grupos, por exemplo, pode ser dada de $\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i}$ formas distintas.

Assim, alguns algoritmos de agrupamento não-hierárquico produzem grupos pela otimização de uma função critério definida localmente, em um subconjunto de padrões (Jain *et al.*, 1999). Geralmente iniciam de uma partição inicial ou de um conjunto inicial de *seed points* que irão formar os núcleos dos grupos. A partição inicial ou os *seed points* vão sendo modificados de acordo com a função critério, até que esta seja otimizada, portanto, devem ser escolhidos criteriosamente.

Um exemplo de algoritmo bastante popular de agrupamento não-hierárquico e que busca um ótimo local é o $K-$Médias ou *K-Means* (Hartigan e Wong, 1979; Lloyd, 1982; MacQueen, 1967). Esse algoritmo é o mais simples que emprega um critério de erro quadrático médio (*mean squared error*) (MacQueen, 1967) dado por: $\sum_{j=1}^{k} \sum_{i=1}^{n_j} \|a_i^j - c_j\|^2$, sendo a_i^j o i-ésimo padrão que pertence ao j-ésimo grupo e c_j é o centróide do j-ésimo grupo.

Na versão mais clássica de heurística de otimização para o $K-$Médias (Lloyd, 1982), o algoritmo inicia dividindo os padrões aleatoriamente em k grupos ou de uma partição pré-definida pelo usuário. Então o centro, ou centróide, desses grupos são calculados e os padrões são realocados para os grupos de acordo com a maior proximidade ao centróide. Para tanto, o centróide é definido como a média aritmética de cada uma das dimensões das distâncias de todos os pontos do grupo. O centróide é então recalculado e novamente os padrões são realocados até que se atinja estabilidade, ou seja, não haja realocação de nenhum padrão de um grupo para outro.

O algoritmo geral para o método $K-$Médias é dado por:

1. Escolha os K centróides iniciais
2. Siga pela lista de padrões, inserindo-os ao grupo cujo centróide é o mais próximo.
3. Recalcule o centróide do grupo recebendo os novos padrões e retirando padrões removidos.
4. Repita os passos 2 e 3 até que não haja mais inserções a serem feitas.

Uma grande quantidade de variantes do $K-$Médias foi desenvolvida, por exemplo, buscando selecionar boas partições iniciais, permitindo a separação e união dos grupos resultantes de acordo com a variância ou selecionando uma função critério diferente. Uma versão mais eficiente que a original foi desenvolvida por (Hartigan e Wong, 1979), que seleciona um padrão e o realoca otimamente, considerando a movimentação do centróide pela realocação dos padrões. A heurística de Hartigan se distingue da de Lloyd no critério de realocação, na eficiência e na busca pelo ótimo local.

O algoritmo geral para o método $K-$Médias de Hartigan é dado por:

1. Escolha os K centróides iniciais.
2. Siga pela lista de padrões, inserindo-os ao grupo cujo centróide é o mais próximo.
3. Recalcule o centróide do grupo recebendo os novos padrões e retirando padrões removidos.
4. Para todos os centróides atualizados no passo anterior
   (a) Calcular a soma dos erros quadrados do grupo
   (b) Para todos padrões do grupo
      i. Calcular a soma dos erros quadrados dos outros grupos com a inclusão do padrão
      ii. Se a soma dos erros quadrados for menor com a inclusão desse padrão em outro grupo, atribuir o padrão ao outro grupo

O algoritmo de Lloyd realoca um padrão do grupo k_1 ao grupo k_2, caso esteja mais próximo do centróide de k_2 do que do centróide de k_1. Já o algoritmo de Hartigan adota a distância Euclideana e segue a conclusão do trabalho de (Sparks, 1973). Esse trabalho propõe que é mais efetivo realocar um padrão do grupo k_1 ao k_2 apenas caso o quadrado da distância Euclideana, ou soma dos erros quadrados, ao centro do grupo k_2 seja menor do que a do centro de k_1, mesmo quando se recalcula o centróide simultaneamente aos reposicionamentos dos padrões. Desta forma há menor número de realocações.

Já quanto à eficiência, o algoritmo de Lloyd é menos eficiente porque calcula as distâncias Euclideanas quadradas entre todos os padrões e todos centróides até que se atinja a convergência. Em contraste, o algoritmo de Hartigan recalcula o quadrado da distância Euclideana entre o padrão e os centróides caso haja possibilidade de realocação. Essa possibilidade de realocação é dada por um conjunto chamado *live set* que contém todos centróides que possuem possibilidade de serem alterados.

Finalmente, o algoritmo de Hartigan consegue amenizar o problema de convergência para um mínimo local ao qual os algoritmos de $K-$Médias são susceptíveis (Telgarsky e Vattani, 2010). O $K-$Médias é susceptível ao problema de convergência para um mínimo local, pois é sensível a escolha da partição inicial, que pode levar a escolha de um mínimo local no espaço de todas partições possíveis, devido a natureza gulosa do algoritmo de realocação (Bradley e Fayyad, 1998; Grim *et al.*, 1998; Moore, 1999). No entanto, como em Hartigan a movimentação dos centróides é considerada na realocação de padrões e um padrão não é sempre atribuído ao grupo com centróide mais próximo, o conjunto de mínimos locais de Hartigan é um subconjunto de mínimos locais de Lloyd e portanto é menos sensível à partição inicial dos dados. Assim, em geral, a heurística de Hartigan chega a um ótimo local melhor do que a heurística de Lloyd (Hartigan e Wong, 1979).

2.1.2.4 *Biclustering*

Uma outra abordagem de agrupamento que também tem sido utilizada na área de bioinformática é o método chamado de *biclustering*, *co-clustering* ou *two-mode clustering* (Mechelen *et al.*, 2004b). No *biclustering* os dados são representados no formato de uma matriz, geralmente tendo linhas como os objetos e colunas como os atributos do objeto, e o agrupamento é realizado simultaneamente nas linhas e colunas baseando-se em algum critério de homogeneidade. Os grupos, chamados *biclusters*, são subconjuntos de linhas e colunas da matriz, formando submatrizes que contêm elementos que exibem comportamento similar dentro deste subconjunto de linhas ou colunas. Por conseguinte, algoritmos de *biclustering* não obrigam que padrões em um mesmo bicluster tenham todos atributos similares, mas que tenham um subconjunto de atributos específicos similares.

Assim, dada a matriz de dados A com n linhas e m colunas, composta dos elementos a_{ij}, com conjunto de índice de linhas $I = \{1, \ldots, n\}$ e índice de colunas $J = \{1, \ldots, m\}$, A pode ser representada por $A(I, J)$. Sendo $I' \subseteq I$ e $J' \subseteq J$, $A_{I'J'} = (I', J')$ representa um uma submatriz da matriz A contendo os elementos a_{ij} com de índice de linhas em I' e índice de colunas em J'. Dessa forma, o método de *biclustering* busca um conjunto de *biclusters*, ou submatrizes, $\{(I_1, J_1), ..., (I_k, J_k)\}$, com $I_i \subseteq I$ e $J_i \subseteq J$, da matriz A, com cada *bicluster* seguindo um dado critério de homogeneidade.

Um *bicluster* pode ser caracterizado como um conjunto de linhas representando um comportamento comum em um conjunto de colunas, ou, um conjunto de colunas representando um comportamento comum em um conjunto de linhas (Kriegel *et al.*, 2009). De acordo com o algoritmo empregado, *biclusters* podem ser exclusivos em linha, quando uma linha só participa de um *bicluster*; em coluna, quando uma coluna só participa de um *bicluster*; em linha e coluna, quando uma linha e uma coluna só participam de um *bicluster*. Pode ainda ser sobreponível se a_{ij} participa de mais de um *bicluster*; exaustivo em linha, quando toda linha pertence a pelo menos um *bicluster* ou em coluna, quando toda coluna participa de pelo menos um *bicluster* ou em linha e coluna quando uma linha e uma coluna participam de pelo menos um *bicluster*. Não há restrições quanto a organização dos *biclusters* e a falta de restrições estruturais nas soluções de *biclustering* permite uma grande liberdade mas é consequentemente mais vulnerável ao *overfitting*(Tanay *et al.*, 2003).

Os algoritmos de *biclustering* podem usar diversos tipos de critérios de homogeneidade, de acordo com os dados, o problema e o objetivo da análise. Um *bicluster* pode consistir, por exemplo, de valores constantes nas linhas e colunas, ou seja, $a_{ij} = a$, sendo a um valor típico, ou aproximadamente constantes, $a_{ij} \approx a$. A constância de valores em colunas, ou $a_{ij} = a + c_j$, sendo c_j o valor de ajuste para coluna j, também pode ser usada como critério de homogeneidade, bem como a constância de linhas, ou $a_{ij} = a + l_i$, sendo l_i o valor de ajuste para linha i. Considerando-se a possibilidade de covariância entre linhas e colunas, o critério de homegeneidade de $a_{ij} = a + l_i + c_j$ pode ser usado. Finalmente, *biclusters* podem ter valores que se modificam ao longo das linhas e/ou colunas não importando o valor exato da alteração.

O problema da busca por *biclusters* de tamanhos máximos em matrizes na forma mais básica (binária contendo 0s e 1s) é equivalente a busca por bicliques com "maximum edge biclique"em um grafo bipartido. Como esse problema é conhecido por ser NP completo (Peeters, 2003), a maioria dos algoritmos utiliza abordagens heurísticas e suposições para encontrar *biclusters*. Uma forma de diminuir o espaço de busca é, por exemplo, procurar por agrupamentos de subespaços de eixos paralelos, como nos algoritmos de busca por *biclusters* de valores constantes ou valores constantes em colunas ou em hiperplanos paralelos aos eixos de atributos irrelevantes, como nos *biclusters* com valores coerentes (Kriegel *et al.*, 2009).

Além de suposições sobre o espaço de busca dos *biclusters*, heurísticas também são usadas em algoritmos de *biclustering*. São comuns as heurísticas de combinação de agrupamento de linhas e colunas, na qual algoritmos de agrupamento são aplicados a linhas e colunas da matriz de dados separadamente e então os resultados são combinados; de divisão e conquista, na qual o problema é dividido em subproblemas similares ao problema original e as soluções combinadas e busca iterativa gulosa, buscando por soluções ótimas locais em busca de soluções ótimas globais. Além dessas heurísticas também são aplicadas heurísticas de enumeração exaustiva de *biclusters* que realizam buscas exaustivas, mas assumindo restrições no tamanho dos *biclusters* e identificação de parâmetros de distribuição que assumem modelos estatísticos e procuram identificar os melhores parâmetros para cada conjunto específico de dados com a minimização de um dado critério.

2.1.2.5 Bimax

O Bimax é um algoritmo de *biclustering* apresentado em (Prelic *et al.*, 2006) que aplica a divisão e conquista na busca por *biclusters*. O conjunto de dados é representado por m colunas e n linhas em uma matriz binária A^{nxm} e um *bicluster* (I', J') corresponde a um subconjunto de linhas $I' \subseteq 1, ..., n$ que possuem comportamento comum em um subconjunto de colunas $J' \subseteq 1, ..., m$. O par (I', J') define uma submatriz $A_{I'J'}$ para qual todos elementos são iguais a 1 e que é de inclusão maximal, isto é, que não está inteiramente contido em qualquer outro *bicluster*(Prelic *et al.*, 2006).

O algoritmo (Figura 2.2, extraída de (Prelic *et al.*, 2006)) inicialmente divide as colunas da matriz nos subconjuntos J_V e J_U (Figura 2.3, também extraída de (Prelic *et al.*, 2006)), sendo que em J_U temos as colunas com valores 1 de uma das linhas da matriz. As linhas de A são então ordenadas de forma que primeiro tenhamos as linhas com valores 1 em J_U, definindo I_U; depois as linhas com valores 1 em J_U e J_V, definindo I_W e somente então as linhas com valores 1 em J_V, definindo I_V. Finalmente, U e V são definidos como (I_U, J_U) e (I_V, J_V), respectivamente.

A matriz A é, desta forma, particionada em 3 submatrizes, uma contendo células com valores 0 que será dispensada, e outras duas chamadas V e U às quais o algoritmo será reaplicado até que as matrizes contenham apenas valores 1. Se U e V não partilham nenhuma coluna de A, ou seja, I_W é vazio, as duas matrizes podem ser processadas independentemente, caso contrário, é necessário gerar somente os *biclusters* em V que compartilham pelo menos uma coluna em comum com J_V.

O Bimax requer recursos de memória proporcional a $O(nm\ min\{n, m\})$ e possui complexidade de tempo de pior caso de $O(nm\beta)$ para matrizes com biclusters disjuntos e $O(nm\beta\ min\{n, m\})$ para matrizes aleatórias, com β sendo o número de todos *biclusters* de inclusão máxima em A^{nxm}. No entanto, como o número de biclusters de inclusão máxima pode ser exponencial em m e n, geração de todos os grupos pode ser inviável.

```
procedure Bimax(A)
    Z ← ∅
    M ← conquer(A, ({1, ..., n}, {1, ..., m}), Z)
    return M
end procedure

procedure conquer(A, (I, J), Z)
    if ∀i ∈ I, j ∈ J : a_ij = 1 then
        return {(I, J)}
    end if
    (I_U, I_V, I_W, J_U, J_V) = divide(A, (I, J), Z)
    M_U ← ∅, M_V ← ∅
    if I_U ≠ ∅ then
        M_U ← conquer(  , (I_U ∪ I_W, J_U), Z)
    end if
    if I_V ≠ ∅ ∧ I_W = ∅ then
        M_V ← conquer(A, (I_V, J_V), Z)
    else if  W ≠ ∅ then
        Z' ← Z ∪ {J_V}
        M_V ← conquer(A, (I_W ∪ I_V, J_U ∪ J_V), Z')
    end if
    return M_U ∪̇ M_V
end procedure

procedure divide(A, (I, J), Z)
    I' ← reduce(A, (I, J), Z)
    choose i ∈ I' with 0 < Σ_{j∈J} a_ij < |J|
    if such an i ∈ I' exists then
        J_U ← {j | j ∈ J ∧ a_ij = 1}
    else
        J_U = J
    end if
    J_V ← J \ J_U                                   J
    I_U ← ∅, I_V ← ∅, I_W ← ∅
    for each i ∈ I' do
        J* ← {j | j ∈ J ∧ a_ij = 1}
        if J* ⊆ J_U then
            I_U ← I_U ∪ {i}
        else if  * ⊆  V then
            I_V ← I_V ∪ {i}
        else
            I_W ← I_W ∪ {i}
        end if
    end for
    return (I_U, I_V, I_W, J_U, J_V)
end procedure

procedure reduce(A, (I, J), Z)
    I' ← ∅
    for each i ∈ J do
        J* ← {j | j ∈ J ∧ a_ij = 1}
        if J* ≠ ∅ ∧ ∀  + ∈ Z : J^+ ∩ J* ≠ ∅ then
            I' = I' ∪ {i}
        end if
    end for
    return I'
end procedure
```

Figura 2.2: *Algoritmo Bimax (Figura extraída de (Prelic et al., 2006))*

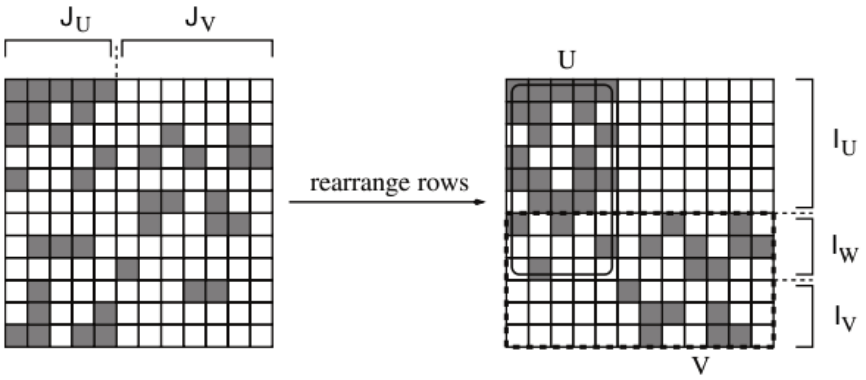

Figura 2.3: *Simulação do algoritmo Bimax (extraída de (Prelic et al., 2006))*

2.2 Trabalhos Relacionados

Trabalhos anteriores empregaram diversas metodologias para examinar sequências de protease e transcriptase reversa em busca de padrões de mutações (Alteri *et al.*, 2009; Doherty *et al.*, 2011; Gonzales *et al.*, 2003; Hoffman *et al.*, 2003; Liu *et al.*, 2008; Reuman *et al.*, 2010; Rhee *et al.*, 2003; Sing *et al.*, 2005; Wu *et al.*, 2003; Yahi *et al.*, 1999).

O objetivo do estudo de (Yahi *et al.*, 1999) foi avaliar mutações em protease e transcriptase reversa relacionadas a resistência em 302 pacientes durante tratamento com distintas combinações de medicamentos. 787 sequências de protease, transcriptase reversa ou ambas proteínas foram analisadas. Teste chi-quadrado, ou teste Kendal para amostras menos numerosas, e teste Fisher *two-tailed* foram utilizados para verificar a significância de associações específicas de mutações e em diferentes subgrupos de carga viral. Em sequências de protease, mutações nas posições 46-10, 46-71, 46-90, 82-71, 82-10, 82-54, 82-90, 90-71, 90-10, 90-46, 90-54 e 90-77. Associações entre mutações na posições 41-210, 210-67, 210-69, 210-219, 210-184, 210-215, 219-67, 219-69 e 219-70 foram encontradas em sequências de transcriptase reversa.

Sequências de proteína de pacientes dos Estados Unidos e Europa que já passaram por tratamento com vários medicamentos foram analisadas em (Gonzales *et al.*, 2003). 485 sequências de protease de pacientes em tratamento com pelo menos 3 inibidores de protease e 487 sequências de transcriptase reversa de pacientes tratados pelo menos 4 inibidores de nucleosídeo foram utilizadas no estudo. Teste exato de Fisher e o procedimento de Benjamini-Hochberg foram aplicados para comparar a prevalência de padrões de mutação envolvendo um, duas ou três posições, considerando-se as posições 10 a 90 de protease e 40 a 240 de transcriptase reversa. Para identificar padrões com mais de três mutações, o agrupamento k-medoids foi usado considerando a presença ou ausência de mutações nas posições 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 88, 82, 84, e 90 de protease e 41, 62, 65, 67, 69, 70, 74, 75, 77, 115, 116, 151, 184, 210, 215 e 219 de transcriptase reversa. Segundo o trabalho, foram escolhidas posições de transcriptase relacionadas a resistência aos NRTIs e posições de protease associadas com a resistência aos inibidores de protease e não polimórficas na ausência de tratamento. Oito agrupamentos de transcriptase reversa foram encontrados, explicando aproximadamente 63% da variabilidade nas posições, incluindo seis grupos compostos primariamente por TAMs (mutações análogas a timidina) (posições 41, 67, 70, 210, 215 e 219), um pelas posições 151, 75, 77 e 116 e outro sem mutações A posição 184 fez parte de 5 centros de grupos. Nove agrupamentos foram encontrados para protease, incluindo um grupo com as mutações 30 e 88. Os nove grupos explicaram aproximadamente 68% da variabilidade nas posições de protease.

Padrões de variabilidade de posições de protease foram examinadas em 1179 sequências traduzidas de protease de HIV subtipo B em (Hoffman *et al.*, 2003). 648 sequências foram extraídas de pacientes fora de tratamento e 531 de pacientes em tratamento, considerando-se apenas posições com aminoácidos variando em pelo menos 5% das sequências, totalizando 30 posições de protease. Informação mútua foi empregada como medida de covariância para todos 465 pares possíveis de posições de protease consideradas variáveis (10, 12, 13, 14, 15, 19, 20, 30, 32, 35, 36, 37, 41, 46, 48, 54, 57, 60, 62, 63, 64, 69, 71, 72, 73, 77, 82, 84, 88, 90 e 93). Um teste de permutação foi utilizado para calcular a significância estatística das interações e uma tabela de contingência indicou associações entre substituições específicas de aminoácidos. 9 pares alcançaram significância em pacientes fora de tratamento e 32 pares em pacientes em tratamento, sendo que 5 desses pares foram significantes em ambos testes. A maioria das associações era positiva e de proximidade na estrutura da proteína. Os autores consideram que algumas associações podem ser dadas pelo fato de que os pacientes passaram por vários tratamentos e por isso possuem grande número de mutações, no entanto, outras provavelmente representam interações cooperativas importantes.

Aproximadamente 6000 sequências de protease e transcriptase reversa foram analisadas em (Rhee *et al.*, 2003), 27 do subtipo C, 15 do subtipo A e 7 do subtipo D. Foram consideradas as posições 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 82, 84, 88 e 90 de protease e 41, 44, 62, 65, 67, 69, 70, 74, 115, 116, 118, 151, 184, 210, 215 e 219 de transcriptase reversa. De acordo com a combinação de padrões específicos de mutação, dados de susceptibilidade foram obtidos do *Stanford RT and Protease Database*. Os 30 padrões de mutações mais frequentes incluiram 55% das mutações de protease, 46% das mutações ditas relacionadas a resistência aos NRTIs e 66% das mutações ditas relacionadas a resistência aos NNRTIs.

2244 sequências de protease subtipo B com históricos de tratamento bem documentados foram avaliados em (Wu *et al.*, 2003), sendo 1004 de pacientes que não passaram por tratamento e 637 de pacientes que receberam um inibidor de protease e 603 que receberam um ou mais inibidores de protease. Com intuito de investigar a correlação entre posições, os coeficientes de correlação binomial (phi) foram calculados para ocorrência simultânea de duas mutações separadamente nos grupos de pacientes em tratamento (45 pares) e fora de tratamento (17 pares). A análise de componentes principais das posições também foi realizada nos dados de pacientes em tratamento, usando a matriz de coeficientes de correlação binomial como medida de similaridade. Uma busca exaustiva foi usada para agrupar os pares de resíduos considerados covariantes.

Para explicar pares de posições covariantes que não se localizavam próximo na estrutura tridimensional do proteína,uma análise de cadeia de Markov foi realizada para encontrar a menor cadeia contendo resíduos correlacionados a menos de 8 $\AA$. Com intuito de verificar se as cadeias eram estatísticamente significativas, uma análise de permutação foi realizada. Nessa análise, pares de posições eram escolhidos aleatoriamente e verificava-se se esses resíduos podiam estar ligados por uma cadeia. 115 pares de posições de protease tiveram covariação estatísticamente significantes, dentre os quais 59 estão próximos na conformação da proteína.

(Sing *et al.*, 2005) considera que existem três mais importantes complexos de mutações relacionados a mecanismos de resistência em sequências de transcriptase reversa, o complexo TAM1, o complexo TAM2 e o complexo com a posição 151 como principal. Assim, o agrupamento hierárquico com coeficiente de correlação de Matthews como medida de similaridade é aplicado a sequências de transcriptase reversa de 1355 pacientes com falha terapêutica. O objetivo da análise é identificar quais mutações agrupam com qual complexo de mutações. Adicionalmente, usando teste exato de Fisher foram procurados pares de interações associadas com tratamento com análogos de nucleosídeo inibidores de transcriptase reversa (NRTI), incluindo TAMs. Ainda, com escalonamento multidimensional (MDS) as interações positivas das mutações com as mutações TAM1 e TAM2 foram testadas. O teste exato de Fisher obteve resultados compatível com a existência dos complexos de mutações citados, no entanto, há interações entre os complexos TAM1 e TAM2. Os resultados do agrupamento hierárquico e do MDS mostram que a maioria das posições que o trabalho considera foram agrupadas com o complexo TAM1 e um novo grupo é formado a partir de mutações que aparecem mais frequentemente em grupos sem tratamento.

(Liu *et al.*, 2008) aplica a análise de mutações correlacionadas (CMA) para identificar grupos de posições de protease altamente covariantes em sequências de 7758 pacientes em tratamento e 8761 pacientes sem tratamento. A CMA quantifica a covariância em alinhamento múltiplo de sequências usando informação mútua como medida de correlação, seguido por agrupamento espectral. O objetivo dessa análise é identificar ligações que são diretamente relevantes para função e estrutura da proteína. No entanto, a covariância pode também ser resultado de sinais evolutivos. Também foi aplicado o particionamento K-way usando k valendo 3, 4 e 5. A abordagem distinguiu um grupo de mutações que conferem resistência a mais de um medicamento e outro relacionado a diferenças entre subtipos.

213 sequências de protease e transcriptase reversa de HIV subtipo B foram examinadas em (Alteri *et al.*, 2009). Foram consideradas as posições 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 82, 84, 88, 90 de protease e 41, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 115, 116, 151, 181, 184, 188C, 190, 210, 215, 219, 225, 230, 236 de transcriptase reversa. Com essas posições, foram calculados todos os pares de coeficiente de correlação binomial e para identificar os pares de combinações significantivos no teste de múltiplas hipóteses, foi utilizado o procedimento de Benjamini-Hochberg. Após identificação dos pares, foi realizado agrupamento dos mesmos incluindo grupos com duas ou mais posições como forma de análise de estrutura de covariância de mutações. Os pares mais correlacionados encontrados pelo estudo foram 210-41, 210-215 e 210-103. O agrupamento indicou a existência de um grupo contendo mutações importantes em pacientes recém diagnosticados de subtipo B, incluindo mutações nas posições 41, 210, 215, 103 e 60.

13039 sequências de trancriptase reversa de vírus do grupo M de pacientes que já receberam apenas efavirenz ou nevirapine como inibidor de transcriptase reversa não análogo a nucleosídeo (NNRTI) foram analisadas em (Reuman *et al.*, 2010). 30 posições de transcriptase reversa foram consideradas: 90, 94, 98, 100, 101, 102, 103, 105, 106, 108, 138, 139, 178, 179, 181, 188, 190, 221, 223, 225, 227, 230, 232, 234, 236, 237, 238, 241, 242 e 318. O coeficiente de similaridade de Jaccard foi aplicado na identificação de correlação de pares com o método de Holm para controlar a taxa de erro da família para múltiplas comparações em pares. Plot poissoness foi utilizado para modelar as distribuições de mutações em cada sequência e um boostrap paramétrico usando simulação multivariada de Poison para se acessar a distribuição de correlações positivas e negativas. Padrões de três ou mais mutações NNRTI entre as selecionadas foram buscadas e sua frequência de ocorrência calculada. 29 de 1288 pares de mutações possuiam covariâncias positivas e 116 pares covariância negativa. Foram identificados 57 agrupamentos.

Padrões de resistência fenotípica multi-PI (inibidor de protease) foram buscados com a aplicação de agrupamento baseado em programação inteira globalmente ótima em (Doherty *et al.*, 2011). 398 sequências de protease fenotipadas para amprenavir, atazanavir, indinavir, lopinavir, nelfinavir, ritonavir, saquinavir, tipranavir e darunavir foram agrupadas com base nos valores de resistência a cada um dos medicamentos. Cada sequência foi posicionada em um ponto no espaço de resistência dos medicamentos de dimensão 9. A medida de similaridade empregada foi composta pela distância euclideana fenotípica somada à genotípica, dada pelo número de diferentes aminoácidos em cada posição de protease. O agrupamento foi validado pela sua efetividade em prever o nível de resistência a um medicamento, baseado na resistência dos outros medicamentos com validação cruzada *n-fold*. Esse estudo tinha como objetivo encontrar sequências e mutações representativas para fenótipos de resistência elucidando suas assinaturas genotípicas. A análise de agrupamento gerou agrupamentos com sequências resistentes a todos medicamentos, todos, com exceção

de um medicamento, um grande subconjunto de medicamentos e somente um medicamento. As sequências representativas corroboram com observações realizadas previamente ao trabalho.

2.2.1 Conclusão dos Trabalhos Relacionados

Tais estudos sobre os padrões de mutações em protease e transcriptase reversa são importantes porque a interação entre mutações podem resultar em diferentes respostas a tratamentos. Por exemplo, uma mutação pode compensar pela perda de *fitness* causada por outra mutação que confere resistência aos medicamentos. No entanto, alguns do estudos anteriores apenas investigam interações entre pares de mutações e a maioria deles apenas analisam sequências de vírus de subtipo B. Além disso, estudam diferentes conjuntos de posições de proteína e não é possível comparar os resultados. Portanto, padrões de mutações ainda não foram bem caracterizados em sequências de protease e transcriptase reversa. A caracterização desses padrões pode levar ao melhor entendimento da interação entre essas mutações e classificação dessas sequências, correlacionando genótipo e resistência aos medicamentos permitindo a personalização do tratamento de pacientes e do desenvolvimento de novos medicamentos mais eficazes contra variantes resistentes. Entendendo a relação entre o genótipo e a resistência aos medicamentos é possível esclarecer um pouco mais a dinâmica do desenvolvimento de resistência cruzada, importante fator de influência no sucesso no tratamento.

O esclarecimento da relação entre genótipo e a suceptibilidade também pode colaborar no desenvolvimento de vacinas contra o HIV (Korber *et al.*, 1993; Wu *et al.*, 2003). Isso porque os padrões de mutações do HIV podem influenciar na elucidação de mecanismos de escape do sistema imune (Carlson *et al.*, 2008), que são relevantes para a pesquisa de vacinas (Brumme *et al.*, 2009).

Dessa forma, no presente estudo, um grande número de posições de proteína (44 de protease e 38 de transcriptase reversa) de vírus dos subtipos B, C e F foram aplicados a métodos de agrupamento e *biclustering* e sequências foram classificadas de acordo com a ocorrência de padrões de mutações. Esses agrupamentos e *biclusters* foram comparados com padrões de mutações estudados anteriormente e com o conhecimento atual em padrões de mutação.

Capítulo 3

Desenvolvimento do framework

3.1 Introdução

O R Project for Statistical Computing (Team, 2008) é uma linguagem e um ambiente que disponibiliza uma grande variedade de técnicas e gráficos estatísticos (http://www.r-project.org/). Dentre os pacotes disponibilizados pelo R estão uma grande quantidade de implementações de algoritmos de agrupamento e *biclustering*(como pode ser visto em: http://cran.cnr.berkeley.edu/web/views/Cluster.html).

Neste projeto utilizamos os pacotes *The R Stats Package* (http://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html) e *biclust* (http://cran.r-project.org/web/packages/biclust/). O pacote *The R Stats Package* contém funções para cálculos estatísticos e geração aleatória de números, incluindo a implementação do $K-$Médias de (Forgy, 1965), (Hartigan e Wong, 1979), (Lloyd, 1982) e (MacQueen, 1967). Já o pacote *biclust* possui a implementação de diversos algoritmos de *biclustering* como Cheng and Church (Cheng e Church, 2000), Spectral (Kluger *et al.*, 2003), Plaid Model (Lazzeroni e Owen, 2002), Xmotifs (Murali e Kasif, 2003) and Bimax (Prelic *et al.*, 2006).

Além da disponibilização de implementações de algoritmos de reconhecimento de padrões, o R Project for Statistical Computing também fornece métodos de visualização de agrupamentos e *biclusters*. Para o $K-$Médias pode-se utilizar, por exemplo, as funções *plot* (Figura 3.1), *plotCluster* (Figura 3.2) ou *silhouette* (Figura 3.3) e para o Bimax *bubbleplot* (Figura 3.4), *biclustbarchart* 3.5) ou *biclustmember* (Figura 3.6).

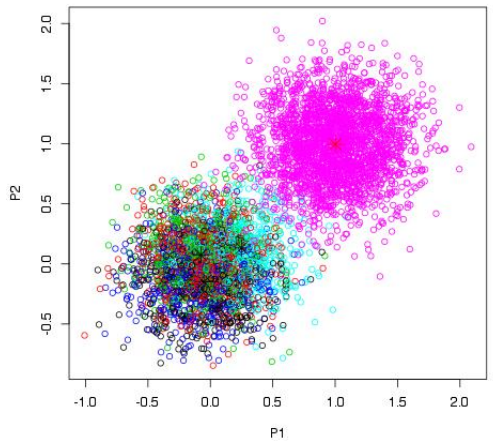

Figura 3.1: *Exemplo de gráfico gerado pelo plot representando 6 grupos encontrados pelo $K-$Médias em uma matriz de dados de dimensões* 5000×20

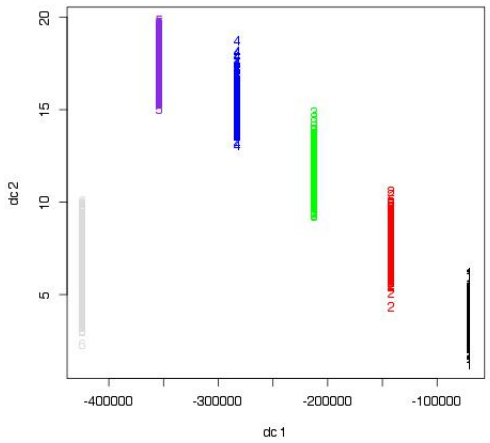

Figura 3.2: *Exemplo de gráfico gerado pelo plotCluster representando 6 grupos encontrados pelo $K-$Médias em uma matriz de dados de dimensões* 5000×20

No entanto a visualização para grandes volumes de dados e com grande dimensões, pode ser complexa. Por exemplo, nas Figuras 3.1, 3.2, 3.3, 3.4, 3.5 e 3.6 mostramos a aplicação do $K-$Médias e Bimax a uma matriz de dimensões 5000×20 gerada aleatoriamente. Por essas figuras é possível observar que nem todos os gráficos são capazes de apresentar toda a informação em uma única imagem e alguns não são adequados para tal volume de dados.

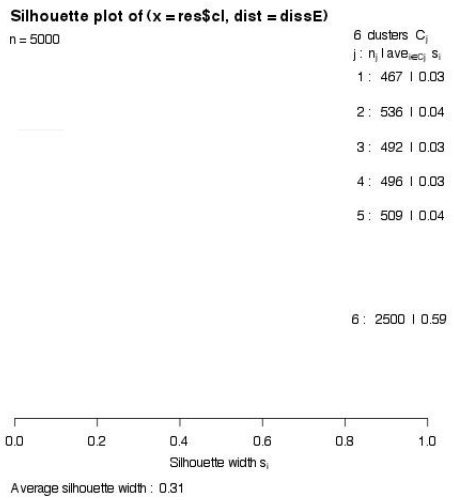

Figura 3.3: *Exemplo de gráfico gerado pelo silhouette representando 6 grupos encontrados pelo $K-$Médias em uma matriz de dados de dimensões* 5000×20

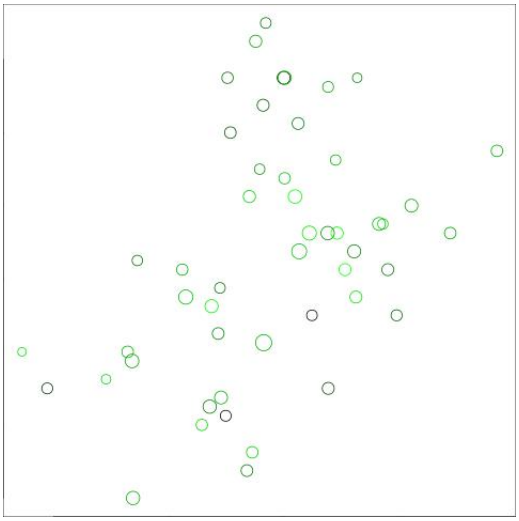

Figura 3.4: *Exemplo de gráfico gerado pelo bubbleplot representando biclusters encontrados pelo Bimax em uma matriz de dados de dimensões* 5000×20

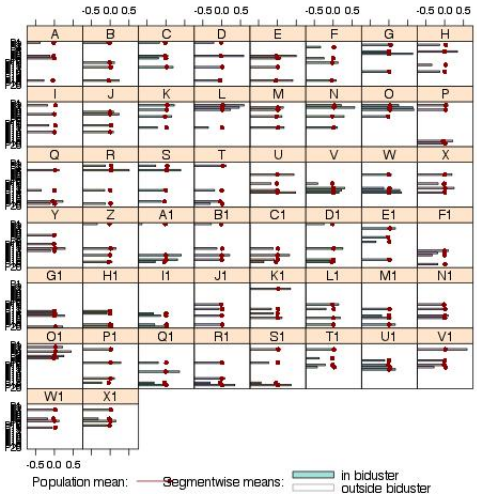

Figura 3.5: *Exemplo de gráfico gerado pelo biclustbarchart representando biclusters encontrados pelo Bimax em uma matriz de dados de dimensões* 5000×20

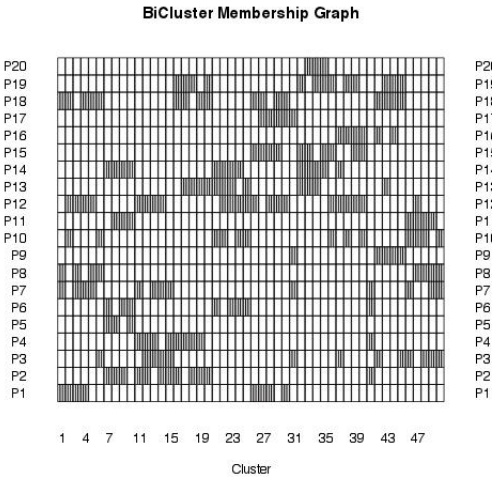

Figura 3.6: *Exemplo de gráfico gerado pelo biclustmember representando biclusters encontrados pelo Bimax em uma matriz de dados de dimensões* 5000×20

Além disso, é importante que a imagem possa ser facilmente compreendida e interpretada por especialistas no domínio dos dados. Por esses motivos, implementamos *scripts* para geração de imagens apropriadas para a aplicação com dados binários e capazes de representar e sintetizar a informação contida nos grupos e *biclusters*.

Desta forma, na busca por grupos e *biclusters*, *scripts* em Perl foram implementados para binarização das sequências e os algoritmos $K-$Médias e Bimax dos pacotes *The R Stats Package* e *biclust* foram utilizados, bem como gráficos *biclustmember* para os *biclusters*.

Para criação de histogramas também foram implementados *scripts* em Perl com chamadas ao *software* Gnuplot (Williams *et al.*, 2012), que emprega linhas de comando na geração de gráficos de funções matemáticas. Após a geração dos histogramas, mais *scripts* foram criados em Perl para elaboração das imagens binárias no formato *Portable Pixmap* (PPM), tabelas descrevendo os grupos e *biclusters* e imagens coloridas no formato PPM.

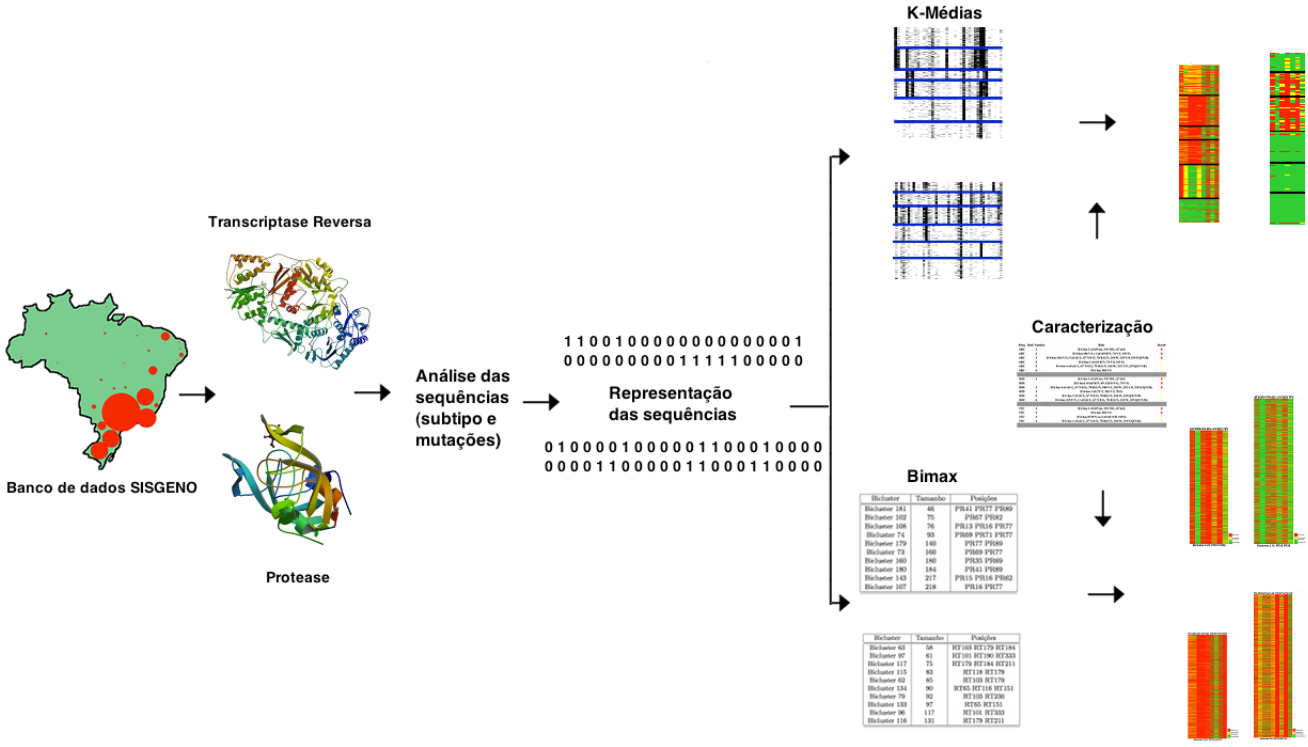

Figura 3.7: *Pipeline resumindo o framework proposto. Sequências de protease e transcriptase reversa foram reunidas de pacientes do Brasil inteiro, foram alinhadas a sequências consenso e foi determinado seu subtipo. Em seguida foram binarizadas e submetidas aos algoritmos de agrupamento e biclustering. Os grupos e biclusters foram caracterizados e comparados com as predições da look-up table brasileira.*

3.2 Pipeline

A Figura 3.7 resume o *framework* criado para analisar as sequências de protease e transcriptase reversa (disponível em www.ime.usp.br/~mcintho). Esse *framework* agrega as funcionalidades adicionais implementadas e acopladas ao R de modo a viabilizar as classificações das mutações de HIV. Inicialmente, sequências de pacientes distribuídos em 27 estados brasileiros foram extraídas do banco de dados do SISGENO (Sistema de Controle de Exames de Genotipagem). O grande número de locais de origem dos pacientes é importante para amostragem de variabilidade do vírus. A porcentagem de sequências de cada estado com relação à totalidade dos dados foi representada na Figura 3.7 pelo mapa do Brasil com o auxílio dos círculos vermelhos.

As sequências foram então analisadas pelo *webservice* Sierra (http://hivdb.stanford.edu/DR/webservices/index.html) do *HIV Drug Resistance Database* (http://hivdb.stanford.edu/), indicando seu subtipo e a presença de mutações. A partir da análise de mutações, as sequências foram transformadas para forma binária por meio de código implementado em linguagem Perl. Em seguida foram submetidas a algoritmos de agrupamento e *biclustering* dos pacotes *The R Stats Package* e *biclust* do R. Com os grupos e *biclusters* gerados, foram criados histogramas para representá-los por meio de *scripts* em Perl e Gnuplot.

Além dos histogramas, imagens e tabelas também foram criadas por meio de *scripts* em Perl. Os grupos e *biclusters* foram caracterizados de acordo com a ocorrência de mutações e comparados com a predição de resistência aos medicamentos da *look-up table* brasileira também por imagens geradas por *scripts* em Perl. A metodologia é descrita com mais detalhes nas seções seguintes.

3.3 Análise das Sequências

No Brasil, o Programa Nacional de AIDS do ministério da saúde oferece o teste de genotipagem a pacientes em falha terapêutica através de uma rede de laboratórios chamada Rede Nacional de Genotipagem (RENAGENO). Em apoio a essa rede, o SISGENO (http://www.aids.gov.br/sisgeno/), um sistema informativo foi desenvolvido. Essa ferramenta permite que a equipe médica e laboratorial mantenha informações sobre resultados de exames e armazene dados para análises futuras. Entre as informações armazenadas no SISGENO estão os resultados de testes de genotipagem de pacientes de todo Brasil. Esse conjunto de sequências é uma fonte importante de conhecimento sobre as linhagens circulantes de HIV no país.

O presente estudo utilizou 11.454 sequências de protease e transcriptase reversa obtidas através do SISGENO. Todas as sequências foram alinhadas com sequências HIV-1 subtipo B consenso (representadas nas Figuras 3.8 e 3.9) mantidas em Los Alamos HIV Sequence Database (http://hiv-web.lanl.gov). Esse banco de dados possui sequências anotadas de HIV, mutações de resistência aos medicamentos, epítopos de HIV e resultados de testes de vacinas, bem como as sequências consenso utilizadas. As sequências consenso são derivadas de um alinhamento de sequências do subtipo B e são utilizadas para comparação com novas sequências.

```
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGI
GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
```

Figura 3.8: *Sequência consenso da protease subtipo B utilizada*

```
PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKI
GPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGL
KKKKSVTVLDVGDAYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNVLP
QGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRT
KIEELRQHLLRWGFTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEKD
SWTVNDIQKLVGKLNWASQIYAGIKVKQLCKLLRGTKALTEVIPLTEEAE
LELAENREILKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQEPFKNLK
TGKYARMRGAHTNDVKQLTEAVQKIATESIVIWGKTPKFKLPIQKETWEA
WWTEYWQATWIPEWEFVNTPPLVKLWYQLEKEPIVGAETFYVDGAANRET
KLGKAGYVTDRGRQKVVSLTDTTNQKTELQAIHLALQDSGLEVNIVTDSQ
YALGIIQAQPDKSESELVSQIIEQLIKKEKVYLAWVPAHKGIGGNEQVDK
LVSAGIRKVL
```

Figura 3.9: *Sequência consenso da transcriptase reversa subtipo B utilizada*

Os alinhamentos das sequências com as sequências consenso foram realizados por meio do *webservice* Sierra (http://hivdb.stanford.edu/DR/webservices/index.html) do *HIV Drug Resistance Database* (http://hivdb.stanford.edu/). O *HIV Drug Resistance Database* é um banco de dados utilizado para representar, armazenar e analisar dados de HIV relacionados à resistência aos medicamentos. Já o *webservice* Sierra foi criado para facilitar a entrada de grandes volumes de dados ao HIVdb (http://sierra2.stanford.edu/sierra/servlet/JSierra). Dentre os possíveis métodos de análise de dados do HIVdb está o algoritmo de alinhamento local de sequências de nucleotídeos para aminoácidos de Huang *et al.* (1997). Esse alinhamento foi empregado para determinar as posições das sequências de proteínas que contém mutações em relação às sequências referência.

Além do alinhamento das sequências com sequências consenso, o *webservice* também fornece dados sobre os subtipos das sequências dos vírus. A determinação do subtipo é dada pela comparação com sequências referência, como descrito em Gifford *et al.* (2006). As referências incluem sequências HIV-1 do grupo *Main* de subtipos A, B, C, D, F, G, H, J, K, CRF01_AE e CRF02_AG. A determinação do subtipo das sequências pelo Sierra possibilitou a divisão dos dados em 10.229 sequências de subtipo B, 424 de subtipo C e 801 de subtipo F.

3.4 Representação das Sequências

Os algoritmos de agrupamento e *biclustering* particionam um conjunto de dados em grupos ou *biclusters* utilizando uma medida de similaridade como parâmetro de comparação. Essa medida de similaridade informa ao algoritmo as características que devem ser consideradas na tomada de decisão do particionamento dos dados, ou seja, ajuda na distinção entre dados similares que devem pertencer a um mesmo grupo ou *bicluster* e dados dissimilares que devem pertencer a grupos ou *biclusters* distintos.

No caso de sequências de proteína e do estudo de seus padrões de mutações, a informação essencial para o agrupamento dos dados está na presença ou ausência dessas mutações. Considerando-se a presença ou a ausência de mutações, a maneira mais simples de representação dessa informação é o mapeamento binário das sequências ou o mapeamento bitmap. Esse mapeamento é dado utilizando-se apenas os valores 0 e 1, por exemplo, com 0 indicando a ausência de mutação e 1 indicando a presença de mutação em cada posição de aminoácido das proteínas.

Assim, para simplificar a representação e comparação das sequências de protease e transcriptase reversa, um mapeamento bitmap foi empregado. Nesse mapeamento, se uma sequência possuísse o mesmo aminoácido da sequência selvagem, ele seria trocado pelo valor 0 e quando a sequência possuísse um aminoácido diferente,

era substituído pelo valor 1, como em Reuman *et al.* (2010). Os dados podem portanto ser interpretados como vetores binários em um espaço N dimensional, com N valendo 99 para protease e 335 para transcriptase reversa, o número de aminoácidos considerados das duas sequências.

No entanto, quando se trabalha com padrões em espaços de alta dimensão, a maldição da alta dimensionalidade se torna um problema. A maldição da alta dimensionalidade faz com que todas distâncias sejam parecidas em espaços de grandes dimensões (Kriegel *et al.*, 2009). Consequentemente, padrões localizados em pontos próximos no espaço têm valores de distâncias parecidas com a de padrões em pontos distantes no espaço. Por conseguinte, não é possível medir as similaridades entre os padrões e nem realizar o seu particionamento.

Uma maneira comum de se evitar a maldição da alta dimensionalidade dos dados em casos em que vários atributos dos dados são correlacionados ou somente algumas características são relevantes é o emprego da seleção de características (Kriegel *et al.*, 2009). Na seleção de características, um subconjunto do conjunto de atributos dos dados são selecionados, de forma a diminuir a dimensionalidade dos dados.

Assim, 38 posições de transcriptase reversa e 44 posições de protease representadas na Tabela 3.1 foram selecionadas a partir das posições utilizadas na *look-up table* brasileira e indicações de especialistas. Essas posições são conhecidas por estarem relacionadas a resistência aos medicamentos e foram selecionadas de um total de 335 aminoácidos da transcriptase reversa e 99 da protease. Desta forma, as sequências de protease foram representadas por vetores binários em $N = 44$ e a transcriptase em $N = 38$. O processo de representação dos dados está resumido nas Figuras 3.10 e 3.11.

Tabela 3.1: *Posições selecionadas das sequências de Protease e Transcriptase Reversa*

	Proteína	Posição na proteína	Proteína	Posição na proteína
1	Transcriptase Reversa	41	Protease	8
2	Transcriptase Reversa	44	Protease	10
3	Transcriptase Reversa	50	Protease	11
4	Transcriptase Reversa	65	Protease	13
5	Transcriptase Reversa	67	Protease	15
6	Transcriptase Reversa	69	Protease	16
7	Transcriptase Reversa	70	Protease	20
8	Transcriptase Reversa	74	Protease	24
9	Transcriptase Reversa	75	Protease	30
10	Transcriptase Reversa	77	Protease	32
11	Transcriptase Reversa	98	Protease	33
12	Transcriptase Reversa	100	Protease	34
13	Transcriptase Reversa	101	Protease	35
14	Transcriptase Reversa	103	Protease	36
15	Transcriptase Reversa	106	Protease	41
16	Transcriptase Reversa	108	Protease	43
17	Transcriptase Reversa	115	Protease	45
18	Transcriptase Reversa	116	Protease	46
19	Transcriptase Reversa	118	Protease	47
20	Transcriptase Reversa	151	Protease	48
21	Transcriptase Reversa	157	Protease	50
22	Transcriptase Reversa	179	Protease	53
23	Transcriptase Reversa	180	Protease	54
24	Transcriptase Reversa	181	Protease	57
25	Transcriptase Reversa	184	Protease	58
26	Transcriptase Reversa	188	Protease	60
27	Transcriptase Reversa	190	Protease	62
28	Transcriptase Reversa	208	Protease	63
29	Transcriptase Reversa	210	Protease	67
30	Transcriptase Reversa	211	Protease	69
31	Transcriptase Reversa	214	Protease	70
32	Transcriptase Reversa	215	Protease	71
33	Transcriptase Reversa	219	Protease	73
34	Transcriptase Reversa	225	Protease	74
35	Transcriptase Reversa	227	Protease	76
36	Transcriptase Reversa	230	Protease	77
37	Transcriptase Reversa	236	Protease	82
38	Transcriptase Reversa	333	Protease	83
39			Protease	84
40			Protease	85
41			Protease	88
42			Protease	89
43			Protease	90
44			Protease	93

Além da seleção de posições das proteínas, outro método aplicado para evitar a maldição da alta dimensionalidade dos dados foi a representação com mapemanto bitmap. Em trabalhos anteriores (Beerenwinkel *et al.* (2005); Sing *et al.* (2005)), uma outra representação binária foi utilizada, mapeando cada bit a diferentes possibilidades de mutações para diferentes aminoácidos. Nessa representação ao invés de uma posição de proteína ser representada por um bit, era representada por n bits, sendo n o número de diferentes aminoácidos que ocorriam no conjunto de dados naquela posição. Essa representação fornece mais detalhes sobre as mutações, mas também aumenta a dimensionalidade dos vetores de dados. Por esse motivo, as representações da protease e transcriptase reversa com mapemanto bitmap e em espaço de dimensões 42 e 38 (posições descritas na Tabela 3.1), respectivamente, foram preferidas.

```
PQITLWQRPLVKIKVGGQVKEALLDTGADDTVLEEMNLPGKWKPKMIGGI
GGFIKVRQYDQISIEICGHKAIGTVLVGPTPINIIGRNLLTQIGCTLNF
                          ↓
0010000000010010001000000000000000000010001000000000
0000000000001000000000000000000010000000000000000000
                          ↓
  00001000000000100000000000010000000010000000000
```

Figura 3.10: *Mapeamento bitmap e seleção de posições da protease*

```
PISPIETVPVKLKPGMDGPRVKQWPLTEEKIKALVEICAELEKEGKISKIGPENPYNTPIFAIRKKDS
TKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDPDFRKYTAFTIPSTN
NETPGVRYQYNVLPQGWKGSPAIFQSSMTRILEPFRKQNPEIIIYQYVDDLYVASDLEIGQHRRKIEE
LRQHLSRWGFFTPDKKHQKEPPFLWMGYELHPDTWTVQPIVLPEKDSWTVNDIQKLVGKLNWASQIYA
GIKVKQLCKLLRGTKALTEVlPLTEEAELELAENREILKEPVHGVYYDPSKDLIAEIQKQGQG
                          ↓
0000000000000000000100000000000000000010100000000000000000010001000
0000000000000000000000000000000000000000000000000000001000000000000
1000000100000000000000000000000100000000001010000100000100000000010
0000000010001100000000000000000000000100000000000000000000000000000
0000000010000000000000000100000000000000000000000000000000000000000
                          ↓
          1000000000000000000001001010101011000000
```

Figura 3.11: *Mapeamento bitmap e seleção de posições da transcriptase reversa*

3.5 $K-$Médias

Com intuito de gerar grupos de sequências de protease e transcriptase reversa usando algoritmos de reconhecimento de padrões, o algoritmo $K-$Médias do R Project for Statistical Computing (Team, 2008) com a implementação de Hartigan (Hartigan e Wong, 1979) foi aplicado a 10.229 sequências de protease e transcriptase reversa de subtipo B, 424 de subtipo C e 801 de subtipo F, separados por subtipo e proteína. O $K-$Médias foi utilizado para buscar por agrupamentos de sequências com k variando de 2 a 16, ou seja, agrupando os dados em 2 a 16 grupos distintos. Para valores de k maiores que 16 o $K-$Médias não foi capaz de convergir e por isso foi usado o valor máximo de 16.

Uma vez que a simulação de até 100 partições aleatórias iniciais gerou grupos com somas de mínimos quadrados parecidas com a simulação de 10 partições aleatórias, para cada valor de k foram simuladas 10 configurações de partições aleatórias iniciais e o agrupamento com maior minimização de soma de quadrados foi escolhido. A utilização de diferentes partições aleatórias iniciais foi empregada como forma de se aumentar a probabilidade de se obter um agrupamento ótimo global, uma vez que o algoritmo pode não alcançar um ótimo global, particularmente para dados em espaços de grandes dimensões.

3.6 Caracterização do Agrupamento

Um problema que surgiu da geração dos agrupamentos foi a visualização e interpretação dos grupos no domínio das mutações de HIV. Esse problema é consequência do grande número de sequências, que dificulta a visualização e a caracterização dos agrupamentos. Imagens podem ser utilizadas para resolver esse problema, já que representam uma ferramenta de visualização informativa e intuitiva para defender e validar resultados, bem como formular e testar hipóteses. Quando a pesquisa envolve análise de dados, o uso de imagens se torna ainda mais importante, já que o volume de dados faz com que seja mais difícil manipular e visualizar diretamente os dados. Portanto, imagens podem ser uma forma de resumir dados e resultados.

Desta forma, com objetivo de analisar os grupos, verificar se seguem padrões de mutações e caracterizar esses padrões, imagens binárias foram criadas, inspiradas na visualização de dados de microarranjo. Para tanto, as sequências de proteínas foram representadas como linhas e as posições de aminoácido como colunas. As sequências, ou linhas, foram agrupadas de acordo com o rótulo de grupo designado e os grupos separados por linhas azuis. Quando uma sequência possuía valor 1 na posição de aminoácido, a posição era representada por pixels pretos e quando a posição possuía valor 0, a posição era representada por pixels brancos. Assim, seis imagens foram criadas para cada valor combinado de k, proteína e subtipo.

Os pixels brancos e pretos utilizados foram importantes na distinção entre grupos, acentuando as diferenças e descrevendo-os. Os pixels também colaboraram na visualização das posições de aminoácido responsáveis pela representação e caracterização dos grupos, resumindo a informação contida nas sequências e grupos.

Para extrair mais detalhes, histogramas foram gerados para cada grupo, proteína e subtipo, mostrando a porcentagem das sequências em cada grupo contendo mutações em uma dada posição da proteína. Assim, cada barra no histograma representa uma posição de aminoácido e a porcentagem de sequências no grupo com mutações na posição.

Além da visualização dos grupos e da caracterização quanto aos seus padrões de mutações, outro problema proveniente do grande volume de sequências e da grande dimensionalidade dos dados foi a comparação da classificação obtida com a predição de resistência aos medicamentos proveniente da *look-up table* brasileira. Para resolver esse problema, outra imagem foi gerada a partir das análises realizadas pelo software HIVDAG (Araújo *et al.*, 2008).

O HIVDAG interpreta as regras na *look-up table* no contexto das sequências e produz predições relacionadas à resistência aos medicamentos. As regras da *look-up table* brasileira são criadas com base em estudos científicos que relacionam a ocorrência de uma mutação e a resistência a um dado medicamento. Assim, como novos estudos são constantemente realizados, todos os anos são atualizadas as regras, de modo a acompanhar a evolução do conhecimento. Nesse estudo foi empregada a versão 11 do algoritmo.

A tabela referencia os medicamentos ATV/R, DRV/R, FPV/R, IDV/R, LPV/R, SQV/R e TPV/R com a protease como alvo e os medicamentos 3TC, ABC, AZT, d4T, ddI, TDF, EFV, ETV, NVP com a transcriptase reversa como alvo. O software classifica as sequências como possuindo resistência (R), resistência intermediária (I) ou susceptibilidade (S) a cada um dos medicamentos.

Consequentemente, as sequências de protease receberam 7 rótulos de predição de resistência aos medicamentos (ATV/R, DRV/R, FPV/R, IDV/R, LPV/R, SQV/R e TPV/R) e as sequências de transcriptase reversa receberam 9 rótulos (3TC, ABC, AZT, d4T, ddI, TDF, EFV, ETV e NVP). Para representar os três possíveis resultados, as cores vermelha, amarela e verde representaram resistência, resistência intermediária e susceptibilidade, respectivamente. Portanto, como nas imagens binárias, as linhas representaram as sequências de proteína e as colunas as predições de resistência aos medicamentos geradas pela *look-up table*.

Por conseguinte, nas imagens coloridas, linhas verticais mostrando uma cor dominante em um grupo indicam que as sequências no grupo possuem a mesma predição de resistência aos medicamentos. Se os grupos apresentarem linhas verticais vermelhas, amarelas ou verdes nas diferentes colunas do grupo, há correspondência entre a predição da *look-up table* e os grupos do $K-$Médias.

3.7 Bimax

Para realizar a busca por padrões de mutações em sequências de protease e transcriptase reversa, o algoritmo Bimax do R Project for Statistical Computing Team (2008) foi aplicado a 10.229 sequências de protease e transcriptase reversa de subtipo B, 424 de subtipo C e 801 de subtipo F, separados por subtipo e proteína. Os parâmetros do Bimax foram ajustados para que buscasse por até 500 *biclusters*, sendo que nenhum dos conjuntos de dados atingiu tal número de *biclusters*, com mínimos de 2 a 10 colunas, uma vez que não houve convergência com mais de 10 colunas, repetindo a aplicação do algoritmo 100 vezes. O Bimax foi escolhido entre os algoritmos de *biclustering* por ser adequado a dados binários, gerar *biclusters* de inclusão máxima e contendo apenas valores 1, ou seja, *biclusters* com posições de sequências nas quais ocorrem mutações.

3.8 Caracterização dos *Biclusters*

A caracterização dos *biclusters* gerados pelo Bimax difere da caracterização dos grupos gerados pelo $K-$Médias porque o Bimax em sua busca pelos *biclusters* determina os atributos dos dados que são responsáveis pela formação do *bicluster*. Assim, o algoritmo recebe as sequências como entrada e devolve o número de *biclusters* encontrados, o seu tamanho, as sequências que os compõem e as posições das proteínas que os definem.

Portanto, tabelas foram criadas para apresentar o número de *biclusters* encontrados para cada subtipo de vírus e proteína. Também foram criadas tabelas descrevendo as posições que definem cada *bicluster* bem como a quantidade de sequências que abrangem.

Além das tabelas, foram gerados gráficos mostrando quais posições determinam a composição dos *biclusters*. Esse gráfico é disponibilizado pelo R Project for Statistical Computing Team (2008), no pacote *biclust* (http://cran.r-project.org/web/packages/biclust/index.html). No eixo x do gráfico são representados os *biclusters* encontrados e no eixo y as colunas da matriz de dados, ou as posições das proteínas, que os caracterizam. Esses gráficos possibilitam a visualização das posições que contribuem para a geração dos *biclusters*.

Já com intuito de relacioná-los com a predição de resistência aos medicamentos da *look-up table* brasileira, foram criadas imagens para cada *bicluster*. Como nas imagens geradas para caracterização dos grupos do $K-$Médias, as sequências foram representadas como linhas e as predições de resistência aos medicamentos como colunas. Também foram utilizadas as cores vermelha, amarela e verde para representar resistência, resistência intermediária e susceptibilidade, respectivamente.

Capítulo 4

Resultados e Discussão

4.1 $K-$Médias

Para os distintos valores de k, imagens em preto e branco foram criadas caracterizando os grupos de cada combinação de subtipo e proteína. As Figuras 4.1 a 4.6 representam o resultado do agrupamento para $k = 6$, subtipos B, C e F de protease e transcriptase reversa (as figuras com outros valores de k estão disponíveis em http://www.ime.usp.br/~mcintho/).

Na progressão dos valores de k, inicialmente os grupos foram divididos em um grupo de sequências com muitas mutações e outro com poucas. Então, com o aumento progressivo do valor de k, o grupo com muitas mutações foi sendo repetidamente dividido. Dentre os valores de k de 2 a 16, os grupos gerados com $k = 6$ representaram melhor o conhecimento atual dos padrões de mutações e relação entre mutações.

Como pode ser observado nas figuras, os grupos encontrados pelo $K-$Médias possuem diferentes padrões de mutações nas análises de ambas proteínas. O $K-$Médias foi capaz de gerar grupos de acordo com a presença de diferentes mutações, mostrando, portanto, que é possível obter uma classificação para sequências de proteína de HIV por meio de algoritmos de agrupamento, de acordo com a ocorrência simultânea de mutações. Pelas figuras também é possível verificar que os grupos gerados para as mesmas proteínas, mas diferentes subtipos possuem padrões de mutação parecidos, principalmente para transcriptase reversa.

Com intuito de caracterizar os grupos quanto aos padrões de mutações, histogramas como os mostrados nas Figuras 4.7 a 4.12 para $k = 6$ foram construídos (os histogramas com outros valores de k estão disponíveis em http://www.ime.usp.br/~mcintho/). As figuras mostram a porcentagem de ocorrência de mutações em cada posição de aminoácido em cada grupo.

Na caracterização dos padrões são importantes posições com altas frequências de mutações em um grupo e baixas frequências em outros, uma vez que possibilitam a distinção entre as sequências e grupos. Adicionalmente, as posições com altas frequências de mutações concomitantes em um grupo são aquelas cuja ocorrência pode estar relacionada.

Os diferentes padrões de mutações são apresentados nas figuras nas diferentes porcentagens de mutações presentes em cada posição de proteína para cada grupo de sequências. Pode-se observar que algumas posições são mais importantes para caracterização e descrição dos grupos, tais como as posições PR10, PR82 e PR90 da protease e RT67, RT70 e RT219 da transcriptase reversa.

Para comparar os grupos com a predição de resistência aos medicamentos dada pela *look-up table* brasileira, imagens foram criadas. As imagens dos grupos de protease, como nas Figuras 4.13 a 4.15 para $k = 6$ (as figuras com outros valores de k estão disponíveis em http://www.ime.usp.br/~mcintho/), mostram a divisão em grupos com sequências com predição de sensibilidade, alguns com predição de resistência intermediária e outros com predição de resistência à maioria dos medicamentos.

Os grupos de transcriptase reversa também apresentaram predição de resistência aos medicamentos NNRTI parecidas. Contudo, mostraram diferentes combinações de predições para os medicamentos NRTI, como pode ser observado nas Figuras 4.16 a 4.18 para $k = 6$. Diferentemente dos medicamentos que possuem como alvo a protease ou medicamentos NNRTI, os grupos possuem sequências com diferentes predições de resultados para o tratamento com medicamentos NRTI nos distintos grupos.

Assim, os padrões de mutação encontrados pelo algoritmo de agrupamento para sequências de protease foram capazes de produzir grupos apenas com padrões de predição de resistência ou susceptibilidade à maioria dos medicamentos, como pode ser observado nas figuras. O $K-$Médias separou as sequências em grupos com predição de susceptibilidade à maioria dos medicamentos inibidoras de protease (grupos B6.1, B6.4, B6.5, C6.2, C6.3, F6.4 e F6.6), grupos de predição de resistência intermediária (grupo B6.3, C6.1, C6.4, F6.1 e F6.2) e grupos de predição de resistência (grupo B6.2, B6.6, C6.5, C6.6, F6.3 e F6.5).

Já para as sequências de transcriptase reversa é possível observar que para os medicamentos EFV, ETV e NVP, que compõem o conjunto dos medicamentos chamados NNRTIs, não há grandes diferenças entre os grupos. No entanto, por exemplo, a maioria das sequências do grupo B6.2 do subtipo B, segundo a *look-up table* brasileira, deve ser resistente a 3TC, deve ter resistência intermediária a ABC e ddI e deve ser susceptível a AZT, d4T e TDF. Já a maioria das sequências do grupo B6.5 do subtipo B deve ser susceptível a 3TC, ABC, AZT, d4T, ddI e TDF. Logo ambos grupos representam diferentes padrões de predição. As posições de transcriptase reversa que determinam o agrupamento dessas sequências devem ser importantes preditoras de resposta a tratamento com medicamentos NRTI.

Os resultados para o agrupamento de sequências de subtipos C e F foram similares aos resultados para o subtipo B e estão resumidos nas Tabelas 4.1 e 4.2. As tabelas também representam uma síntese dos grupos e apresentam a informação essencial necessária para seu entendimento e comparação. Nessas tabelas foram apresentadas para cada grupo as posições nas quais pelo menos 50% das sequências possuem mutações.

Os grupos dos subtipos B, C e F são similares quanto às posições em cada grupo que possuem maiores frequências de mutações, excluindo posições cuja mutação ocorre mais frequentemente em um dado subtipo, nesse conjunto de sequências analisado. Por exemplo, as posições PR15, PR20, PR36, PR41, PR69, PR89 e PR93 do subtipo C em protease; PR15, PR35, PR36, PR41 e PR89 para subtipo F em protease; e posição RT211 para subtipo C e F em transcriptase reversa. Além disso, os conjuntos de dados de subtipos C e F eram muito menores do que o conjunto de dados de subtipo B e, portanto, podem não representar toda variabilidade da população desses subtipos. O subtipo C foi o que apresentou maior diferença na comparação dos grupos, mas a correspondência entre as posições que são mais importantes para a formação dos grupos ainda existiu.

Essa correspondência entre os grupos pode ser observada, por exemplo, nos grupos de protease:

- B6.2, C6.5 e F6.3 que possuem altas porcentagens de mutações nas posições PR10, PR54, PR82 e PR90 (como descrito em (Wu *et al.*, 2003; Yahi *et al.*, 1999))
- B6.3, C6.4 e F6.1, nas posições PR30 e PR88 (como descrito em (Deforche *et al.*, 2007; Gonzales *et al.*, 2003; Hoffman *et al.*, 2003; Liu *et al.*, 2008; Rhee *et al.*, 2004; Wu *et al.*, 2003))
- B6.1, B6.4, B6.5, C6.2, C6.3, F6.4 e F6.6 com poucas mutações e predição de susceptibilidade à maioria dos medicamentos.

Nos grupos de transcriptase reversa a correspondência é vista nos grupos:

- B6.3, B6.4, C6.5, C6.6 e F6.5 que possuem altas porcentagens de mutações nas posições RT67, RT70 e RT219 (como descrito em (Rhee *et al.*, 2004; Shafer *et al.*, 2000b; Sing *et al.*, 2005)), com predição de resistência a AZT, d4T ddI, susceptibilidade a TDF e resistência ou resistência intermediária a 3TC e ABC
- B6.6, C6.3 e F6.3 nas posições RT41, RT67 e RT210 (como descrito em (Yahi *et al.*, 1999)), com predição de resistência a AZT, d4T, ddI e TDF e resistência ou resistência intermediária a 3TC e ABC
- B6.2, C6.4 e F6.6 nas posições RT184, RT214, com predição de resistência a 3TC, susceptibilidade a AZT, d4T e TDF e resistência intermediária a ABC e ddI
- B6.1, C6.2 e F6.1 nas posições RT41, RT184 e RT215 (como descrito em (Gonzales *et al.*, 2003; Rhee *et al.*, 2004)), com predição de resistência a 3TC, ABC,AZT,d4T e ddI e susceptibilidade a TDF
- B6.5, C6.1, F6.2 e F6.4 com poucas mutações e predição de susceptibilidade à maioria dos medicamentos

Portanto, os grupos sugerem que mutações nas posições PR10, PR54, PR82 e PR90 e nas posições PR30 e PR88 da protease estão relacionadas e frequentemente ocorrem juntas. Também frequentemente ocorrem juntas mutações nas posições RT67, RT70 e RT219, nas posições RT41, RT67 e RT210, nas posições RT184 e RT214 e nas posições RT41, RT184 e RT215 da transcriptase reversa. Por conseguinte, esses padrões também observados em (Deforche *et al.*, 2007; Gonzales *et al.*, 2003; Hoffman *et al.*, 2003; Liu *et al.*, 2008; Rhee *et al.*, 2004; Shafer *et al.*, 2000b; Sing *et al.*, 2005; Wu *et al.*, 2003; Yahi *et al.*, 1999) são importantes quando se investiga genótipo e fenótipo (resistência aos medicamentos) e no desenvolvimento de novos medicamentos.

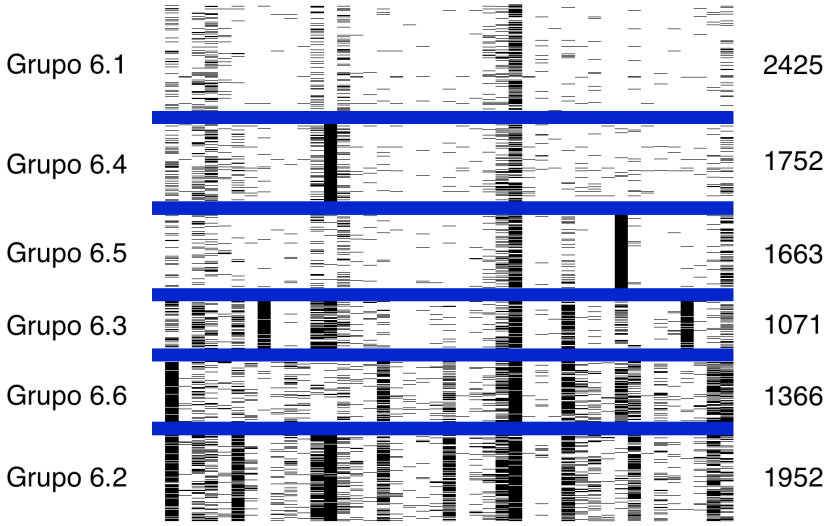

Figura 4.1: *Figura em preto e branco dos grupos para sequências de protease subtipo B. As colunas na figura representam as posições de aminoácido selecionadas e as linhas, as sequências de proteína. Os seis grupos são delimitados por linhas azuis.*

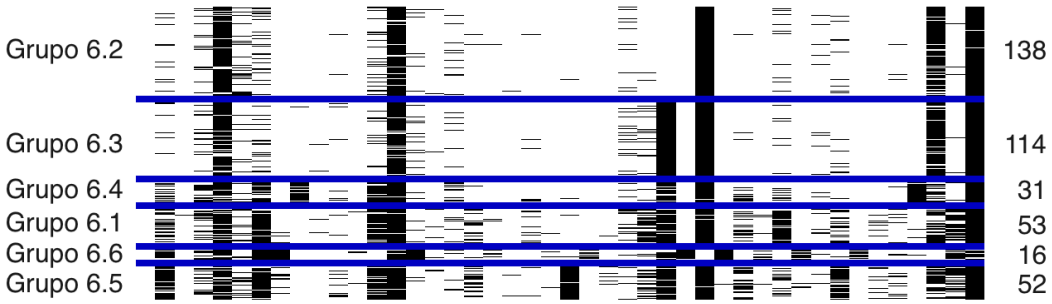

Figura 4.2: *Figura em preto e branco dos grupos para sequências de protease subtipo C. As colunas na figura representam as posições de aminoácido selecionadas e as linhas, as sequências de proteína. Os seis grupos são delimitados por linhas azuis.*

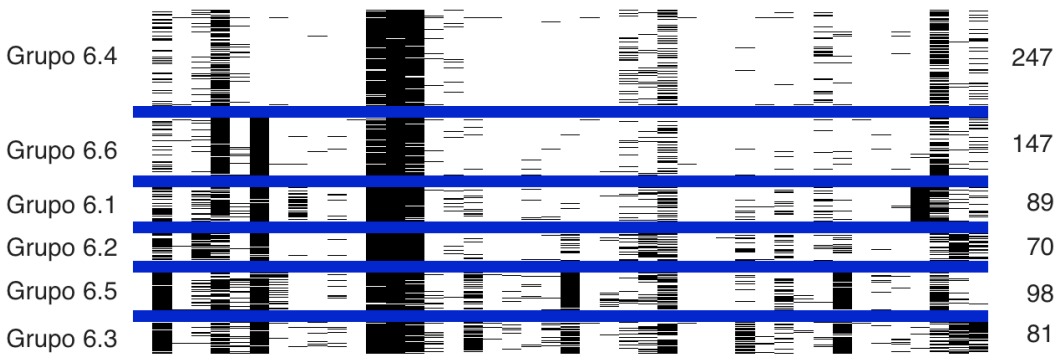

Figura 4.3: *Figura em preto e branco dos grupos para sequências de protease subtipo F. As colunas na figura representam as posições de aminoácido selecionadas e as linhas, as sequências de proteína. Os seis grupos são delimitados por linhas azuis.*

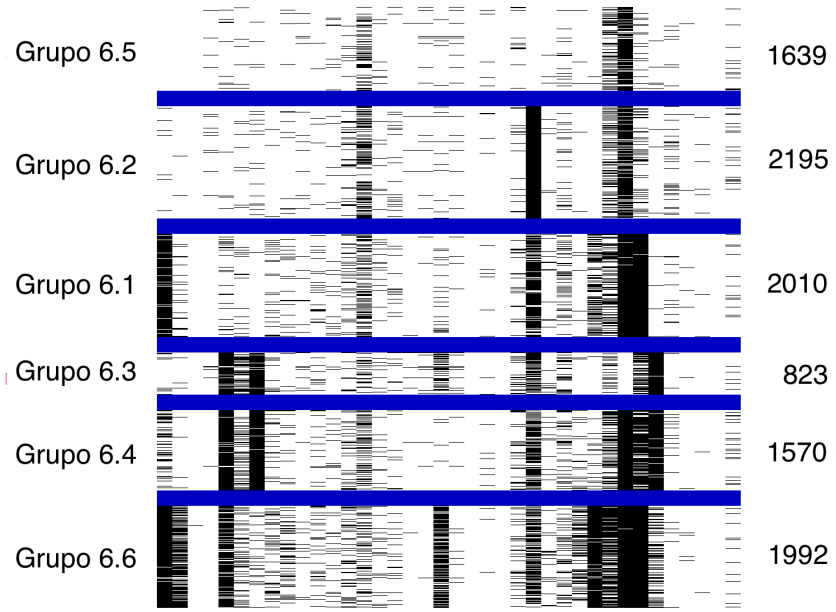

Figura 4.4: *Figura em preto e branco dos grupos para sequências de transcriptase reversa subtipo B. As colunas na figura representam as posições de aminoácido selecionadas e as linhas, as sequências de proteína. Os seis grupos são delimitados por linhas azuis.*

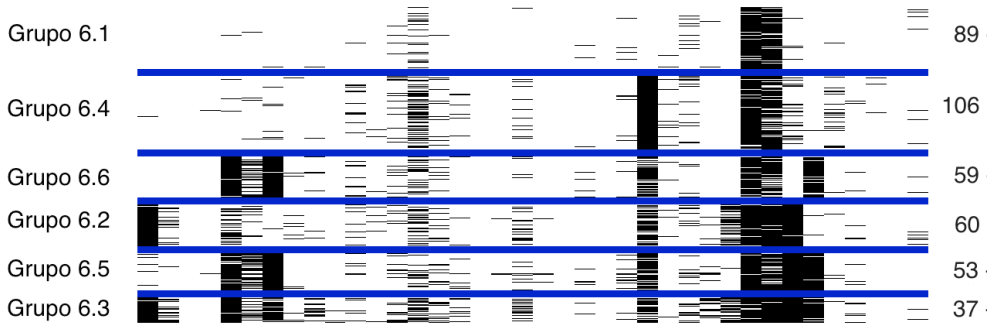

Figura 4.5: *Figura em preto e branco dos grupos para sequências de transcriptase reversa subtipo C. As colunas na figura representam as posições de aminoácido selecionadas e as linhas, as sequências de proteína. Os seis grupos são delimitados por linhas azuis.*

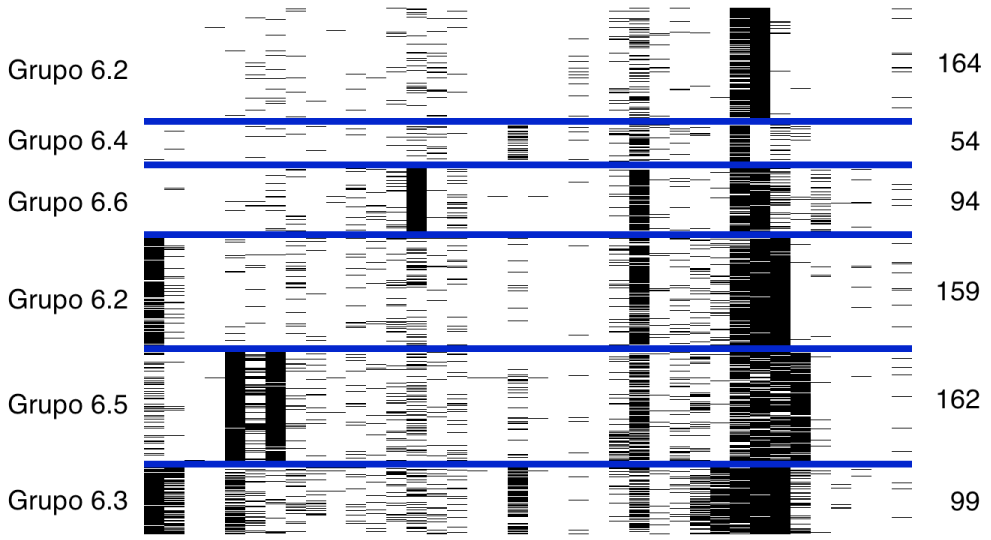

Figura 4.6: *Figura em preto e branco dos grupos para sequências de transcriptase reversa subtipo F. As colunas na figura representam as posições de aminoácido selecionadas e as linhas, as sequências de proteína. Os seis grupos são delimitados por linhas azuis.*

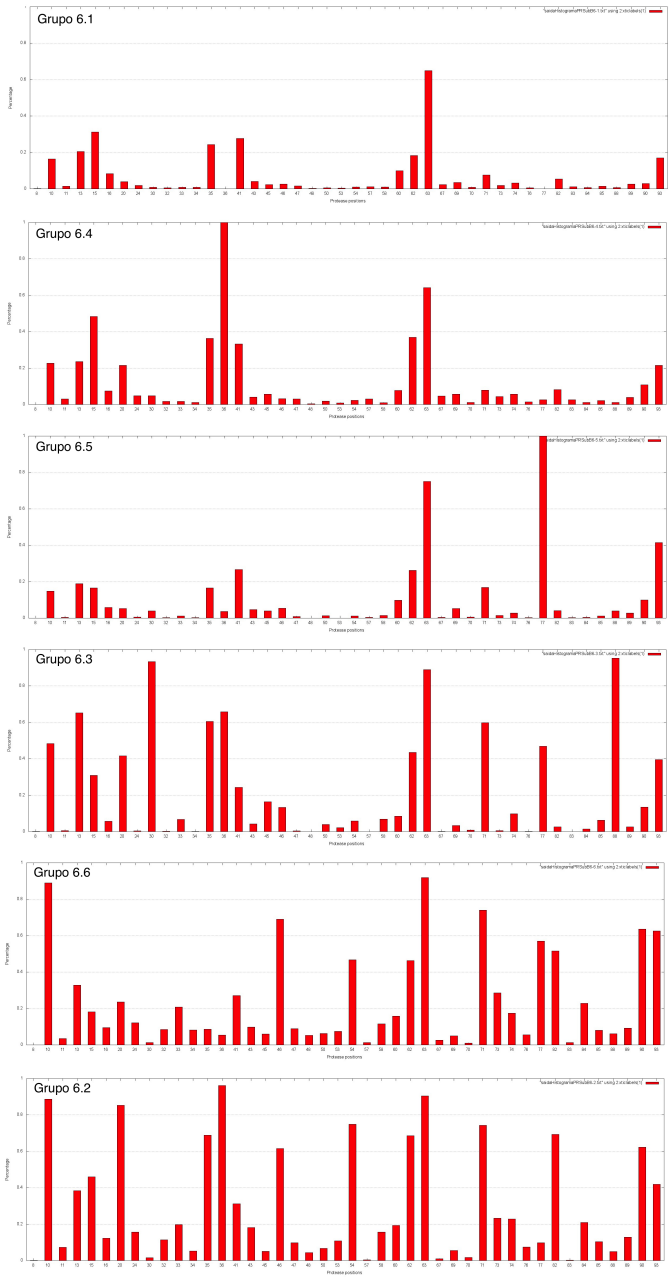

Figura 4.7: *Histogramas de frequência de mutações por grupos. Histogramas contendo as frequências de mutações para cada posição de aminoácido selecionada na protease em cada um dos seis grupos com sequências de subtipo B e* $k = 6$.

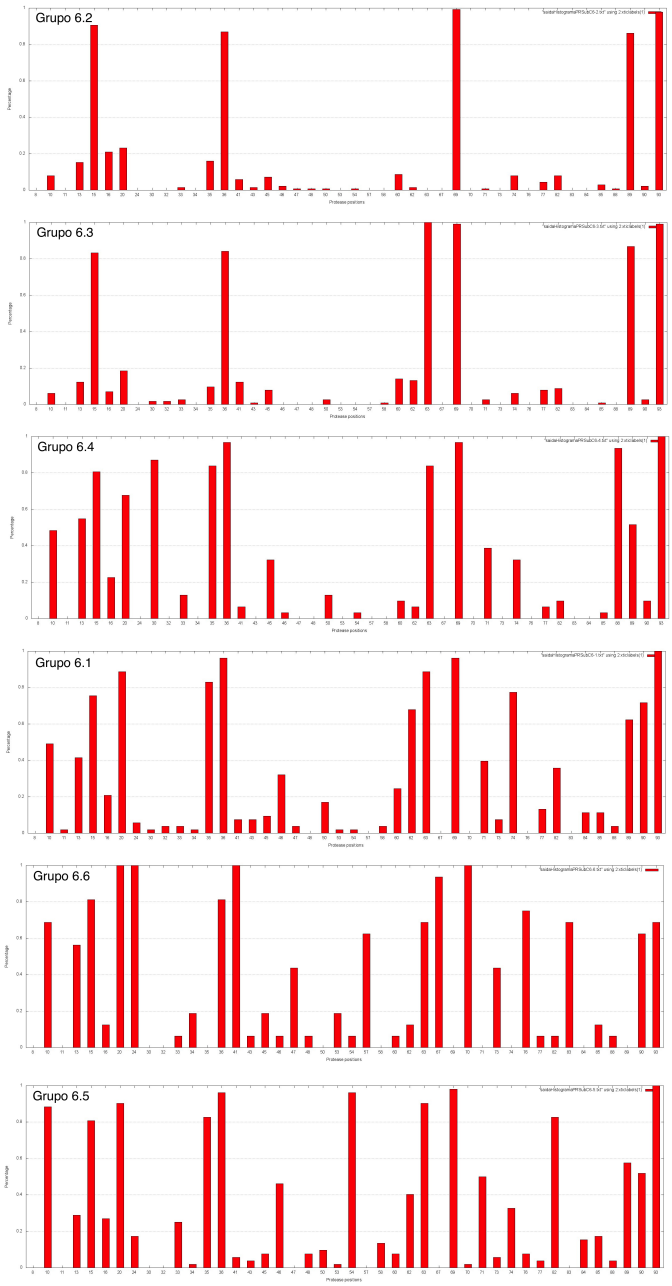

Figura 4.8: *Histogramas de frequência de mutações por grupos. Histogramas contendo as frequências de mutações para cada posição de aminoácido selecionada na protease em cada um dos seis grupos com sequências de subtipo C e $k = 6$.*

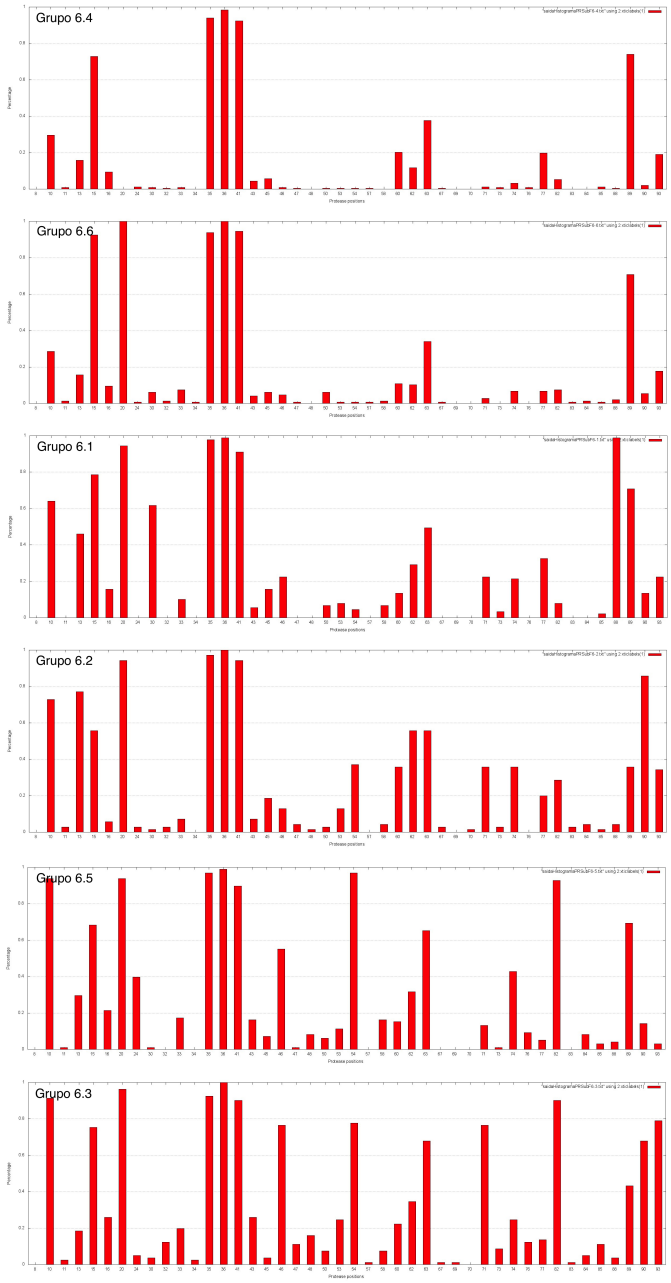

Figura 4.9: *Histogramas de frequência de mutações por grupos. Histogramas contendo as frequências de mutações para cada posição de aminoácido selecionada na protease em cada um dos seis grupos com sequências de subtipo F e $k = 6$.*

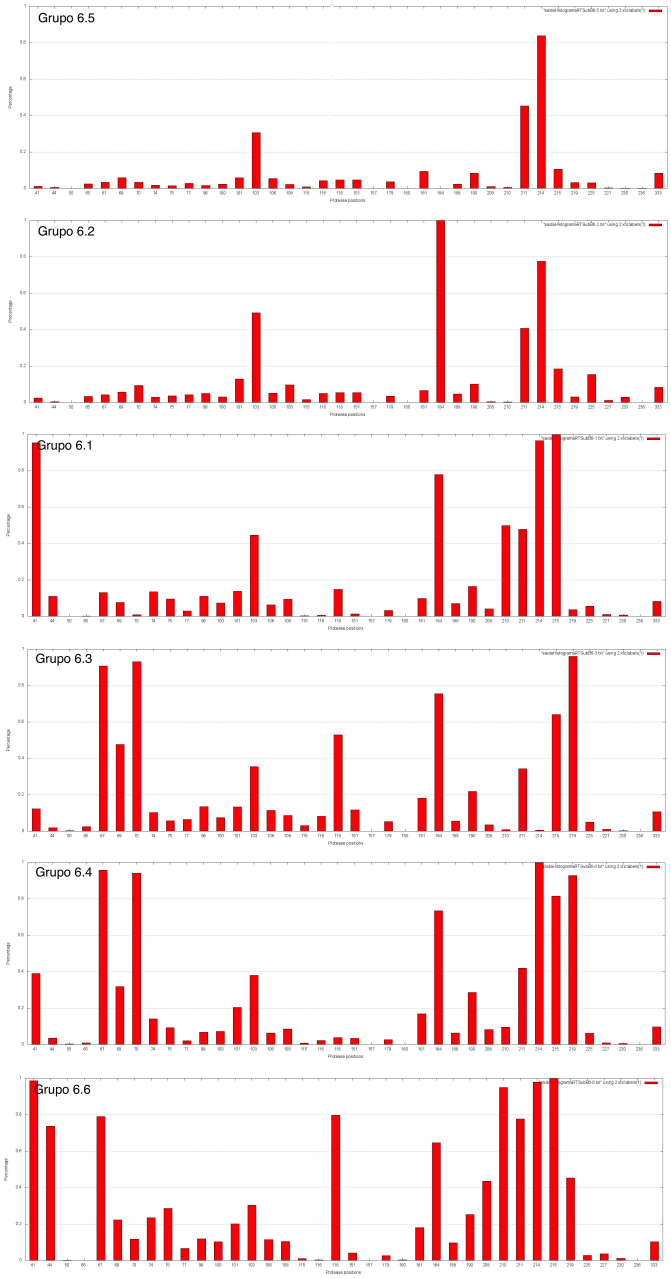

Figura 4.10: *Histogramas de frequência de mutações por grupos. Histogramas contendo as frequências de mutações para cada posição de aminoácido selecionada na transcriptase reversa em cada um dos seis grupos com sequências de subtipo B e $k = 6$.*

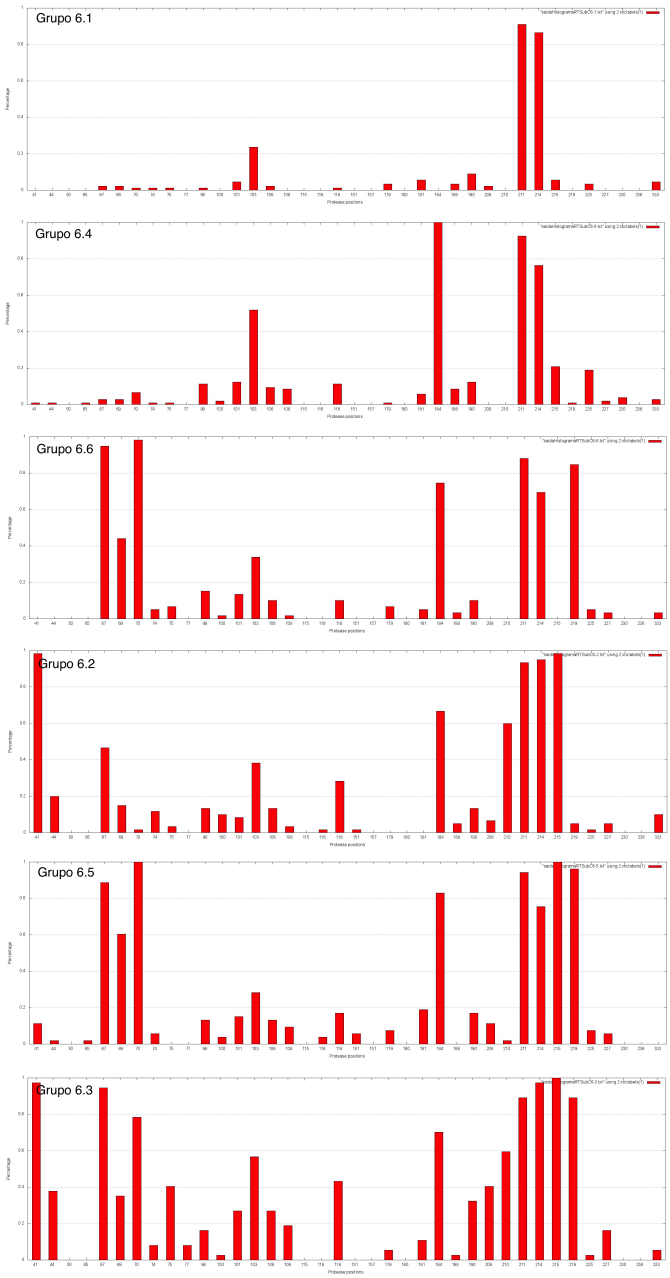

Figura 4.11: *Histogramas de frequência de mutações por grupos. Histogramas contendo as frequências de mutações para cada posição de aminoácido selecionada na transcriptase reversa em cada um dos seis grupos com sequências de subtipo C e $k = 6$.*

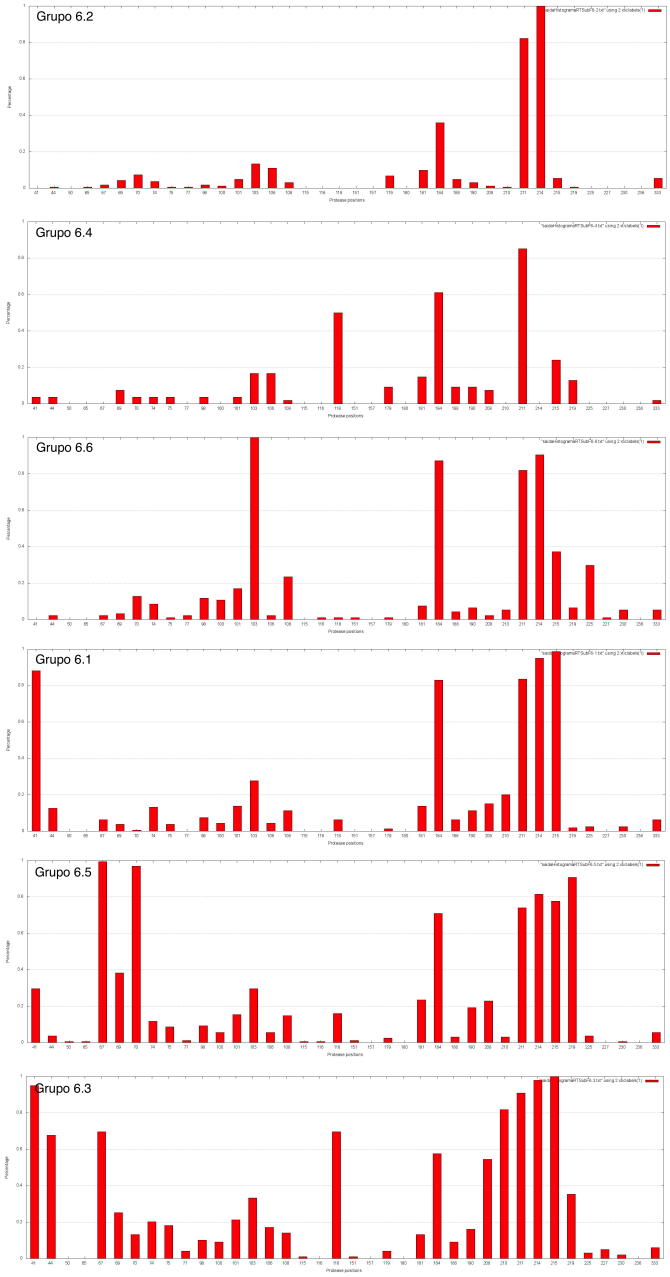

Figura 4.12: *Histogramas de frequência de mutações por grupos. Histogramas contendo as frequências de mutações para cada posição de aminoácido selecionada na transcriptase reversa em cada um dos seis grupos com sequências de subtipo F e $k = 6$.*

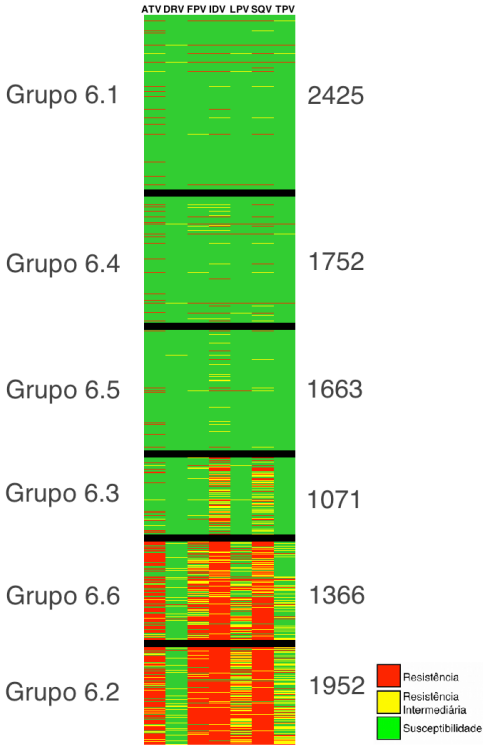

Figura 4.13: *Imagem colorida dos grupos para sequências de protease subtipo B. As colunas na Figura representam os nove medicamentos da look-up table brasileira (ATV/R, DRV/R, FPV/R, IDV/R, LPV/R, SQV/R and TPV/R, nessa ordem) e as linhas, as sequências de proteína. Os grupos são delimitados por linhas pretas.*

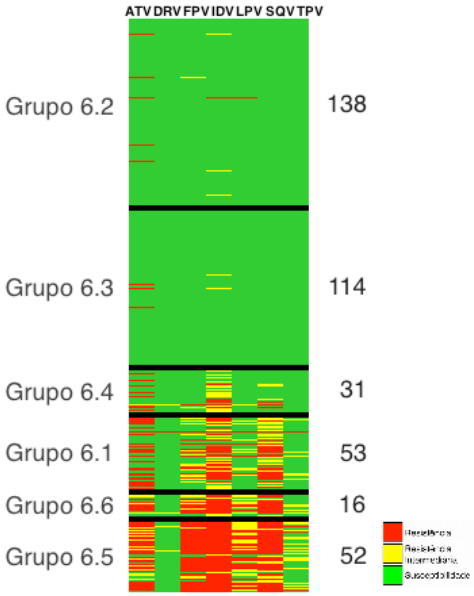

Figura 4.14: *Imagem colorida dos grupos para sequências de protease subtipo C. As colunas na Figura representam os nove medicamentos da look-up table brasileira (ATV/R, DRV/R, FPV/R, IDV/R, LPV/R, SQV/R and TPV/R, nessa ordem) e as linhas, as sequências de proteína. Os grupos são delimitados por linhas pretas.*

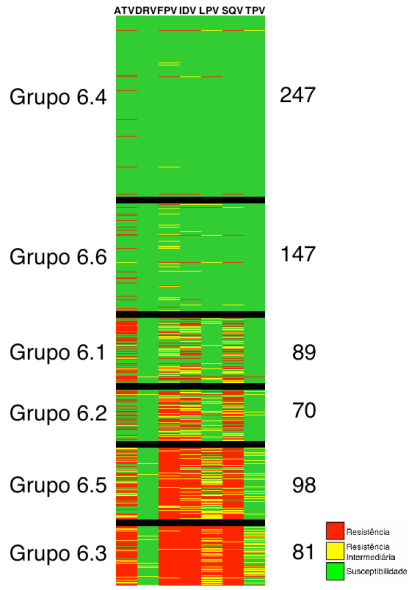

Figura 4.15: *Imagem colorida dos grupos para sequências de protease subtipo F. As colunas na Figura representam os nove medicamentos da look-up table brasileira (ATV/R, DRV/R, FPV/R, IDV/R, LPV/R, SQV/R and TPV/R, nessa ordem) e as linhas, as sequências de proteína. Os grupos são delimitados por linhas pretas.*

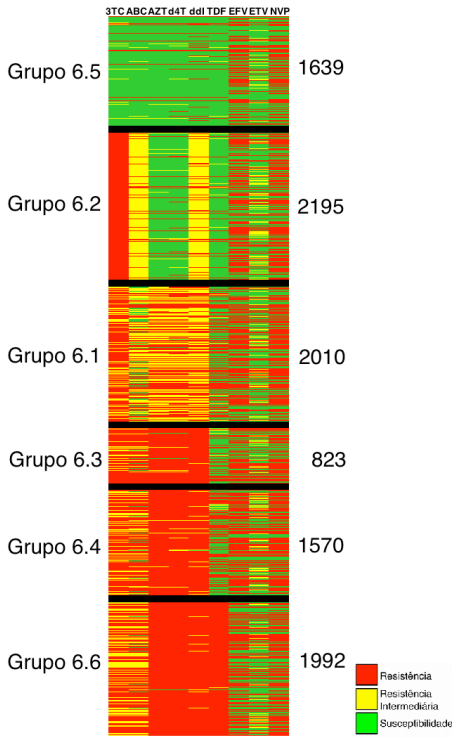

Figura 4.16: *Imagem colorida dos grupos para sequências de transcriptase reversa subtipo B. As colunas na Figura representam os nove medicamentos da look-up table brasileira (3TC, ABC, AZT, d4T, ddI, TDF, EFV, ETV and NVP, nessa ordem) e as linhas, as sequências de proteína. Os grupos são delimitados por linhas pretas.*

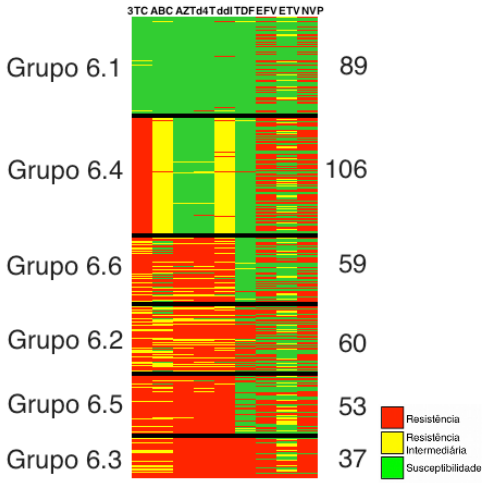

Figura 4.17: *Imagem colorida dos grupos para sequências de transcriptase reversa subtipo C. As colunas na Figura representam os nove medicamentos da look-up table brasileira (3TC, ABC, AZT, d4T, ddI, TDF, EFV, ETV and NVP, nessa ordem) e as linhas, as sequências de proteína. Os grupos são delimitados por linhas pretas.*

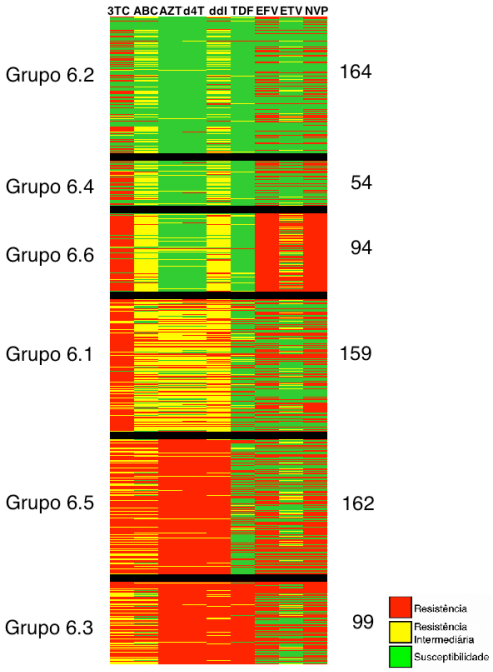

Figura 4.18: *Imagem colorida dos grupos para sequências de transcriptase reversa subtipo F. As colunas na Figura representam os nove medicamentos da look-up table brasileira (3TC, ABC, AZT, d4T, ddI, TDF, EFV, ETV and NVP, nessa ordem) e as linhas, as sequências de proteína. Os grupos são delimitados por linhas pretas.*

Tabela 4.1: *Posições da Protease com mutações em pelo menos 50% das sequências para cada grupo.*

		Posição da Protease																	
Grupo	Tamanho	10	13	15	20	30	35	36	41	46	54	62	63	71	82	88	89	90	93
Grupo B6.1	2425												X						
Grupo B6.2	1952	X			X		X	X		X	X	X	X	X	X			X	
Grupo B6.3	1071		X			X	X	X					X	X		X			
Grupo B6.4	1752							X					X						
Grupo B6.5	1663												X						
Grupo B6.6	1366	X								X			X	X	X			X	X
Grupo C6.1	53			X	X		X	X				X	X				X	X	
Grupo C6.2	138			X				X									X		X
Grupo C6.3	114			X				X									X		X
Grupo C6.4	31		X	X	X	X	X	X					X			X	X		
Grupo C6.5	52	X		X	X		X	X			X		X	X	X		X	X	
Grupo C6.6	16	X	X	X				X					X					X	X
Grupo F6.1	89	X		X	X	X	X	X	X							X	X		
Grupo F6.2	70	X	X	X	X		X		X			X	X					X	
Grupo F6.3	81	X		X	X		X		X	X	X		X	X	X			X	X
Grupo F6.4	247			X			X	X	X								X		
Grupo F6.5	98	X		X	X		X	X	X	X	X		X		X		X		
Grupo F6.6	147			X			X		X								X		

Tabela 4.2: *Posições da Transcriptase Reversa com mutações em pelo menos 50% das sequências para cada grupo.*

		Posição da Transcriptase Reversa										
Grupo	Tamanho	41	67	69	70	103	184	210	211	214	215	219
Grupo B6.1	2010	X					X			X	X	
Grupo B6.2	2195						X			X		
Grupo B6.3	823		X		X		X				X	X
Grupo B6.4	1570		X		X		X			X	X	X
Grupo B6.5	1639									X		
Grupo B6.6	1992	X	X				X	X	X	X	X	
Grupo C6.1	89								X	X		
Grupo C6.2	60	X					X	X	X	X	X	
Grupo C6.3	37	X	X		X	X	X	X	X	X	X	X
Grupo C6.4	106					X	X		X	X		
Grupo C6.5	53		X	X	X		X		X	X	X	X
Grupo C6.6	59		X		X		X		X	X		X
Grupo F6.1	159	X					X		X	X	X	
Grupo F6.2	164								X	X		
Grupo F6.3	99	X	X				X	X	X	X	X	
Grupo F6.4	54						X		X			
Grupo F6.5	162		X		X		X		X	X	X	X
Grupo F6.6	94					X	X		X	X		

4.2 Bimax

Para os distintos subtipos, proteínas e número mínimo de colunas, as Tabelas 4.3 a 4.5 mostram a quantidade de *biclusters* encontrados. Para a proteína transcriptase reversa de subtipo C, o Bimax encontrou *biclusters* com no máximo 8 posições de proteína, para transcriptase reversa subtipo F, com no máximo 9 posições e para as outras proteínas e valores de mínimo de colunas, o Bimax encontrou padrões de até 10 posições.

O Bimax foi capaz de encontrar um grande número de padrões de mutações para protease e transcriptase reversa quando comparado com trabalhos anteriores. O subtipo com maior número de *biclusters* foi o subtipo B, no entanto, isso pode apenas ser consequência do maior número de sequências usadas na análise e da maior representatividade das sequências com relação às variações circulantes no país.

Além do grande número de padrões de mutação encontrados, os *biclusters* são compostos por grandes números de posições de proteína. Como a maioria dos estudos que procuram por padrões de mutação em HIV realizam análises de correlação que buscam pares de posições correlacionadas (Alteri *et al.* (2009); Hoffman *et al.* (2003); Sing *et al.* (2005); Wu *et al.* (2003), como descrito na seção de trabalhos relacionados), a maioria dos padrões já reportados em trabalhos são de pares. Com a aplicação de métodos de reconhecimento de padrões, foi possível encontrar *biclusters* relacionando a ocorrência de 2 até 10 posições de proteína.

As Tabelas 4.6 a 4.14 apresentam os 30 maiores *biclusters* encontrados para cada valor mínimo de posições da proteína protease de subtipo B (as Tabelas com outros valores de mínimo de colunas e subtipos estão disponíveis em http://www.ime.usp.br/~mcintho/). O maior *bicluster* encontrado para a protease de subtipo B com mínimo de 2 colunas inclui 3857 sequências e o menor 181 sequências. Para subtipo B e mínimo de 10 colunas, o maior inclui 225 e o menor 10 sequências.

As Tabelas 4.15 a 4.23 apresentam os 30 maiores *biclusters* encontrados para transcriptase reversa de subtipo B para cada um dos valores mínimos de colunas (as Figuras com outros valores de mínimo de colunas e subtipo estão disponíveis em http://www.ime.usp.br/~mcintho/). O maior *bicluster* encontrado para a transcriptase reversa subtipo B com mínimo de 2 colunas inclui 5675 sequências e o menor 58 sequências. Para subtipo B e mínimo de 10 colunas o maior inclui 268 e o menor 10 sequências.

Dentre o *biclusters* encontrados estão alguns que já foram reportados em trabalhos anteriores como, por exemplo, os *biclusters* de protease:

- os *biclusters* número 63 (PR71 e PR90), 55 (PR10 e PR82), 45 (PR10 e PR90), e 23 (PR10 e PR46) reportados previamente por (Hoffman *et al.*, 2003; Wu *et al.*, 2003; Yahi *et al.*, 1999)
- os *biclusters* 11 (PR35 e PR36), 21 (PR10 e PR71), 12 (PR63 e PR90), 13 (PR36 e PR62), 17 (PR20 e PR36), 14 (PR10 e PR93), 46 (PR10 PR54), e 158 (PR77 e PR93) por (Hoffman *et al.*, 2003; Wu *et al.*, 2003)
- o *bicluster* 20 (PR54 e PR82) por (Yahi *et al.*, 1999)
- os *biclusters* 25 (PR71 e PR93), 69 (PR62 e PR71) por (Hoffman *et al.*, 2003)
- o *bicluster* 6 (PR63 e PR71) por (Wu *et al.*, 2003) e
- os *biclusters* com três posições de protease (PR15, PR20 e PR36); (PR20, PR36 e PR62); e (PR20, PR35 e PR36) foram reportados por (Wu *et al.*, 2003).

São exemplos de *biclusters* de transcriptase reversa já reportados em outros trabalhos:

- os *biclusters* 25 (RT210 e RT215) e 59 (RT41 e RT210) já foram citados por (Alteri *et al.*, 2009; Sing *et al.*, 2005; Yahi *et al.*, 1999)
- o *bicluster* 19 (RT70 e RT219) por (Sing *et al.*, 2005; Yahi *et al.*, 1999)
- o *bicluster* 20 (RT67 e RT 70) por (Gonzales *et al.*, 2003; Yahi *et al.*, 1999)
- o *bicluster* 3 (RT41 e RT215) por (Gonzales *et al.*, 2003; Sing *et al.*, 2005)
- os *biclusters* 64 (RT184 e RT21O) e 12 (RT67 e RT219) por (Yahi *et al.*, 1999)
- os *biclusters* 26, 32, 128 e 8 por (Gonzales *et al.*, 2003)
- os *biclusters* de transcriptase reversa de três posições (RT41, RT 210 e RT215); (RT41, RT184 e RT215); (RT67, RT70 e RT184) e de quatro posições (RT67, RT 70, RT184 e RT219); e (RT67, RT70, RT215 e RT219) já foram citados por (Gonzales *et al.*, 2003).

Com intuito de verificar a contribuição das distintas posições das proteínas com relação à definição dos *biclusters*, gráficos como os mostrados em 4.19 a 4.36 para subtipo B foram criados relacionando os *biclusters* e as posições que os definem. Os gráficos mostram que algumas das posições selecionadas de protease e transcriptase reversa não participam dos *biclusters*. As posições que não compõem *biclusters* são apresentadas nas Tabelas 4.24 a 4.29.

Pelas Figuras 4.19 a 4.27 e Tabelas 4.24 a 4.29 é possível verificar a participação das posições de proteína nos padrões de mutação. Apesar de algumas posições possuírem maior contribuição para a definição de *biclusters* que outras, apenas as posições de 50, 180 e 157 de protease não participam de nenhum *bicluster*

de sequências de nenhum subtipo e todas as posições de transcriptase reversa participam de algum *bicluster* de sequências de algum subtipo.

Para relacionar os *biclusters* gerados pelo Bimax com a predição da *look-up table* brasileira, foram geradas imagens como as das Figuras 4.37 a 4.53 para cada um dos *biclusters* encontrados. As Figuras 4.37 a 4.53 são exemplos de *biclusters* de sequências de subtipo B e mínimo de 2 colunas (as Figuras com outros valores de mínimo de colunas e subtipos estão disponíveis em http://www.ime.usp.br/~mcintho/).

A partir das figuras é possível observar que os *biclusters* apresentam diferentes predições de resistência aos medicamentos para protease e transcriptase reversa. Diferentemente dos resultados para o $K-$Médias, o Bimax foi capaz de encontrar padrões de mutação que gerassem distintas predições de resistência aos medicamentos PI e aos medicamentos NNRTI (como pode ser visto em 4.50), além das NRTI.

Por exemplo, comparando as Figuras 4.37 e 4.38, podemos notar que o *bicluster* PR10 PR46 abrange sequências que, segundo a *look-up table* brasileira, devem possuir resistência a alguns medicamentos como ATV/R, FPV/R e IDV/R, diferentemente do *bicluster* PR35 PR36 que abrange maior número de sequências com predição de susceptibilidade. Nos *biclusters* de transcriptase reversa, o *bicluster* RT184 RT214 possui sequências com predição de susceptibilidade a AZT, d4T e TDF, e o *bicluster* RT41 RT210 já possui sequências com predição de resistência a esses medicamentos.

Também é possível observar que alguns padrões de mutações distintos levam à predições parecidas. Por exemplo, os *biclusters* PR10 PR46 e PR10 PR54 que possuem predições de resistência a ATV, FPV, IDV e SQV e susceptibilidade a DRV e TPV. Essa similaridade pode ser decorrente do fato de que diferentes mutações ou combinações de mutações levam à resistência a um mesmo medicamento.

Tabela 4.3: *Número de biclusters Subtipo B encontrados para cada valor de número mínimo de colunas*

Colunas	*Biclusters* Protease
2	183
3	326
4	371
5	407
6	452
7	462
8	460
9	458
10	456

Colunas	*Biclusters* Transcriptase Reversa
2	140
3	202
4	272
5	281
6	332
7	384
8	393
9	412
10	379

Tabela 4.4: *Número de biclusters Subtipo C encontrados para cada valor de número mínimo de colunas*

Colunas	*Biclusters* Protease
2	54
3	67
4	113
5	163
6	209
7	244
8	226
9	213
10	103

Colunas	*Biclusters* Transcriptase Reversa
2	98
3	133
4	168
5	177
6	159
7	88
8	15

Tabela 4.5: *Número de biclusters Subtipo F encontrados para cada valor de número mínimo de colunas*

Colunas	*Biclusters* Protease
2	45
3	77
4	144
5	197
6	249
7	307
8	341
9	316
10	209

Colunas	*Biclusters* Transcriptase Reversa
2	107
3	143
4	205
5	223
6	247
7	189
8	84
9	16

Tabela 4.6: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 2 colunas*

Bicluster	Tamanho	Posições
Bicluster 1	3857	PR10 PR63
Bicluster 5	3585	PR36 PR63
Bicluster 6	3362	PR63 PR71
Bicluster 3	3325	PR62 PR63
Bicluster 4	3108	PR63 PR93
Bicluster 9	3081	PR35 PR63
Bicluster 8	2610	PR63 PR77
Bicluster 11	2584	PR35 PR36
Bicluster 7	2539	PR13 PR63
Bicluster 21	2534	PR10 PR71
Bicluster 39	2520	PR20 PR63
Bicluster 16	2466	PR10 PR36
Bicluster 12	2424	PR63 PR90
Bicluster 2	2390	PR15 PR63
Bicluster 13	2367	PR36 PR62
Bicluster 17	2347	PR20 PR36
Bicluster 144	2343	PR10 PR63 PR71
Bicluster 18	2307	PR10 PR62
Bicluster 66	2231	PR46 PR63
Bicluster 48	2230	PR35 PR36 PR63
Bicluster 24	2199	PR41 PR63
Bicluster 30	2129	PR63 PR82
Bicluster 111	2127	PR10 PR36 PR63
Bicluster 14	2119	PR10 PR93
Bicluster 32	2081	PR10 PR20
Bicluster 23	2062	PR10 PR46
Bicluster 128	2052	PR10 PR62 PR63
Bicluster 112	2034	PR36 PR71
Bicluster 46	2029	PR10 PR54
Bicluster 26	2018	PR54 PR63

Tabela 4.7: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 3 colunas*

Bicluster	Tamanho	Posições
Bicluster 29	2343	PR10 PR63 PR71
Bicluster 9	2230	PR35 PR36 PR63
Bicluster 8	2052	PR10 PR62 PR63
Bicluster 3	2003	PR36 PR62 PR63
Bicluster 35	1988	PR20 PR36 PR63
Bicluster 1	1884	PR10 PR63 PR93
Bicluster 24	1880	PR10 PR46 PR63
Bicluster 285	1861	PR36 PR63 PR71
Bicluster 6	1844	PR10 PR54 PR63
Bicluster 25	1823	PR10 PR63 PR90
Bicluster 32	1821	PR10 PR20 PR63
Bicluster 34	1768	PR62 PR63 PR71
Bicluster 7	1751	PR10 PR63 PR82
Bicluster 246	1705	PR63 PR71 PR93
Bicluster 69	1671	PR10 PR20 PR36
Bicluster 148	1669	PR20 PR63 PR71
Bicluster 21	1668	PR10 PR35 PR63
Bicluster 23	1586	PR20 PR35 PR36
Bicluster 127	1583	PR10 PR35 PR36
Bicluster 5	1582	PR10 PR54 PR82
Bicluster 87	1580	PR35 PR63 PR71
Bicluster 233	1542	PR46 PR63 PR71
Bicluster 17	1540	PR54 PR63 PR82
Bicluster 12	1539	PR35 PR62 PR63
Bicluster 180	1521	PR10 PR54 PR71
Bicluster 159	1514	PR10 PR36 PR62
Bicluster 14	1492	PR62 PR63 PR93
Bicluster 13	1469	PR10 PR13 PR63
Bicluster 26	1464	PR35 PR36 PR62
Bicluster 252	1463	PR10 PR36 PR71

Tabela 4.8: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 4 colunas*

Bicluster	Tamanho	Posições
Bicluster 6	1471	PR10 PR20 PR36 PR63
Bicluster 35	1443	PR10 PR54 PR63 PR82
Bicluster 4	1427	PR10 PR54 PR63 PR71
Bicluster 8	1419	PR10 PR35 PR36 PR63
Bicluster 2	1402	PR10 PR63 PR71 PR90
Bicluster 13	1387	PR20 PR35 PR36 PR63
Bicluster 24	1359	PR10 PR36 PR63 PR71
Bicluster 10	1356	PR10 PR46 PR63 PR71
Bicluster 5	1344	PR10 PR36 PR62 PR63
Bicluster 64	1339	PR10 PR62 PR63 PR71
Bicluster 63	1335	PR20 PR36 PR63 PR71
Bicluster 62	1326	PR35 PR36 PR63 PR71
Bicluster 34	1300	PR10 PR63 PR71 PR82
Bicluster 3	1294	PR35 PR36 PR62 PR63
Bicluster 323	1292	PR10 PR20 PR63 PR71
Bicluster 353	1255	PR10 PR36 PR54 PR63
Bicluster 11	1222	PR10 PR46 PR54 PR63
Bicluster 151	1219	PR20 PR36 PR62 PR63
Bicluster 71	1192	PR10 PR63 PR71 PR93
Bicluster 60	1181	PR10 PR46 PR63 PR82
Bicluster 85	1177	PR10 PR54 PR71 PR82
Bicluster 23	1172	PR36 PR62 PR63 PR71
Bicluster 36	1164	PR54 PR63 PR71 PR82
Bicluster 22	1145	PR10 PR20 PR35 PR36
Bicluster 286	1145	PR10 PR36 PR63 PR82
Bicluster 165	1136	PR10 PR20 PR36 PR71
Bicluster 186	1133	PR10 PR20 PR36 PR54
Bicluster 1	1126	PR10 PR62 PR63 PR90
Bicluster 40	1123	PR10 PR20 PR62 PR63
Bicluster 9	1102	PR10 PR46 PR63 PR90
Bicluster 73	1102	PR10 PR46 PR62 PR63

Tabela 4.9: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 5 colunas*

Bicluster	Tamanho	Posições
Bicluster 13	1105	PR10 PR54 PR63 PR71 PR82
Bicluster 6	1046	PR10 PR20 PR36 PR63 PR71
Bicluster 8	1031	PR10 PR20 PR35 PR36 PR63
Bicluster 18	1022	PR10 PR20 PR36 PR54 PR63
Bicluster 58	998	PR10 PR36 PR54 PR63 PR82
Bicluster 12	979	PR10 PR46 PR54 PR63 PR82
Bicluster 10	972	PR20 PR35 PR36 PR63 PR71
Bicluster 265	962	PR10 PR36 PR54 PR63 PR71
Bicluster 64	953	PR10 PR35 PR36 PR63 PR71
Bicluster 14	947	PR10 PR46 PR54 PR63 PR71
Bicluster 1	920	PR10 PR35 PR36 PR62 PR63
Bicluster 16	909	PR10 PR54 PR62 PR63 PR82
Bicluster 19	902	PR10 PR54 PR63 PR71 PR90
Bicluster 381	899	PR10 PR20 PR36 PR63 PR82
Bicluster 217	897	PR10 PR36 PR62 PR63 PR71
Bicluster 143	894	PR10 PR54 PR62 PR63 PR71
Bicluster 33	890	PR10 PR46 PR63 PR71 PR82
Bicluster 279	888	PR10 PR20 PR36 PR54 PR82
Bicluster 55	886	PR10 PR20 PR54 PR63 PR71
Bicluster 22	882	PR20 PR35 PR36 PR62 PR63
Bicluster 139	875	PR10 PR36 PR54 PR62 PR63
Bicluster 290	874	PR10 PR20 PR63 PR71 PR90
Bicluster 315	872	PR10 PR20 PR54 PR63 PR82
Bicluster 5	870	PR10 PR46 PR63 PR71 PR90
Bicluster 282	867	PR10 PR62 PR63 PR71 PR90
Bicluster 57	865	PR20 PR36 PR63 PR71 PR90
Bicluster 335	864	PR10 PR20 PR36 PR63 PR90
Bicluster 43	863	PR20 PR36 PR54 PR63 PR71
Bicluster 60	850	PR20 PR36 PR54 PR63 PR82
Bicluster 41	849	PR10 PR35 PR36 PR54 PR63

Tabela 4.10: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 6 colunas*

Bicluster	Tamanho	Posições
Bicluster 147	805	PR10 PR20 PR36 PR54 PR63 PR82
Bicluster 290	805	PR10 PR20 PR36 PR54 PR63 PR71
Bicluster 422	754	PR10 PR36 PR54 PR63 PR71 PR82
Bicluster 2	751	PR10 PR20 PR35 PR36 PR63 PR71
Bicluster 8	751	PR10 PR46 PR54 PR63 PR71 PR82
Bicluster 148	731	PR10 PR20 PR35 PR36 PR54 PR63
Bicluster 281	707	PR10 PR20 PR36 PR54 PR62 PR63
Bicluster 5	702	PR10 PR20 PR35 PR36 PR62 PR63
Bicluster 323	699	PR10 PR36 PR54 PR62 PR63 PR82
Bicluster 410	696	PR10 PR20 PR36 PR63 PR71 PR90
Bicluster 39	690	PR10 PR54 PR62 PR63 PR71 PR82
Bicluster 1	686	PR10 PR20 PR36 PR62 PR63 PR71
Bicluster 28	675	PR10 PR35 PR36 PR54 PR63 PR82
Bicluster 226	673	PR10 PR36 PR46 PR54 PR63 PR82
Bicluster 425	668	PR10 PR35 PR36 PR54 PR63 PR71
Bicluster 24	650	PR10 PR54 PR63 PR71 PR82 PR90
Bicluster 29	641	PR10 PR35 PR36 PR62 PR63 PR71
Bicluster 291	632	PR10 PR36 PR46 PR54 PR63 PR71
Bicluster 363	627	PR10 PR20 PR36 PR46 PR63 PR71
Bicluster 146	626	PR10 PR20 PR35 PR36 PR63 PR82
Bicluster 195	625	PR10 PR20 PR35 PR36 PR63 PR90
Bicluster 413	621	PR10 PR20 PR36 PR54 PR63 PR90
Bicluster 145	616	PR20 PR35 PR36 PR54 PR63 PR82
Bicluster 373	613	PR10 PR54 PR62 PR63 PR71 PR90
Bicluster 294	611	PR10 PR20 PR36 PR54 PR62 PR82
Bicluster 161	607	PR10 PR46 PR54 PR62 PR63 PR82
Bicluster 26	600	PR10 PR20 PR36 PR62 PR63 PR90
Bicluster 149	598	PR10 PR20 PR35 PR54 PR63 PR82
Bicluster 23	596	PR10 PR20 PR46 PR54 PR63 PR71
Bicluster 40	589	PR20 PR36 PR62 PR63 PR71 PR90

Tabela 4.11: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 7 colunas*

Bicluster	Tamanho	Posições
Bicluster 25	625	PR10 PR20 PR36 PR54 PR63 PR71 PR82
Bicluster 5	584	PR10 PR20 PR35 PR36 PR54 PR63 PR82
Bicluster 59	581	PR10 PR20 PR35 PR36 PR54 PR63 PR71
Bicluster 3	557	PR10 PR20 PR36 PR54 PR62 PR63 PR82
Bicluster 17	547	PR10 PR20 PR36 PR54 PR62 PR63 PR71
Bicluster 258	546	PR10 PR20 PR36 PR46 PR54 PR63 PR82
Bicluster 24	538	PR10 PR20 PR36 PR54 PR63 PR71 PR90
Bicluster 118	521	PR10 PR20 PR35 PR36 PR54 PR62 PR63
Bicluster 7	517	PR10 PR36 PR54 PR62 PR63 PR71 PR82
Bicluster 43	510	PR10 PR20 PR35 PR36 PR62 PR63 PR71
Bicluster 341	509	PR10 PR20 PR35 PR36 PR63 PR71 PR90
Bicluster 443	484	PR10 PR20 PR35 PR36 PR63 PR71 PR82
Bicluster 444	479	PR20 PR35 PR36 PR54 PR63 PR71 PR82
Bicluster 259	463	PR10 PR36 PR46 PR54 PR62 PR63 PR82
Bicluster 260	462	PR10 PR20 PR36 PR46 PR54 PR62 PR63
Bicluster 18	459	PR10 PR35 PR36 PR54 PR62 PR63 PR71
Bicluster 6	456	PR10 PR20 PR36 PR54 PR63 PR82 PR90
Bicluster 70	455	PR10 PR20 PR35 PR36 PR54 PR63 PR90
Bicluster 404	454	PR10 PR20 PR36 PR54 PR62 PR63 PR90
Bicluster 427	454	PR10 PR36 PR54 PR63 PR71 PR82 PR90
Bicluster 21	452	PR10 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 353	449	PR10 PR20 PR36 PR54 PR62 PR71 PR82
Bicluster 379	449	PR20 PR35 PR36 PR62 PR63 PR71 PR90
Bicluster 403	449	PR10 PR36 PR54 PR62 PR63 PR71 PR90
Bicluster 147	447	PR10 PR20 PR54 PR62 PR63 PR71 PR82
Bicluster 413	446	PR10 PR20 PR35 PR36 PR62 PR63 PR90
Bicluster 117	444	PR10 PR20 PR35 PR36 PR54 PR62 PR82
Bicluster 119	435	PR20 PR35 PR36 PR54 PR62 PR63 PR82
Bicluster 183	431	PR10 PR46 PR54 PR63 PR71 PR82 PR90
Bicluster 426	427	PR10 PR20 PR36 PR63 PR71 PR82 PR90

Tabela 4.12: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 8 colunas*

Bicluster	Tamanho	Posições
Bicluster 5	456	PR10 PR20 PR35 PR36 PR54 PR63 PR71 PR82
Bicluster 3	422	PR10 PR20 PR36 PR46 PR54 PR63 PR71 PR82
Bicluster 25	422	PR10 PR20 PR36 PR54 PR62 PR63 PR71 PR82
Bicluster 4	414	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR82
Bicluster 288	402	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR71
Bicluster 2	393	PR10 PR20 PR36 PR54 PR63 PR71 PR82 PR90
Bicluster 53	384	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR82
Bicluster 30	370	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR82
Bicluster 1	366	PR10 PR20 PR35 PR36 PR62 PR63 PR71 PR90
Bicluster 289	365	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR71
Bicluster 423	359	PR10 PR35 PR36 PR54 PR62 PR63 PR71 PR82
Bicluster 11	351	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71
Bicluster 63	346	PR10 PR20 PR36 PR46 PR54 PR63 PR71 PR90
Bicluster 107	344	PR10 PR36 PR46 PR54 PR62 PR63 PR71 PR82
Bicluster 333	343	PR10 PR20 PR35 PR36 PR54 PR63 PR82 PR90
Bicluster 428	341	PR10 PR35 PR36 PR46 PR54 PR63 PR71 PR82
Bicluster 332	340	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR90
Bicluster 31	339	PR10 PR20 PR36 PR54 PR62 PR63 PR82 PR90
Bicluster 75	322	PR10 PR35 PR36 PR54 PR62 PR63 PR71 PR90
Bicluster 280	309	PR10 PR20 PR35 PR36 PR54 PR62 PR71 PR90
Bicluster 24	305	PR10 PR35 PR36 PR46 PR54 PR62 PR63 PR82
Bicluster 125	302	PR10 PR20 PR35 PR46 PR54 PR63 PR71 PR82
Bicluster 126	301	PR10 PR20 PR35 PR54 PR63 PR71 PR82 PR90
Bicluster 109	300	PR10 PR20 PR36 PR46 PR62 PR63 PR71 PR82
Bicluster 15	299	PR10 PR20 PR36 PR46 PR54 PR62 PR71 PR82
Bicluster 32	299	PR10 PR20 PR36 PR54 PR63 PR71 PR82 PR93
Bicluster 7	298	PR10 PR20 PR36 PR46 PR54 PR63 PR82 PR90
Bicluster 65	298	PR10 PR36 PR46 PR54 PR63 PR71 PR82 PR90
Bicluster 108	297	PR10 PR20 PR46 PR54 PR62 PR63 PR71 PR82
Bicluster 170	297	PR10 PR46 PR54 PR62 PR63 PR71 PR82 PR90

Tabela 4.13: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 9 colunas*

Bicluster	Tamanho	Posições
Bicluster 8	315	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR71 PR82
Bicluster 2	297	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR71 PR82
Bicluster 16	296	PR10 PR20 PR35 PR36 PR54 PR63 PR71 PR82 PR90
Bicluster 5	295	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR71 PR90
Bicluster 34	293	PR10 PR20 PR36 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 35	279	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR82
Bicluster 10	268	PR10 PR20 PR35 PR36 PR46 PR54 PR62 PR63 PR82
Bicluster 412	259	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR82 PR90
Bicluster 9	253	PR10 PR20 PR36 PR46 PR54 PR63 PR71 PR82 PR90
Bicluster 36	247	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR90
Bicluster 96	243	PR10 PR35 PR36 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 275	236	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR71 PR90
Bicluster 95	234	PR10 PR20 PR35 PR36 PR54 PR62 PR71 PR82 PR90
Bicluster 401	232	PR10 PR20 PR35 PR36 PR54 PR63 PR71 PR82 PR93
Bicluster 1	220	PR10 PR36 PR46 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 400	217	PR10 PR20 PR36 PR46 PR54 PR63 PR71 PR82 PR93
Bicluster 31	212	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR82 PR90
Bicluster 421	203	PR10 PR20 PR36 PR54 PR62 PR63 PR71 PR82 PR93
Bicluster 37	201	PR10 PR20 PR46 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 145	197	PR10 PR20 PR35 PR36 PR62 PR63 PR71 PR90 PR93
Bicluster 301	197	PR10 PR20 PR35 PR36 PR46 PR62 PR63 PR71 PR90
Bicluster 419	195	PR10 PR20 PR36 PR54 PR63 PR71 PR82 PR90 PR93
Bicluster 211	194	PR10 PR20 PR35 PR36 PR54 PR63 PR71 PR90 PR93
Bicluster 339	189	PR10 PR15 PR20 PR35 PR36 PR54 PR63 PR71 PR82
Bicluster 413	189	PR10 PR15 PR20 PR36 PR54 PR62 PR63 PR71 PR82
Bicluster 29	188	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR71 PR93
Bicluster 212	187	PR10 PR15 PR20 PR35 PR36 PR54 PR63 PR71 PR90
Bicluster 85	186	PR10 PR15 PR20 PR36 PR46 PR54 PR63 PR71 PR90
Bicluster 163	186	PR20 PR35 PR36 PR46 PR54 PR63 PR71 PR82 PR90
Bicluster 447	183	PR10 PR35 PR36 PR46 PR54 PR63 PR71 PR82 PR93

Tabela 4.14: *Posições da Protease Subtipo B que definem os 30 maiores biclusters com mínimo de 10 colunas*

Bicluster	Tamanho	Posições
Bicluster 2	225	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 5	202	PR10 PR20 PR35 PR36 PR46 PR54 PR62 PR63 PR71 PR82
Bicluster 1	187	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 14	179	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR71 PR82 PR90
Bicluster 115	172	PR10 PR20 PR35 PR36 PR46 PR54 PR62 PR63 PR71 PR90
Bicluster 6	164	PR10 PR20 PR35 PR36 PR54 PR62 PR63 PR71 PR82 PR93
Bicluster 21	160	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR71 PR82 PR93
Bicluster 371	157	PR10 PR20 PR35 PR36 PR54 PR63 PR71 PR82 PR90 PR93
Bicluster 52	144	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR82 PR93
Bicluster 4	139	PR10 PR15 PR20 PR36 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 94	138	PR10 PR20 PR35 PR46 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 27	135	PR10 PR15 PR20 PR36 PR46 PR54 PR63 PR71 PR82 PR90
Bicluster 259	134	PR10 PR15 PR20 PR35 PR36 PR54 PR62 PR63 PR71 PR90
Bicluster 30	131	PR10 PR15 PR20 PR35 PR36 PR54 PR63 PR71 PR82 PR90
Bicluster 249	131	PR10 PR15 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR90
Bicluster 207	128	PR10 PR15 PR20 PR35 PR36 PR54 PR62 PR63 PR71 PR82
Bicluster 51	124	PR10 PR20 PR35 PR36 PR46 PR54 PR62 PR63 PR82 PR93
Bicluster 113	121	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR90 PR93
Bicluster 357	117	PR10 PR36 PR46 PR54 PR62 PR63 PR71 PR82 PR90 PR93
Bicluster 31	116	PR10 PR15 PR20 PR35 PR36 PR46 PR54 PR62 PR63 PR82
Bicluster 325	116	PR10 PR20 PR35 PR36 PR46 PR62 PR63 PR71 PR90 PR93
Bicluster 184	112	PR10 PR20 PR35 PR36 PR41 PR54 PR62 PR63 PR71 PR82
Bicluster 7	111	PR10 PR20 PR35 PR36 PR46 PR54 PR62 PR63 PR90 PR93
Bicluster 29	111	PR10 PR15 PR20 PR35 PR36 PR54 PR62 PR63 PR82 PR90
Bicluster 17	108	PR10 PR20 PR36 PR46 PR54 PR62 PR63 PR71 PR73 PR90
Bicluster 436	108	PR10 PR20 PR35 PR36 PR46 PR54 PR63 PR82 PR90 PR93
Bicluster 188	107	PR10 PR15 PR20 PR46 PR54 PR62 PR63 PR71 PR82 PR90
Bicluster 168	106	PR10 PR20 PR35 PR36 PR41 PR46 PR54 PR63 PR71 PR82
Bicluster 251	105	PR10 PR13 PR20 PR35 PR36 PR54 PR63 PR71 PR82 PR90
Bicluster 358	105	PR10 PR15 PR36 PR46 PR54 PR62 PR63 PR71 PR82 PR90

Tabela 4.15: *Posições da Transcriptase Reversa Subtipo B que definem os 30 maiores biclusters com mínimo de 2 colunas*

Bicluster	Tamanho	Posições
Bicluster 1	5675	RT214 RT215
Bicluster 2	5629	RT184 RT214
Bicluster 3	4574	RT41 RT215
Bicluster 8	4553	RT184 RT215
Bicluster 5	4457	RT211 RT214
Bicluster 4	4440	RT41 RT214
Bicluster 9	4363	RT41 RT214 RT215
Bicluster 15	4037	RT184 RT214 RT215
Bicluster 11	3578	RT67 RT215
Bicluster 16	3491	RT211 RT215
Bicluster 18	3455	RT67 RT214
Bicluster 128	3366	RT41 RT184
Bicluster 10	3346	RT184 RT211
Bicluster 6	3340	RT103 RT214
Bicluster 105	3232	RT211 RT214 RT215
Bicluster 138	3063	RT67 RT214 RT215
Bicluster 25	3052	RT210 RT215
Bicluster 24	3019	RT210 RT214
Bicluster 38	3007	RT210 RT214 RT215
Bicluster 29	2999	RT67 RT184
Bicluster 17	2923	RT184 RT211 RT214
Bicluster 59	2890	RT41 RT210
Bicluster 7	2852	RT103 RT184
Bicluster 12	2829	RT67 RT219
Bicluster 56	2756	RT41 RT211
Bicluster 14	2690	RT215 RT219
Bicluster 107	2662	RT41 RT211 RT214
Bicluster 23	2498	RT103 RT215
Bicluster 31	2487	RT214 RT219
Bicluster 110	2456	RT184 RT211 RT215

Tabela 4.16: *Posições da Transcriptase Reversa Subtipo B que definem os 30 maiores biclusters com mínimo de 3 colunas*

Bicluster	Tamanho	Posições
Bicluster 1	4363	RT41 RT214 RT215
Bicluster 2	4037	RT184 RT214 RT215
Bicluster 9	3295	RT41 RT184 RT215
Bicluster 8	3232	RT211 RT214 RT215
Bicluster 5	3200	RT41 RT184 RT214
Bicluster 19	3142	RT41 RT184 RT214 RT215
Bicluster 11	3063	RT67 RT214 RT215
Bicluster 6	3007	RT210 RT214 RT215
Bicluster 3	2923	RT184 RT211 RT214
Bicluster 15	2885	RT41 RT210 RT215
Bicluster 14	2849	RT41 RT210 RT214
Bicluster 24	2845	RT41 RT210 RT214 RT215
Bicluster 17	2717	RT41 RT211 RT215
Bicluster 18	2662	RT41 RT211 RT214
Bicluster 54	2631	RT41 RT211 RT214 RT215
Bicluster 21	2479	RT67 RT184 RT215
Bicluster 26	2463	RT41 RT67 RT215
Bicluster 20	2456	RT184 RT211 RT215
Bicluster 30	2408	RT67 RT184 RT214
Bicluster 4	2378	RT103 RT184 RT214
Bicluster 42	2364	RT41 RT67 RT214
Bicluster 159	2353	RT41 RT67 RT214 RT215
Bicluster 12	2328	RT67 RT215 RT219
Bicluster 46	2274	RT184 RT211 RT214 RT215
Bicluster 10	2236	RT103 RT214 RT215
Bicluster 44	2143	RT214 RT215 RT219
Bicluster 7	2141	RT67 RT70 RT219
Bicluster 91	2135	RT210 RT211 RT214
Bicluster 194	2103	RT67 RT214 RT219
Bicluster 173	2101	RT67 RT184 RT214 RT215

Tabela 4.17: *Posições da Transcriptase Reversa Subtipo B que definem os 30 maiores biclusters com mínimo de 4 colunas*

Bicluster	Tamanho	Posições
Bicluster 2	3142	RT41 RT184 RT214 RT215
Bicluster 3	2845	RT41 RT210 RT214 RT215
Bicluster 1	2631	RT41 RT211 RT214 RT215
Bicluster 5	2353	RT41 RT67 RT214 RT215
Bicluster 4	2274	RT184 RT211 RT214 RT215
Bicluster 18	2126	RT210 RT211 RT214 RT215
Bicluster 14	2101	RT67 RT184 RT214 RT215
Bicluster 247	2061	RT184 RT210 RT214 RT215
Bicluster 20	2056	RT41 RT210 RT211 RT215
Bicluster 24	2033	RT41 RT210 RT211 RT214
Bicluster 249	1988	RT41 RT184 RT210 RT215
Bicluster 248	1963	RT41 RT184 RT210 RT214
Bicluster 6	1916	RT41 RT184 RT211 RT215
Bicluster 7	1884	RT41 RT184 RT211 RT214
Bicluster 13	1843	RT41 RT118 RT214 RT215
Bicluster 12	1840	RT67 RT214 RT215 RT219
Bicluster 28	1826	RT67 RT211 RT214 RT215
Bicluster 51	1765	RT67 RT210 RT214 RT215
Bicluster 151	1709	RT41 RT67 RT210 RT215
Bicluster 195	1709	RT41 RT67 RT184 RT215
Bicluster 46	1702	RT118 RT210 RT214 RT215
Bicluster 117	1693	RT41 RT118 RT210 RT215
Bicluster 8	1687	RT67 RT184 RT215 RT219
Bicluster 31	1677	RT67 RT70 RT215 RT219
Bicluster 59	1675	RT41 RT44 RT214 RT215
Bicluster 10	1646	RT103 RT184 RT214 RT215
Bicluster 194	1631	RT41 RT67 RT184 RT214
Bicluster 23	1595	RT67 RT70 RT184 RT219
Bicluster 9	1593	RT41 RT103 RT214 RT215
Bicluster 69	1550	RT41 RT67 RT211 RT215

Tabela 4.18: *Posições da Transcriptase Reversa Subtipo B que definem os 30 maiores biclusters com mínimo de 5 colunas*

Bicluster	Tamanho	Posições
Bicluster 2	2029	RT41 RT210 RT211 RT214 RT215
Bicluster 3	1962	RT41 RT184 RT210 RT214 RT215
Bicluster 1	1860	RT41 RT184 RT211 RT214 RT215
Bicluster 7	1694	RT41 RT67 RT210 RT214 RT215
Bicluster 8	1673	RT41 RT118 RT210 RT214 RT215
Bicluster 12	1624	RT41 RT67 RT184 RT214 RT215
Bicluster 11	1509	RT41 RT67 RT211 RT214 RT215
Bicluster 9	1480	RT41 RT44 RT210 RT214 RT215
Bicluster 24	1454	RT184 RT210 RT211 RT214 RT215
Bicluster 41	1412	RT41 RT184 RT210 RT211 RT215
Bicluster 42	1395	RT41 RT184 RT210 RT211 RT214
Bicluster 14	1340	RT41 RT67 RT118 RT214 RT215
Bicluster 36	1331	RT41 RT118 RT211 RT214 RT215
Bicluster 13	1327	RT67 RT184 RT214 RT215 RT219
Bicluster 53	1284	RT67 RT210 RT211 RT214 RT215
Bicluster 32	1263	RT67 RT118 RT210 RT214 RT215
Bicluster 146	1261	RT41 RT118 RT210 RT211 RT215
Bicluster 40	1260	RT67 RT184 RT211 RT214 RT215
Bicluster 49	1256	RT41 RT67 RT118 RT210 RT215
Bicluster 221	1247	RT41 RT67 RT118 RT210 RT214
Bicluster 54	1238	RT41 RT44 RT211 RT214 RT215
Bicluster 10	1228	RT41 RT67 RT214 RT215 RT219
Bicluster 5	1215	RT67 RT70 RT184 RT215 RT219
Bicluster 4	1209	RT67 RT70 RT214 RT215 RT219
Bicluster 26	1190	RT67 RT184 RT210 RT214 RT215
Bicluster 47	1188	RT41 RT118 RT184 RT214 RT215
Bicluster 6	1183	RT41 RT103 RT184 RT214 RT215
Bicluster 34	1158	RT41 RT44 RT184 RT214 RT215
Bicluster 48	1155	RT41 RT67 RT184 RT210 RT215
Bicluster 268	1148	RT41 RT44 RT210 RT211 RT215

Tabela 4.19: *Posições da Transcriptase Reversa Subtipo B que definem os 30 maiores biclusters com mínimo de 6 colunas*

Bicluster	Tamanho	Posições
Bicluster 5	1394	RT41 RT184 RT210 RT211 RT214 RT215
Bicluster 1	1246	RT41 RT67 RT118 RT210 RT214 RT215
Bicluster 4	1246	RT41 RT118 RT210 RT211 RT214 RT215
Bicluster 11	1241	RT41 RT67 RT210 RT211 RT214 RT215
Bicluster 6	1143	RT41 RT67 RT184 RT210 RT214 RT215
Bicluster 15	1132	RT41 RT44 RT210 RT211 RT214 RT215
Bicluster 12	1089	RT41 RT118 RT184 RT210 RT214 RT215
Bicluster 7	1081	RT41 RT44 RT118 RT210 RT214 RT215
Bicluster 26	1040	RT41 RT67 RT184 RT211 RT214 RT215
Bicluster 2	1039	RT41 RT44 RT67 RT210 RT214 RT215
Bicluster 10	1019	RT41 RT44 RT184 RT210 RT214 RT215
Bicluster 198	991	RT41 RT67 RT118 RT211 RT214 RT215
Bicluster 27	948	RT41 RT67 RT118 RT210 RT211 RT215
Bicluster 9	907	RT41 RT44 RT67 RT118 RT214 RT215
Bicluster 14	903	RT41 RT67 RT184 RT214 RT215 RT219
Bicluster 3	870	RT67 RT70 RT184 RT214 RT215 RT219
Bicluster 33	869	RT41 RT44 RT67 RT118 RT210 RT215
Bicluster 32	868	RT44 RT67 RT118 RT210 RT214 RT215
Bicluster 44	866	RT41 RT118 RT184 RT211 RT214 RT215
Bicluster 244	864	RT41 RT44 RT118 RT211 RT214 RT215
Bicluster 18	853	RT41 RT67 RT118 RT184 RT214 RT215
Bicluster 49	842	RT41 RT44 RT184 RT211 RT214 RT215
Bicluster 241	839	RT41 RT44 RT118 RT210 RT211 RT215
Bicluster 240	833	RT44 RT118 RT210 RT211 RT214 RT215
Bicluster 245	832	RT41 RT44 RT118 RT210 RT211 RT214
Bicluster 16	799	RT41 RT44 RT67 RT184 RT214 RT215
Bicluster 23	774	RT41 RT44 RT118 RT184 RT214 RT215
Bicluster 62	769	RT41 RT67 RT210 RT214 RT215 RT219
Bicluster 30	755	RT41 RT67 RT211 RT214 RT215 RT219
Bicluster 71	747	RT41 RT208 RT210 RT211 RT214 RT215

Tabela 4.20: *Posições da Transcriptase Reversa Subtipo B que definem os 384 biclusters com mínimo de 7 colunas*

Bicluster	Tamanho	Posições
Bicluster 5	941	RT41 RT67 RT118 RT210 RT211 RT214 RT215
Bicluster 4	866	RT41 RT44 RT67 RT118 RT210 RT214 RT215
Bicluster 11	842	RT41 RT67 RT184 RT210 RT211 RT214 RT215
Bicluster 13	831	RT41 RT44 RT118 RT210 RT211 RT214 RT215
Bicluster 1	816	RT41 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 3	807	RT41 RT67 RT118 RT184 RT210 RT214 RT215
Bicluster 16	797	RT41 RT44 RT67 RT210 RT211 RT214 RT215
Bicluster 6	773	RT41 RT44 RT184 RT210 RT211 RT214 RT215
Bicluster 7	736	RT41 RT44 RT118 RT184 RT210 RT214 RT215
Bicluster 24	726	RT41 RT44 RT67 RT184 RT210 RT214 RT215
Bicluster 133	696	RT41 RT44 RT67 RT118 RT211 RT214 RT215
Bicluster 132	677	RT41 RT44 RT67 RT118 RT210 RT211 RT215
Bicluster 257	619	RT67 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 80	591	RT41 RT44 RT67 RT184 RT211 RT214 RT215
Bicluster 33	582	RT41 RT67 RT208 RT210 RT211 RT214 RT215
Bicluster 327	573	RT41 RT44 RT118 RT184 RT210 RT211 RT215
Bicluster 15	568	RT41 RT67 RT118 RT210 RT214 RT215 RT219
Bicluster 326	567	RT41 RT44 RT118 RT184 RT210 RT211 RT214
Bicluster 66	566	RT41 RT67 RT210 RT211 RT214 RT215 RT219
Bicluster 265	563	RT44 RT67 RT184 RT210 RT211 RT214 RT215
Bicluster 12	562	RT41 RT44 RT208 RT210 RT211 RT214 RT215
Bicluster 8	558	RT41 RT67 RT184 RT211 RT214 RT215 RT219
Bicluster 128	558	RT41 RT67 RT184 RT210 RT214 RT215 RT219
Bicluster 218	544	RT41 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 10	506	RT41 RT103 RT184 RT210 RT211 RT214 RT215
Bicluster 277	505	RT41 RT44 RT210 RT211 RT214 RT215 RT219
Bicluster 198	495	RT41 RT67 RT118 RT208 RT211 RT214 RT215
Bicluster 30	485	RT41 RT44 RT118 RT210 RT214 RT215 RT219
Bicluster 156	474	RT41 RT44 RT67 RT208 RT211 RT214 RT215
Bicluster 2	473	RT41 RT67 RT70 RT184 RT214 RT215 RT219

Tabela 4.21: *Posições da Transcriptase Reversa Subtipo B que definem os 393 biclusters com mínimo de 8 colunas*

Bicluster	Tamanho	Posições
Bicluster 1	674	RT41 RT44 RT67 RT118 RT210 RT211 RT214 RT215
Bicluster 9	613	RT41 RT67 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 11	591	RT41 RT44 RT67 RT118 RT184 RT210 RT214 RT215
Bicluster 8	566	RT41 RT44 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 10	558	RT41 RT44 RT67 RT184 RT210 RT211 RT214 RT215
Bicluster 364	472	RT41 RT44 RT67 RT118 RT184 RT211 RT214 RT215
Bicluster 7	470	RT41 RT67 RT118 RT208 RT210 RT211 RT214 RT215
Bicluster 2	449	RT41 RT44 RT67 RT208 RT210 RT211 RT214 RT215
Bicluster 24	448	RT41 RT44 RT67 RT118 RT208 RT210 RT214 RT215
Bicluster 32	448	RT41 RT44 RT118 RT208 RT210 RT211 RT214 RT215
Bicluster 4	437	RT41 RT67 RT118 RT210 RT211 RT214 RT215 RT219
Bicluster 3	418	RT41 RT67 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 27	417	RT41 RT44 RT67 RT118 RT210 RT214 RT215 RT219
Bicluster 6	411	RT41 RT67 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 23	408	RT41 RT44 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 28	394	RT41 RT67 RT118 RT184 RT210 RT214 RT215 RT219
Bicluster 60	393	RT41 RT118 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 5	382	RT41 RT44 RT118 RT210 RT211 RT214 RT215 RT219
Bicluster 334	382	RT41 RT67 RT118 RT184 RT208 RT210 RT214 RT215
Bicluster 49	379	RT41 RT44 RT67 RT210 RT211 RT214 RT215 RT219
Bicluster 174	366	RT41 RT44 RT67 RT184 RT208 RT210 RT214 RT215
Bicluster 189	366	RT41 RT44 RT118 RT184 RT208 RT210 RT214 RT215
Bicluster 151	344	RT41 RT44 RT67 RT118 RT211 RT214 RT215 RT219
Bicluster 73	342	RT41 RT44 RT67 RT184 RT210 RT214 RT215 RT219
Bicluster 265	341	RT41 RT118 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 382	335	RT41 RT44 RT67 RT118 RT210 RT211 RT215 RT219
Bicluster 264	333	RT41 RT44 RT118 RT184 RT210 RT214 RT215 RT219
Bicluster 263	318	RT41 RT44 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 190	316	RT41 RT44 RT118 RT184 RT208 RT210 RT211 RT214
Bicluster 89	305	RT41 RT67 RT208 RT210 RT211 RT214 RT215 RT219

Tabela 4.22: *Posições da Transcriptase Reversa Subtipo B que definem os 412 biclusters com mínimo de 9 colunas*

Bicluster	Tamanho	Posições
Bicluster 1	460	RT41 RT44 RT67 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 2	387	RT41 RT44 RT67 RT118 RT208 RT210 RT211 RT214 RT215
Bicluster 8	333	RT41 RT44 RT67 RT118 RT210 RT211 RT214 RT215 RT219
Bicluster 6	326	RT41 RT67 RT118 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 3	318	RT41 RT44 RT67 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 13	315	RT41 RT44 RT118 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 215	311	RT41 RT44 RT67 RT118 RT184 RT208 RT210 RT214 RT215
Bicluster 7	301	RT41 RT67 RT118 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 24	289	RT41 RT44 RT67 RT118 RT184 RT210 RT214 RT215 RT219
Bicluster 77	274	RT41 RT44 RT67 RT118 RT184 RT208 RT211 RT214 RT215
Bicluster 78	270	RT41 RT44 RT67 RT118 RT184 RT208 RT210 RT211 RT215
Bicluster 12	269	RT41 RT44 RT67 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 167	257	RT41 RT44 RT118 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 291	257	RT41 RT67 RT118 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 208	239	RT41 RT44 RT67 RT118 RT208 RT210 RT214 RT215 RT219
Bicluster 201	232	RT41 RT44 RT118 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 225	223	RT41 RT67 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 87	213	RT41 RT44 RT67 RT118 RT208 RT211 RT214 RT215 RT219
Bicluster 9	207	RT41 RT44 RT67 RT75 RT118 RT210 RT211 RT214 RT215
Bicluster 115	203	RT41 RT118 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 195	196	RT41 RT44 RT67 RT103 RT118 RT210 RT211 RT214 RT215
Bicluster 65	191	RT41 RT44 RT67 RT118 RT190 RT210 RT211 RT214 RT215
Bicluster 285	190	RT41 RT44 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 50	187	RT41 RT67 RT75 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 166	186	RT41 RT44 RT118 RT184 RT208 RT210 RT214 RT215 RT219
Bicluster 118	184	RT41 RT67 RT118 RT184 RT208 RT210 RT211 RT215 RT219
Bicluster 139	183	RT67 RT118 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 194	183	RT41 RT67 RT103 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 30	176	RT41 RT44 RT103 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 100	174	RT41 RT44 RT67 RT103 RT184 RT210 RT211 RT214 RT215

Tabela 4.23: *Posições da Transcriptase Reversa Subtipo B que definem os 379 biclusters com mínimo de 10 colunas*

Bicluster	Tamanho	Posições
Bicluster 1	268	RT41 RT44 RT67 RT118 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 18	230	RT41 RT44 RT67 RT118 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 3	205	RT41 RT44 RT67 RT118 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 17	181	RT41 RT67 RT118 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 268	167	RT41 RT44 RT67 RT118 RT184 RT208 RT210 RT214 RT215 RT219
Bicluster 305	160	RT41 RT44 RT67 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 79	158	RT41 RT44 RT118 RT184 RT208 RT210 RT211 RT214 RT215 RT219
Bicluster 306	147	RT41 RT44 RT67 RT118 RT184 RT208 RT211 RT214 RT215 RT219
Bicluster 2	141	RT41 RT44 RT67 RT75 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 6	140	RT41 RT44 RT67 RT103 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 331	131	RT41 RT44 RT67 RT118 RT184 RT190 RT210 RT211 RT214 RT215
Bicluster 5	130	RT41 RT44 RT67 RT74 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 133	122	RT41 RT44 RT67 RT103 RT118 RT210 RT211 RT214 RT215 RT219
Bicluster 53	120	RT41 RT44 RT67 RT75 RT118 RT208 RT210 RT211 RT214 RT215
Bicluster 20	117	RT41 RT44 RT67 RT75 RT118 RT210 RT211 RT214 RT215 RT219
Bicluster 27	115	RT41 RT44 RT67 RT103 RT118 RT184 RT210 RT214 RT215 RT219
Bicluster 59	105	RT41 RT67 RT101 RT118 RT184 RT190 RT210 RT211 RT214 RT215
Bicluster 253	105	RT41 RT44 RT67 RT74 RT118 RT210 RT211 RT214 RT215 RT219
Bicluster 336	104	RT41 RT44 RT67 RT103 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 237	102	RT41 RT44 RT67 RT74 RT118 RT184 RT210 RT214 RT215 RT219
Bicluster 236	101	RT41 RT67 RT74 RT118 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 254	101	RT41 RT67 RT75 RT118 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 355	101	RT41 RT44 RT67 RT101 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 379	99	RT41 RT67 RT118 RT184 RT190 RT210 RT211 RT214 RT215 RT219
Bicluster 138	98	RT41 RT44 RT67 RT118 RT184 RT190 RT210 RT214 RT215 RT219
Bicluster 4	96	RT41 RT44 RT75 RT118 RT184 RT208 RT210 RT211 RT214 RT215
Bicluster 19	94	RT41 RT44 RT67 RT69 RT118 RT184 RT210 RT211 RT214 RT215
Bicluster 11	93	RT41 RT44 RT74 RT118 RT184 RT210 RT211 RT214 RT215 RT219
Bicluster 12	93	RT41 RT44 RT67 RT75 RT118 RT184 RT208 RT210 RT214 RT215
Bicluster 73	92	RT41 RT44 RT101 RT118 RT184 RT190 RT210 RT211 RT214 RT215
Bicluster 77	92	RT41 RT44 RT67 RT75 RT118 RT184 RT210 RT214 RT215 RT219

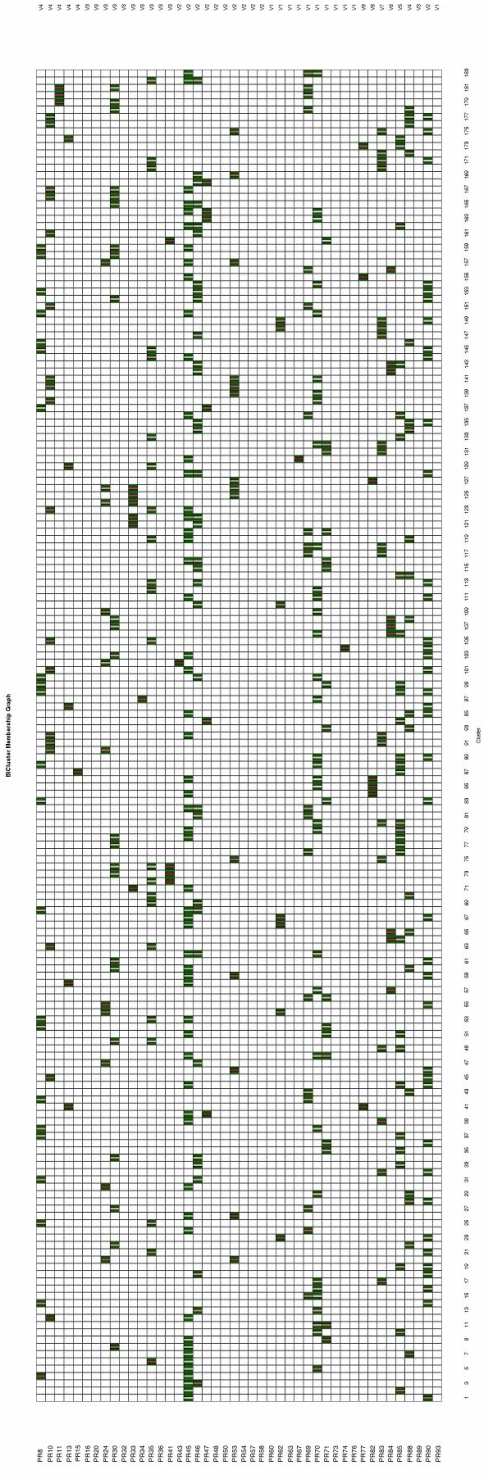

Figura 4.19: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 2 colunas e subtipo B*

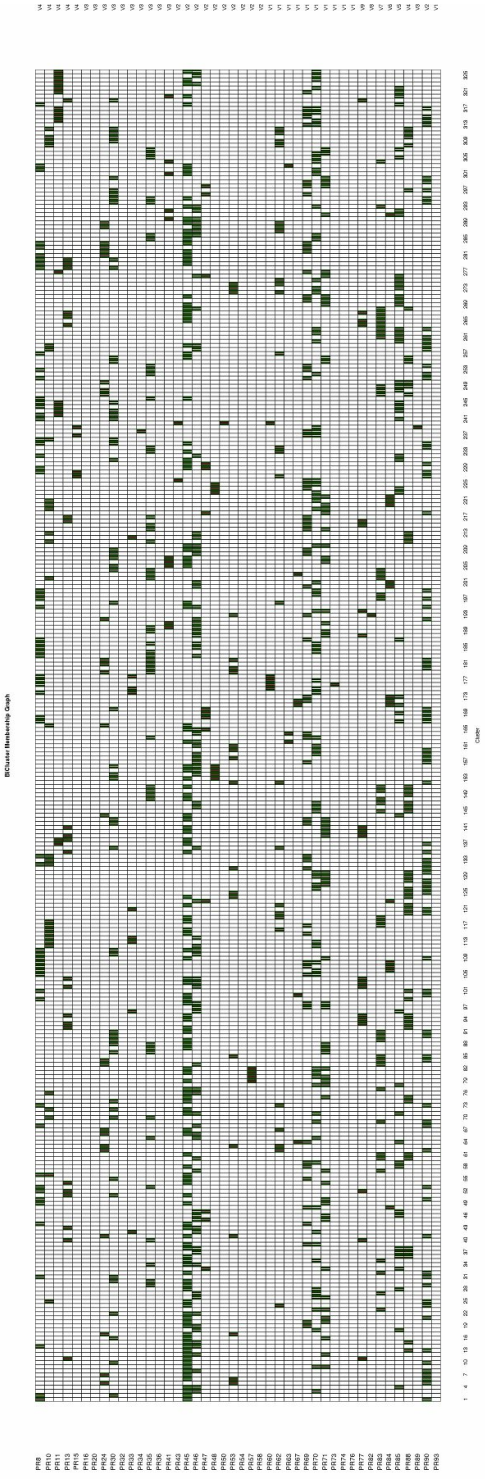

Figura 4.20: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 3 colunas e subtipo B*

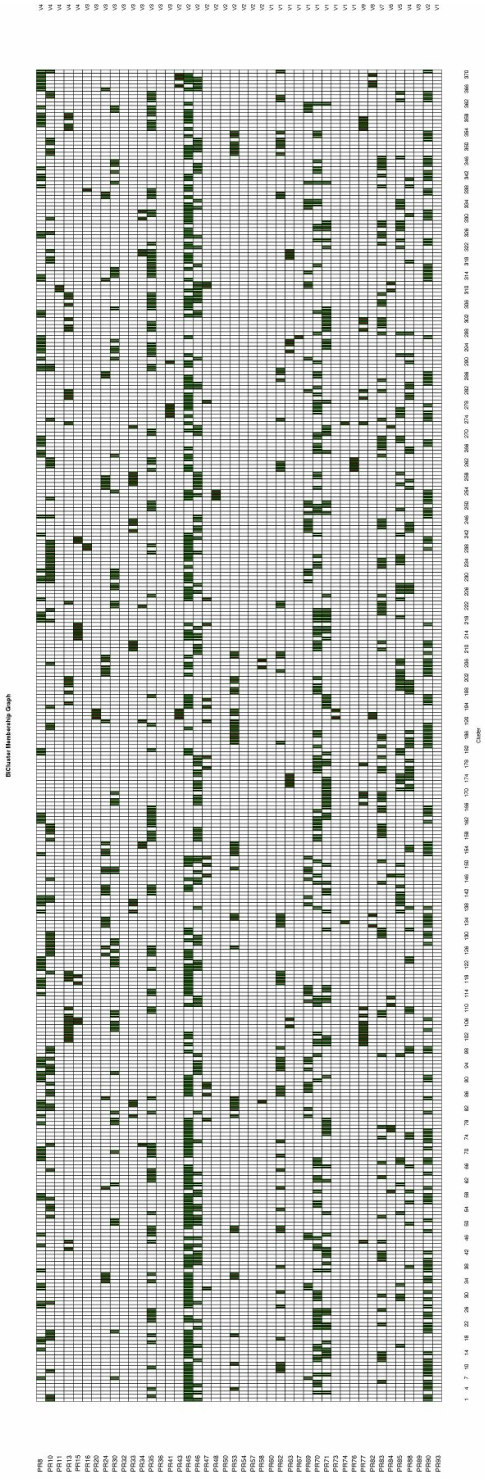

Figura 4.21: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 4 colunas e subtipo B*

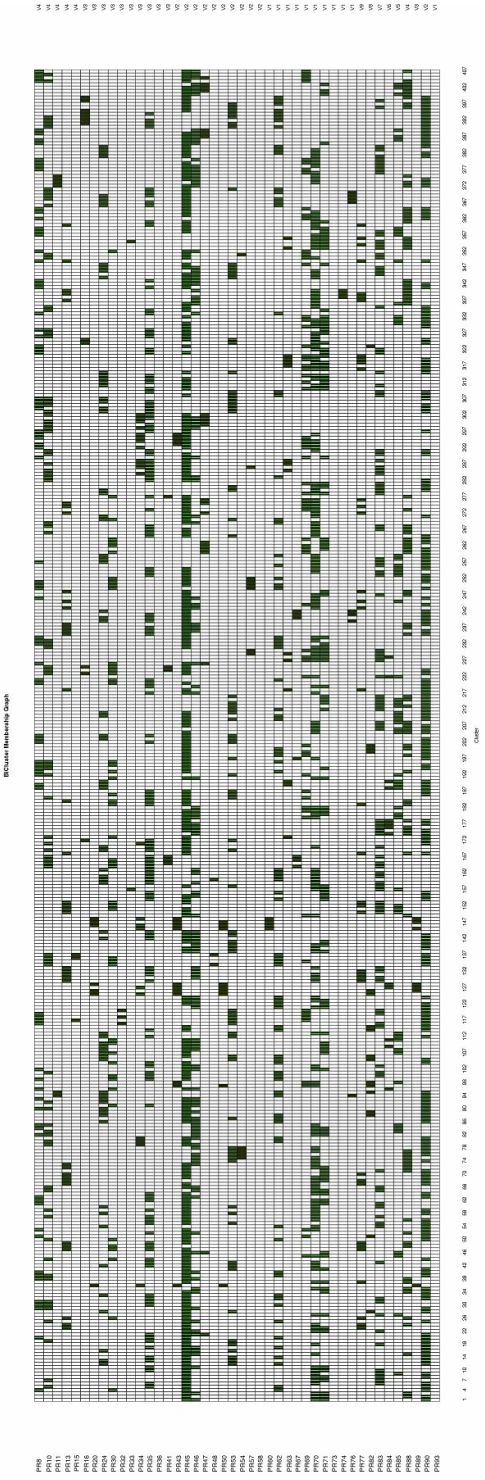

Figura 4.22: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 5 colunas e subtipo B*

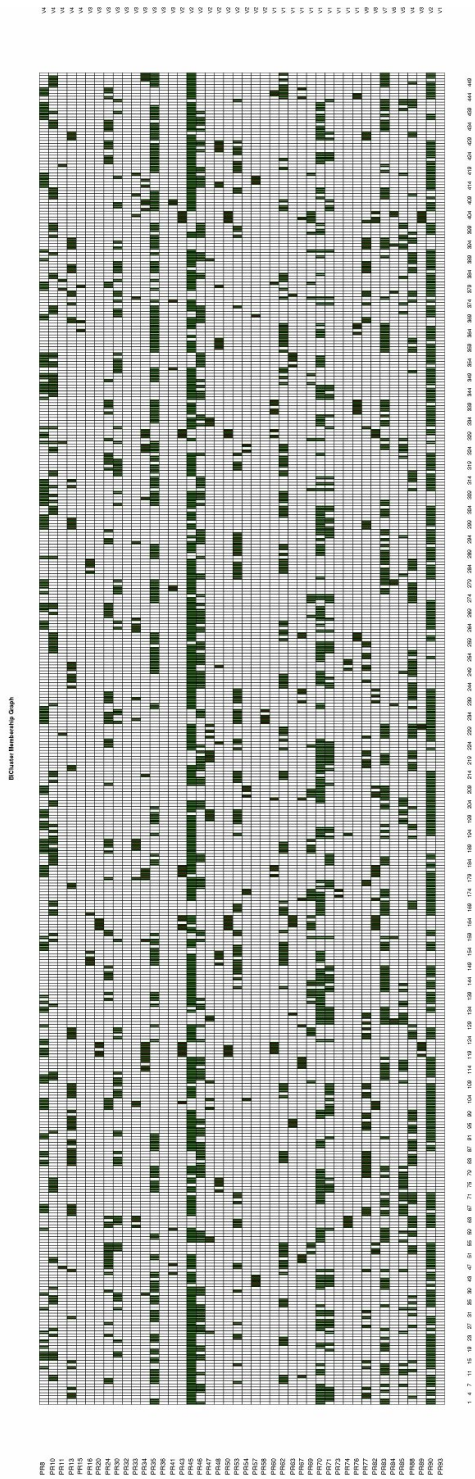

Figura 4.23: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 6 colunas e subtipo B*

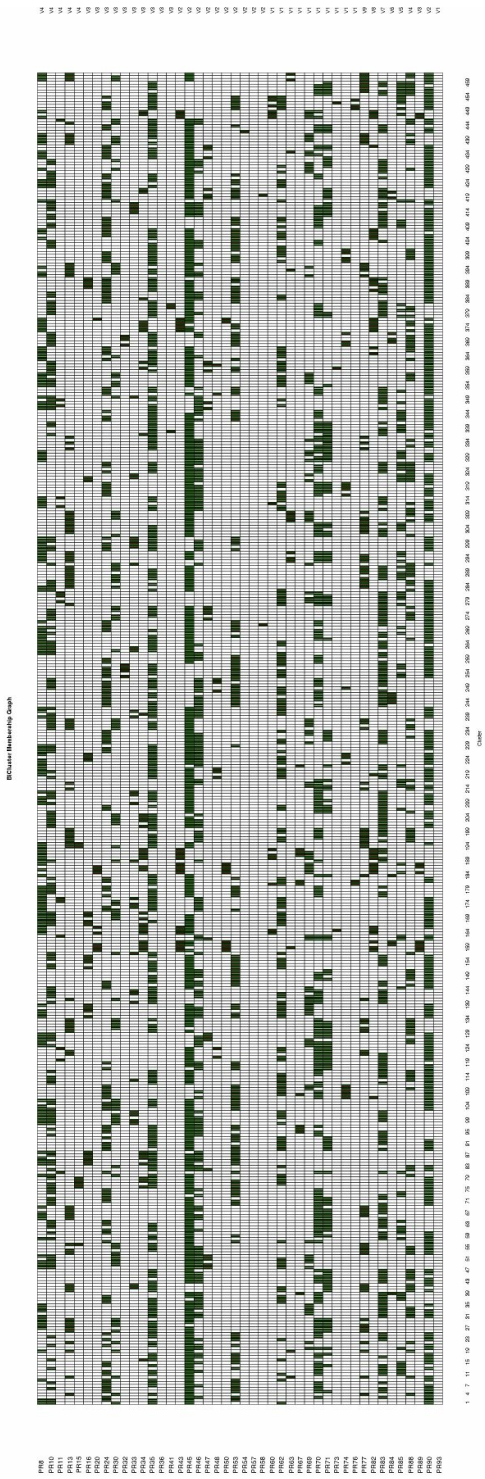

Figura 4.24: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 7 colunas e subtipo B*

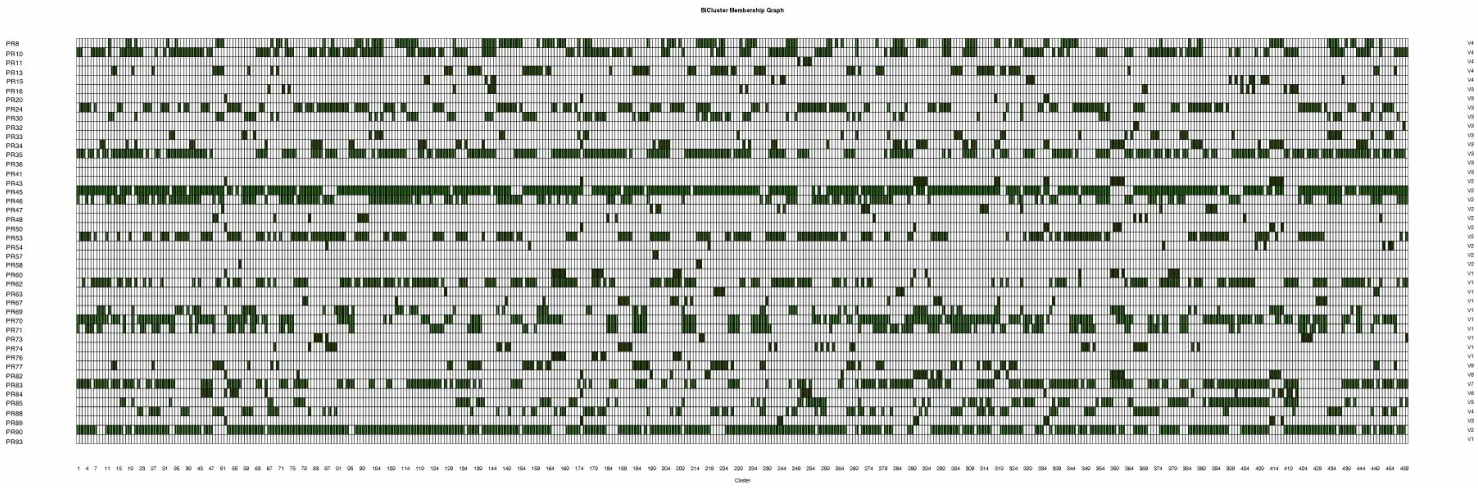

Figura 4.25: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 8 colunas e subtipo B*

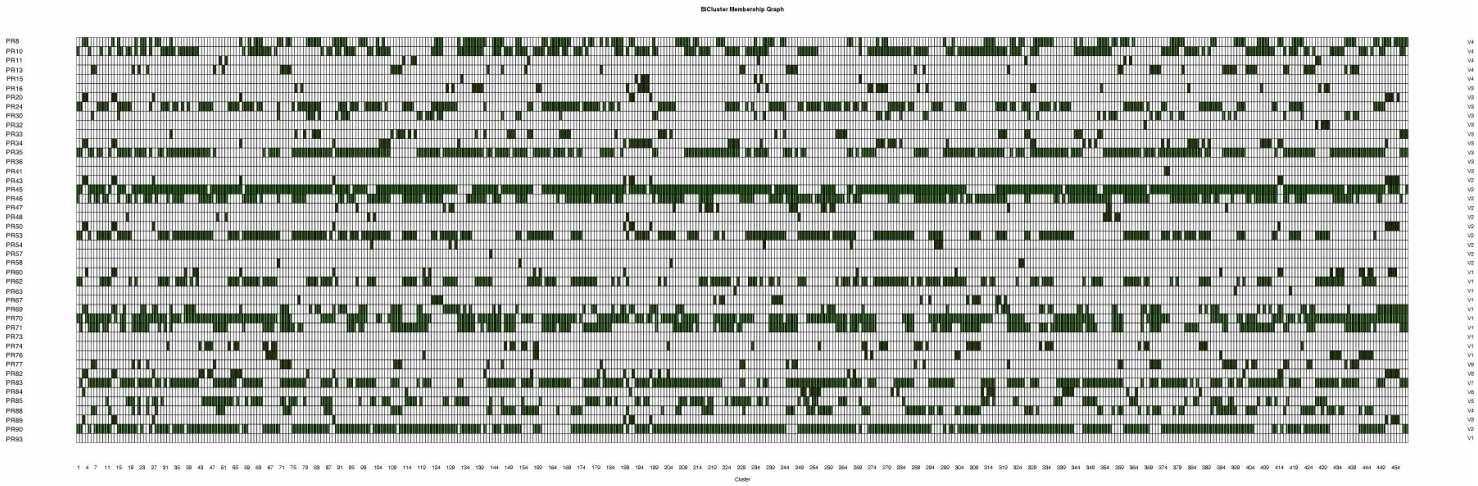

Figura 4.26: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 9 colunas e subtipo B*

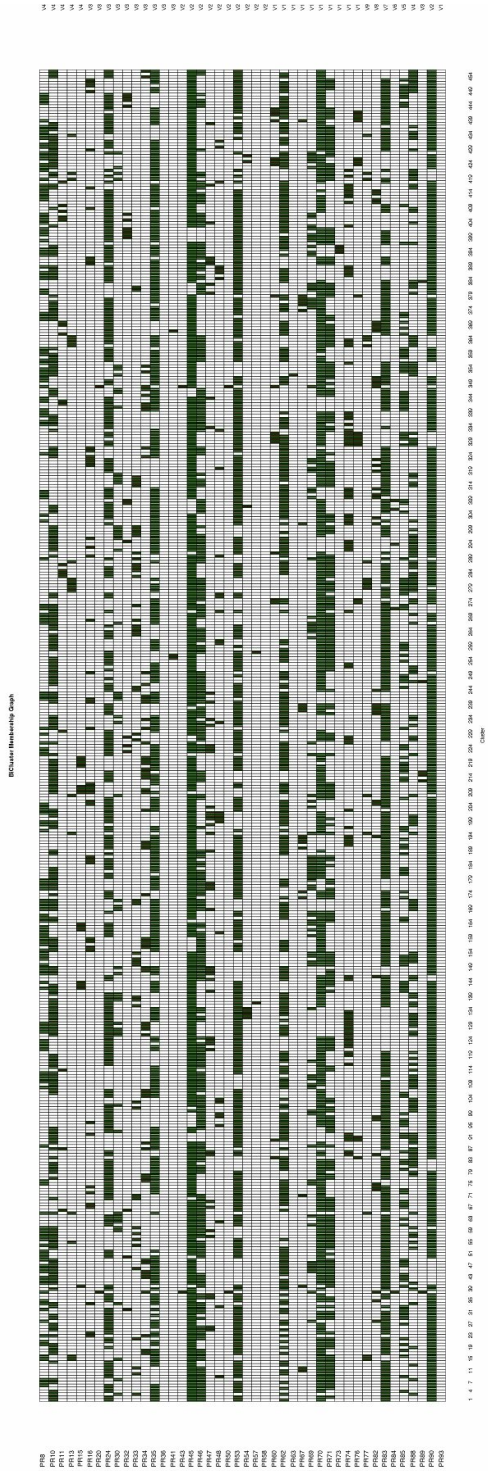

Figura 4.27: *Gráfico da distribuição das posições de protease que definem cada bicluster de mínimo 10 colunas e subtipo B*

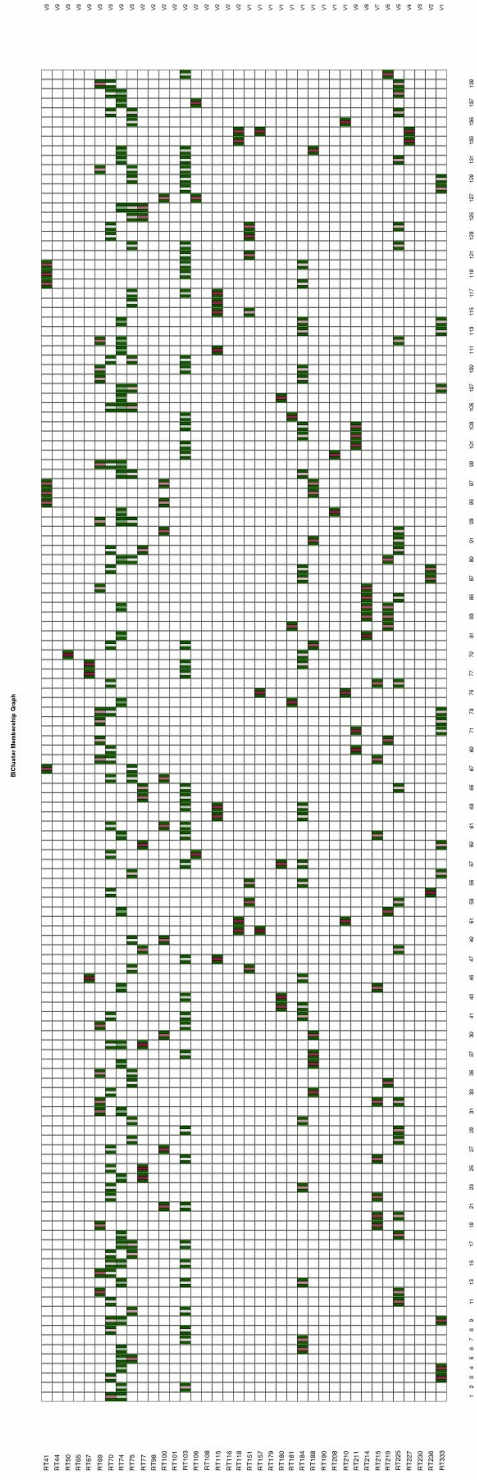

Figura 4.28: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 2 colunas e subtipo B*

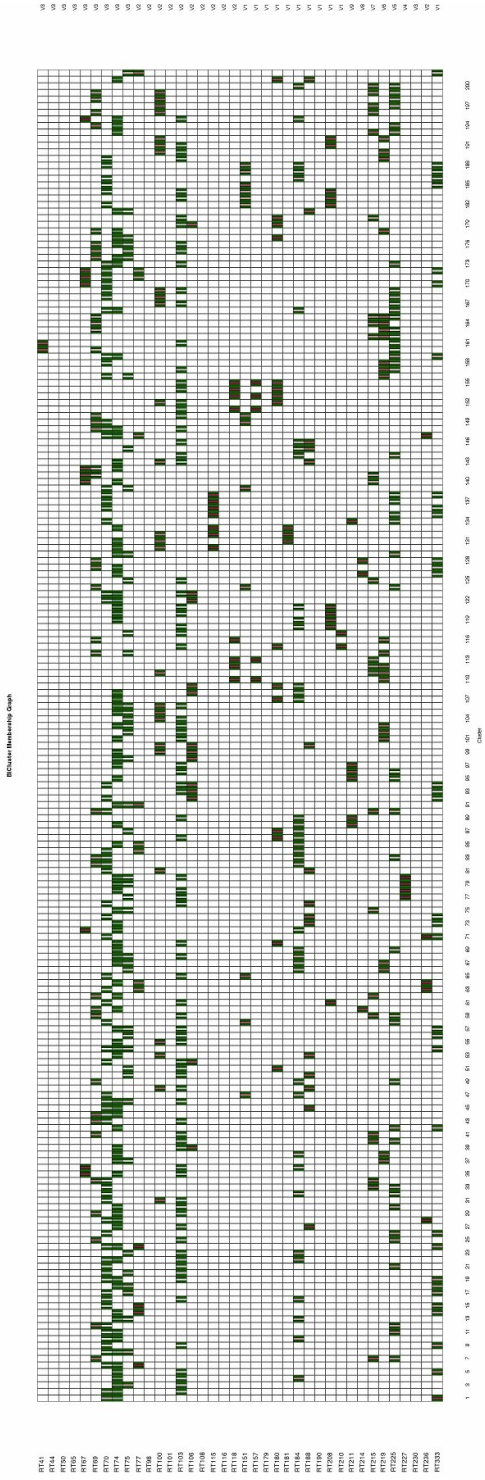

Figura 4.29: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 3 colunas e subtipo B*

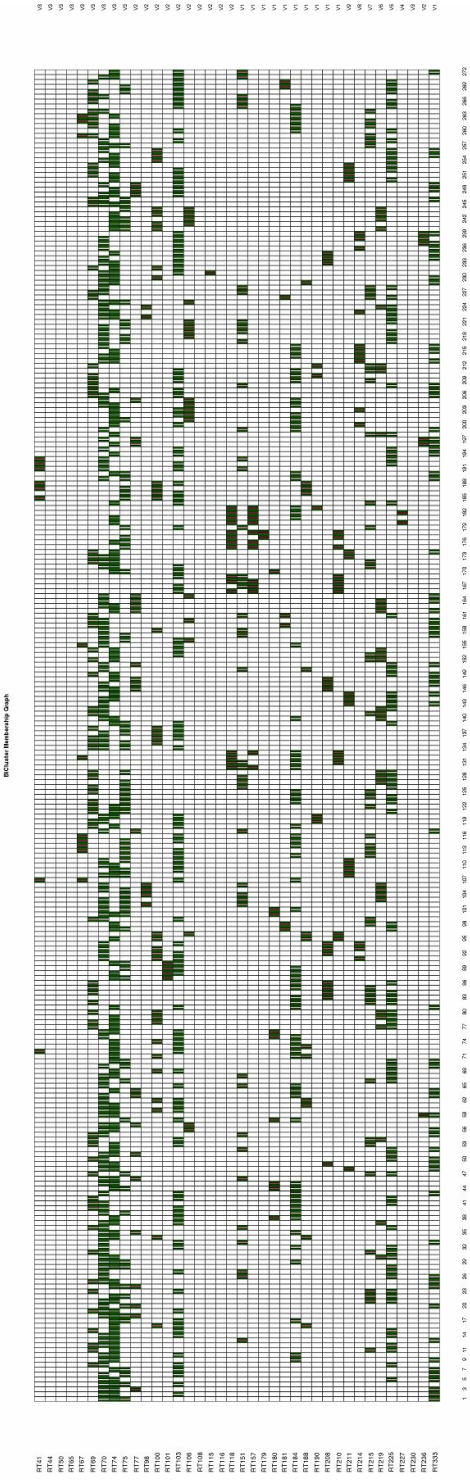

Figura 4.30: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 4 colunas e subtipo B*

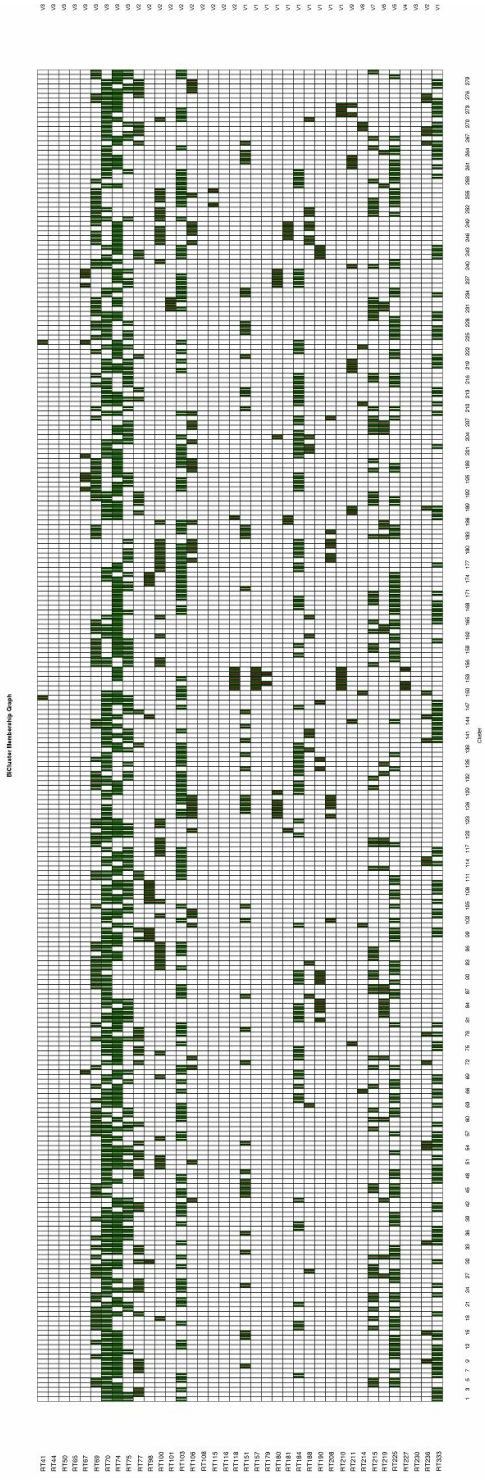

Figura 4.31: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 5 colunas e subtipo B*

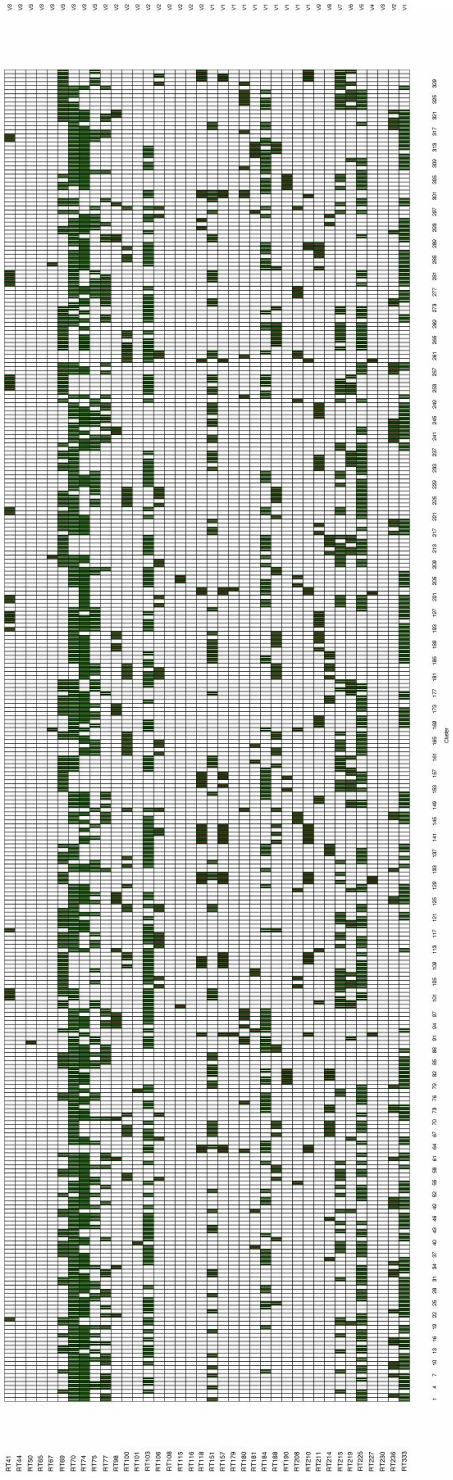

Figura 4.32: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 6 colunas e subtipo B*

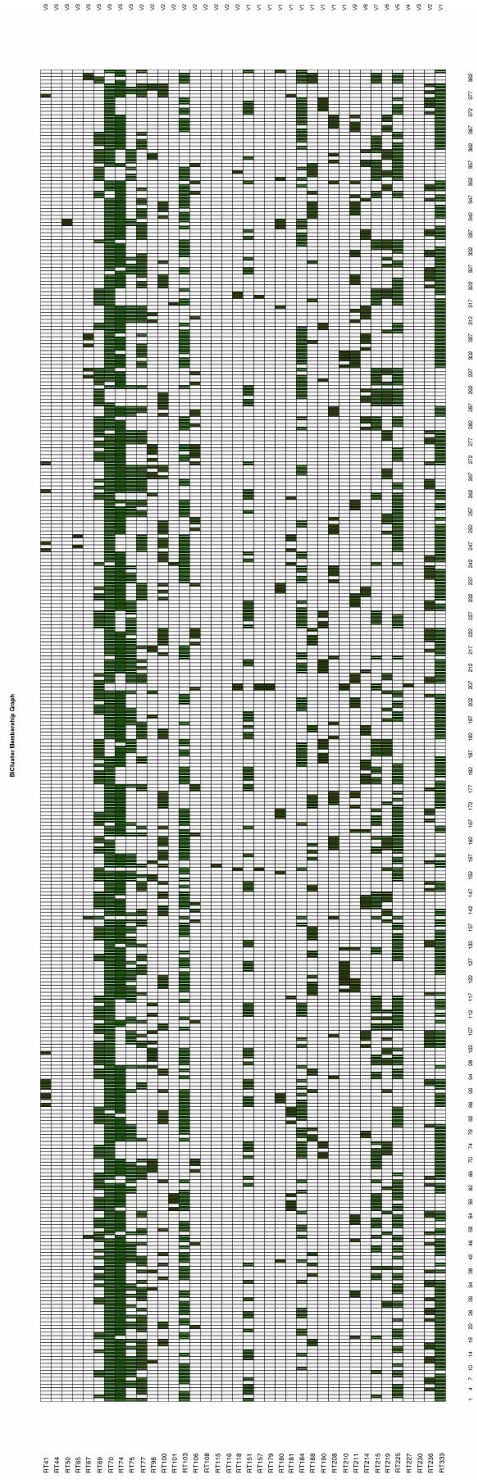

Figura 4.33: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 7 colunas e subtipo B*

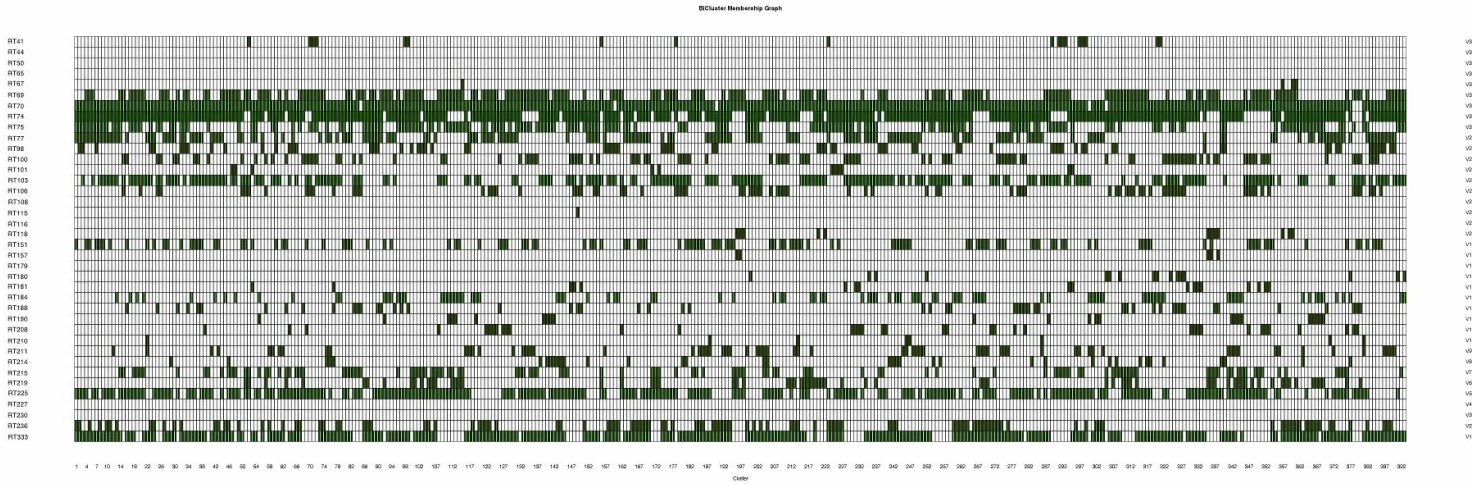

Figura 4.34: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 8 colunas e subtipo B*

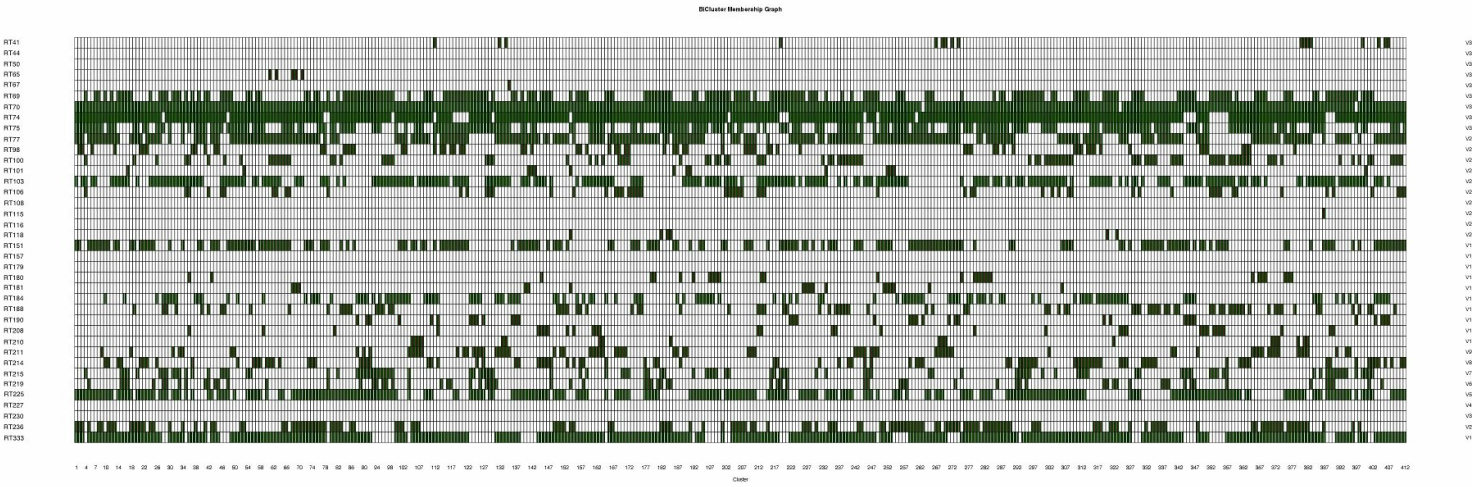

Figura 4.35: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 9 colunas e subtipo B*

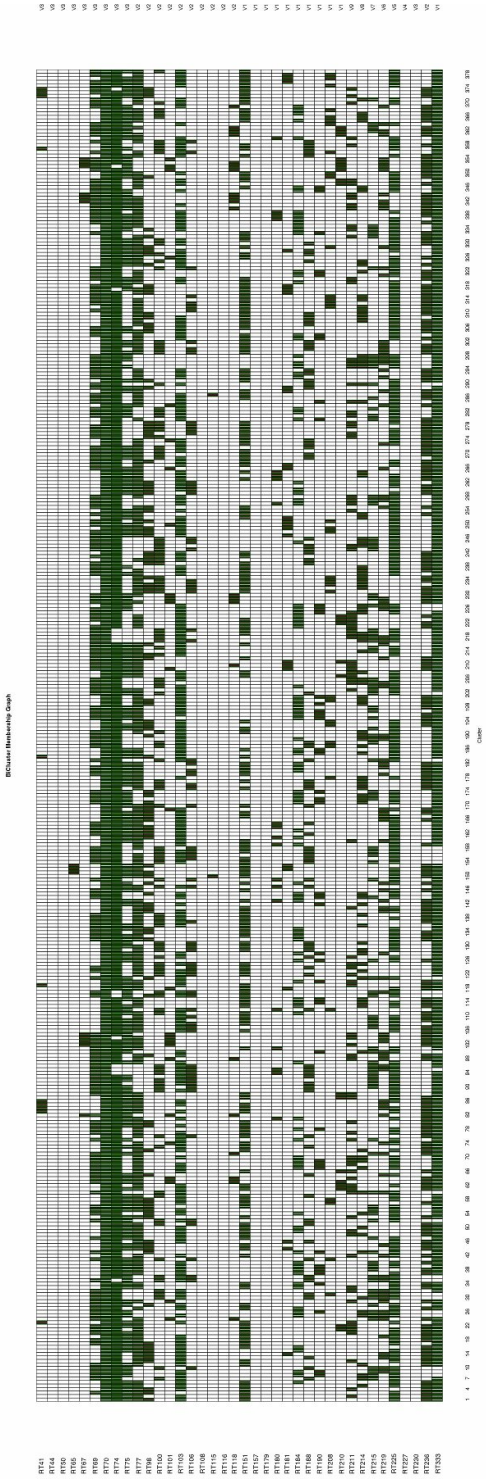

Figura 4.36: *Gráfico da distribuição das posições de transcriptase reversa que definem cada bicluster de mínimo 10 colunas e subtipo B*

Tabela 4.24: *Posições da Protease que não participam de biclusters de subtipo B*

Sub-tipo	Mínimo de colunas	Posições que não participam de *biclusters*
B	2	PR32 PR70 PR84 PR83 PR57 PR11 PR53 PR48 PR50 PR58 PR47 PR8 PR34 PR45 PR76
B	3	PR33 PR32 PR70 PR84 PR83 PR53 PR48 PR8 PR76
B	4	PR70 PR57 PR11 PR53 PR50 PR47 PR8 PR76
B	5	PR70 PR48 PR8 PR34
B	6	PR70 PR8 PR76
B	7	PR70 PR50 PR8
B	8	PR70 PR69 PR8 PR76
B	9	PR70 PR69 PR50 PR8 PR34 PR45 PR76
B	10	PR70 PR69 PR48 PR50 PR8 PR34 PR45

Tabela 4.25: *Posições da Protease que não participam de biclusters de subtipo C*

Sub-tipo	Mínimo de colunas	Posições que não participam de *biclusters*
C	2	PR67 PR33 PR32 PR70 PR88 PR30 PR84 PR74 PR83 PR57 PR24 PR11 PR53 PR48 PR46 PR50 PR58 PR47 PR8 PR34 PR73 PR76 PR43 PR62 PR54
C	3	PR67 PR33 PR32 PR71 PR70 PR88 PR30 PR84 PR74 PR83 PR57 PR11 PR53 PR48 PR46 PR50 PR85 PR41 PR58 PR47 PR8 PR34 PR73 PR45 PR76 PR43 PR54
C	4	PR67 PR32 PR71 PR70 PR88 PR30 PR84 PR83 PR57 PR11 PR53 PR48 PR46 PR50 PR85 PR41 PR58 PR47 PR8 PR34 PR73 PR76 PR43
C	5	PR33 PR32 PR88 PR30 PR84 PR57 PR11 PR53 PR48 PR50 PR58 PR47 PR8 PR34 PR73 PR43 PR54
C	6	PR32 PR84 PR11 PR53 PR48 PR77 PR85 PR58 PR47 PR8 PR34 PR73 PR43
C	7	PR32 PR84 PR83 PR57 PR11 PR53 PR48 PR77 PR50 PR58 PR47 PR8 PR34 PR73 PR45 PR76 PR43
C	8	PR67 PR33 PR32 PR70 PR84 PR83 PR57 PR24 PR11 PR53 PR48 PR77 PR85 PR41 PR58 PR47 PR8 PR60 PR34 PR73 PR45 PR76 PR43
C	9	PR67 PR32 PR70 PR88 PR30 PR16 PR84 PR83 PR57 PR24 PR11 PR53 PR48 PR77 PR85 PR41 PR58 PR47 PR8 PR60 PR34 PR73 PR45 PR76 PR43
C	10	PR67 PR33 PR32 PR70 PR88 PR30 PR16 PR84 PR83 PR57 PR24 PR11 PR53 PR48 PR77 PR50 PR85 PR41 PR58 PR47 PR8 PR60 PR34 PR73 PR45 PR76 PR43

Tabela 4.26: *Posições da Protease que não participam de biclusters de subtipo F*

Sub-tipo	Mínimo de colunas	Posições que não participam de *biclusters*
F	2	PR67 PR32 PR90 PR71 PR70 PR88 PR30 PR16 PR82 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR50 PR85 PR58 PR47 PR8 PR34 PR73 PR45 PR76 PR54
F	3	PR67 PR33 PR32 PR71 PR70 PR88 PR30 PR84 PR83 PR57 PR69 PR24 PR11 PR53 PR48 PR50 PR85 PR47 PR8 PR34 PR73 PR45 PR76 PR43
F	4	PR67 PR33 PR32 PR71 PR70 PR16 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR50 PR85 PR58 PR47 PR8 PR34 PR73 PR45 PR76
F	5	PR67 PR33 PR32 PR70 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR50 PR85 PR58 PR47 PR8 PR34 PR73 PR76
F	6	PR67 PR32 PR70 PR84 PR83 PR57 PR69 PR11 PR48 PR85 PR47 PR8 PR34 PR73 PR76
F	7	PR67 PR32 PR70 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR85 PR47 PR8 PR34 PR73 PR76
F	8	PR67 PR32 PR70 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR50 PR85 PR47 PR8 PR34 PR73 PR45 PR76 PR43
F	9	PR67 PR33 PR32 PR70 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR77 PR50 PR85 PR58 PR47 PR8 PR34 PR73 PR45 PR76
F	10	PR67 PR33 PR32 PR70 PR88 PR30 PR16 PR84 PR83 PR57 PR69 PR11 PR53 PR48 PR77 PR50 PR85 PR58 PR47 PR8 PR60 PR34 PR73 PR45 PR76 PR43

Tabela 4.27: *Posições da Transcriptase Reversa que não participam de biclusters de subtipo B*

Sub-tipo	Mínimo de colunas	Posições que não participam de *biclusters* de subtipo B
B	2	RT227 RT188 RT180 RT100 RT236 RT115 RT208 RT157 RT50
B	3	RT227 RT188 RT180 RT100 RT236 RT230 RT115 RT208 RT157 RT50
B	4	RT227 RT180 RT236 RT230 RT157 RT50
B	5	RT227 RT180 RT236 RT230 RT157 RT50
B	6	RT227 RT180 RT236 RT157 RT50
B	7	RT180 RT236 RT157 RT65 RT50
B	8	RT227 RT180 RT236 RT230 RT115 RT157 RT65 RT50
B	9	RT180 RT116 RT179 RT236 RT230 RT115 RT157 RT65 RT50
B	10	RT180 RT116 RT179 RT236 RT230 RT115 RT157 RT65 RT50

Tabela 4.28: *Posições da Transcriptase Reversa que não participam de biclusters de subtipo C*

Sub-tipo	Mínimo de colunas	Posições que não participam de *biclusters*
C	2	RT227 RT180 RT116 RT44 RT74 RT75 RT236 RT230 RT108 RT115 RT151 RT77 RT157 RT65 RT50
C	3	RT227 RT188 RT180 RT116 RT179 RT44 RT74 RT75 RT236 RT230 RT115 RT151 RT208 RT77 RT157 RT65 RT50
C	4	RT188 RT180 RT116 RT179 RT100 RT74 RT333 RT236 RT230 RT115 RT151 RT77 RT157 RT65 RT50
C	5	RT188 RT180 RT116 RT179 RT100 RT74 RT333 RT236 RT230 RT115 RT151 RT77 RT157 RT65 RT50
C	6	RT227 RT188 RT180 RT116 RT179 RT100 RT74 RT333 RT75 RT236 RT230 RT115 RT151 RT225 RT181 RT77 RT157 RT65 RT50
C	7	RT227 RT188 RT180 RT116 RT179 RT100 RT74 RT333 RT75 RT236 RT230 RT108 RT115 RT151 RT225 RT181 RT77 RT157 RT65 RT50 RT98
C	8	RT227 RT188 RT180 RT116 RT179 RT100 RT74 RT333 RT190 RT75 RT236 RT230 RT108 RT115 RT151 RT225 RT181 RT208 RT77 RT106 RT157 RT65 RT50 RT98 RT101

Tabela 4.29: *Posições da Transcriptase Reversa que não participam de biclusters de subtipo F*

Sub-tipo	Mínimo de colunas	Posições que não participam de *biclusters*
F	2	RT227 RT180 RT116 RT179 RT100 RT44 RT333 RT236 RT230 RT115 RT151 RT77 RT157 RT65 RT50
F	3	RT227 RT180 RT116 RT100 RT74 RT75 RT236 RT115 RT151 RT77 RT157 RT65 RT50
F	4	RT227 RT180 RT116 RT179 RT75 RT236 RT115 RT151 RT77 RT157 RT65 RT50
F	5	RT227 RT180 RT116 RT179 RT236 RT230 RT115 RT151 RT77 RT157 RT65 RT50
F	6	RT227 RT188 RT180 RT116 RT179 RT100 RT333 RT75 RT236 RT230 RT115 RT151 RT225 RT77 RT157 RT65 RT50 RT98
F	7	RT227 RT188 RT180 RT116 RT179 RT100 RT333 RT75 RT236 RT230 RT115 RT151 RT225 RT77 RT157 RT65 RT50 RT98
F	8	RT227 RT188 RT180 RT116 RT179 RT100 RT333 RT75 RT236 RT230 RT115 RT151 RT225 RT77 RT106 RT157 RT65 RT50 RT98 RT101
F	9	RT227 RT188 RT180 RT116 RT179 RT100 RT74 RT333 RT190 RT75 RT236 RT230 RT108 RT115 RT69 RT103 RT151 RT225 RT181 RT77 RT106 RT157 RT65 RT50 RT98 RT101
F	10	RT67 RT118 RT70 RT227 RT188 RT180 RT116 RT179 RT100 RT44 RT74 RT333 RT190 RT75 RT236 RT230 RT108 RT115 RT184 RT69 RT215 RT103 RT151 RT225 RT181 RT208 RT77 RT219 RT106 RT214 RT157 RT65 RT50 RT210 RT41 RT211 RT98 RT101

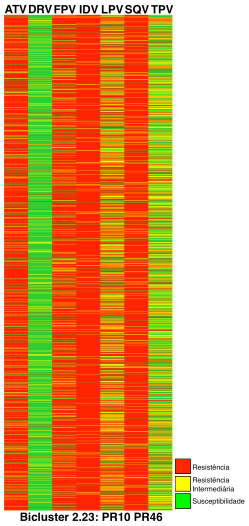

Figura 4.37: *Imagem colorida do bicluster PR10 PR46 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

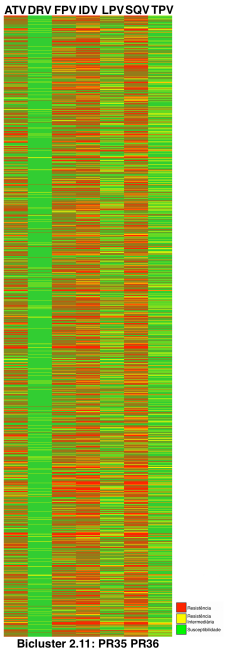

Figura 4.38: *Imagem colorida do bicluster PR35 PR36 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

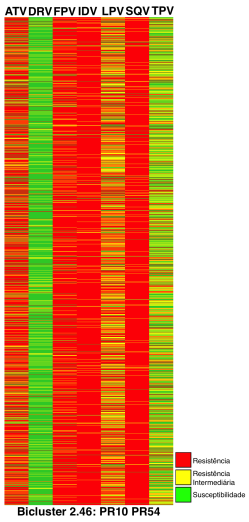

Figura 4.39: *Imagem colorida do bicluster PR10 PR54 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

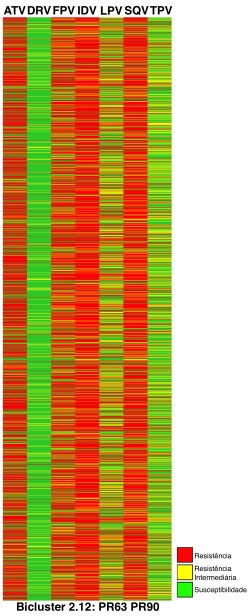

Figura 4.40: *Imagem colorida do bicluster PR63 PR90 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

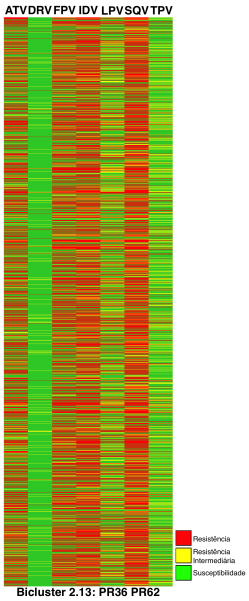

Figura 4.41: *Imagem colorida do bicluster PR36 PR62 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

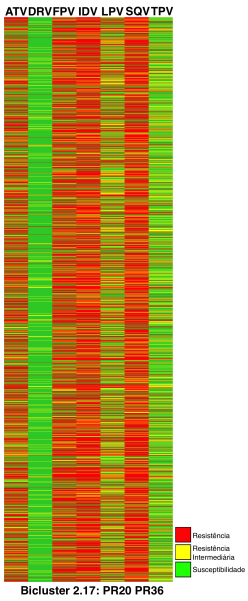

Figura 4.42: *Imagem colorida do bicluster PR20 PR36 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

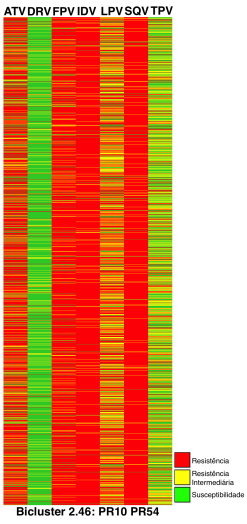

Figura 4.43: *Imagem colorida do bicluster PR10 PR54 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

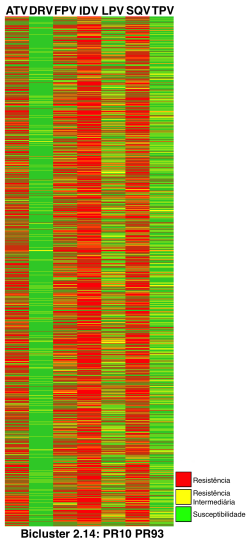

Figura 4.44: *Imagem colorida do bicluster PR10 PR93 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos look-up table brasileira e as linhas, as sequências de proteína.*

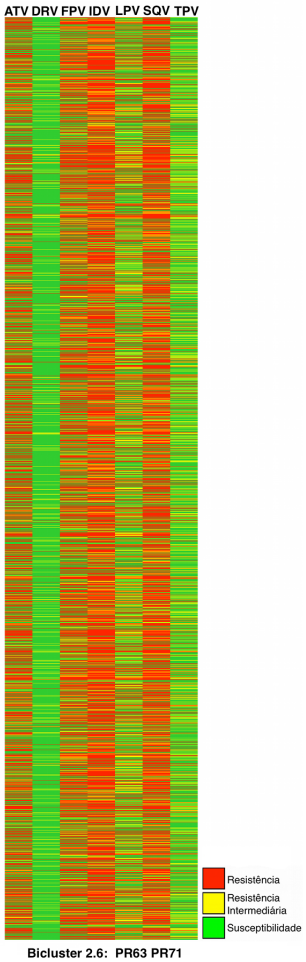

Figura 4.45: *Imagem colorida do bicluster PR63 PR71 para sequências de Protease Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

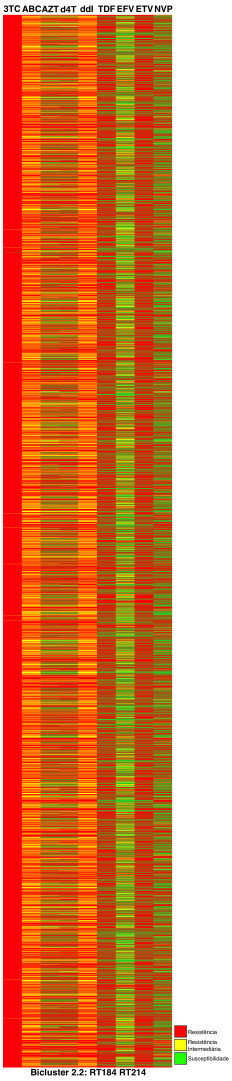

Figura 4.46: *Imagem colorida do bicluster RT184 RT214 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

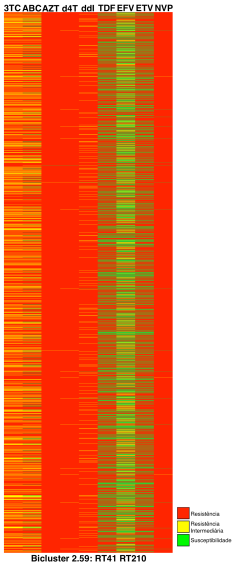

Figura 4.47: *Imagem colorida do bicluster RT41 RT210 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

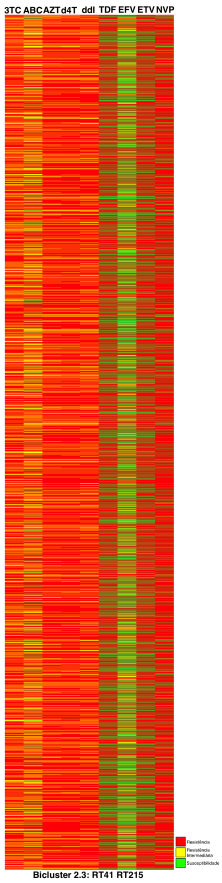

Figura 4.48: *Imagem colorida do bicluster RT41 RT215 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

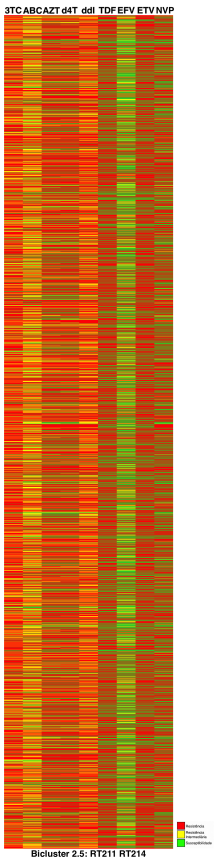

Figura 4.49: *Imagem colorida do bicluster RT211 RT214 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

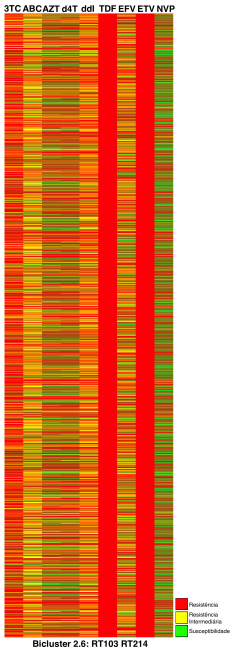

Figura 4.50: *Imagem colorida do bicluster RT103 RT214 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

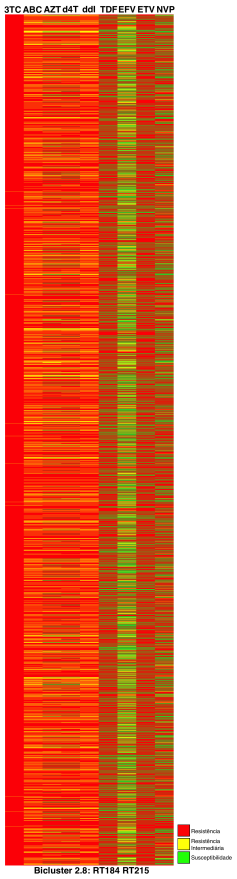

Figura 4.51: *Imagem colorida do bicluster RT184 RT215 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

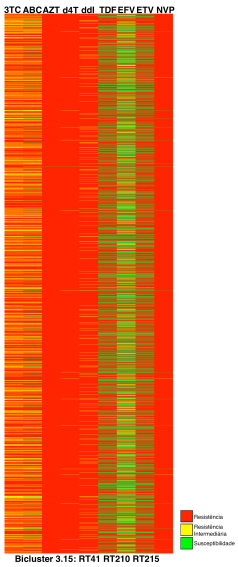

Figura 4.52: *Imagem colorida do bicluster RT41 RT210 RT215 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

Figura 4.53: *Imagem colorida do bicluster RT41 RT184 RT215 para sequências de Transcriptase Reversa Subtipo B. As colunas na Figura representam os medicamentos da look-up table brasileira e as linhas, as sequências de proteína.*

Capítulo 5

Conclusão

Neste trabalho, uma nova abordagem para análise de mutações de HIV é apresentada. Os esquemas de classificação atuais de sequências de HIV são baseados em sistemas baseados em regras, bancos de dados e *look-up tables* que fazem uma compilação da informação contida em estudos científicos. O *framework* proposto se baseia em uma representação bitmap que extrai informação da protease e transcriptase reversa e fornece informação sobre a probabilidade de sucesso no tratamento, auxiliando na elucidação das interações entre mutações.

Um novo esquema de visualização inspirado na análise de dados de microarranjo foi proposto com intuito de entender melhor os grupos no domínio do HIV. As figuras são importantes na visualização e comparação dos grupos contendo vetores binários e grandes volumes de dados. No nosso estudo, as figuras em preto e branco indicam a presença e ausência de mutações nas sequências em cada grupo, portanto, acentuando as diferenças entre os grupos.

O $K-$Médias e o Bimax geraram grupos e *biclusters* com sequências similares representando diferentes padrões de mutações. Os grupos e *biclusters* mostraram algumas mutações que frequentemente ocorrem juntas, que são importantes para a definição dos grupos e que estão presentes em um grande número de sequências. Essas posições precisam ser consideradas quando se infere a probabilidade de resistência aos medicamentos porque afetam um grande número de pacientes e estão relacionados à diferentes padrões de predição de resistência aos medicamentos provenientes da *look-up table* brasileira.

Exemplos de combinações de posições importantes na identificação de padrões de mutações foram indicadas pelo $K-$Médias e incluem as posições PR30 e PR88; PR10, PR54, PR82 e PR90; e RT41, RT67 e RT210. Dentre essas combinações, as posições PR30 e PR88 já foram previamente relacionadas à resistência a nelfinavir (Rhee *et al.*, 2003) e as posições RT67, RT70 e RT219, à susceptibilidade diminuída a ZDV e D4T (Rhee *et al.*, 2003), o que é coerente com os resultados das Figuras 4.13 e 4.16.

As posições RT41 e RT210 definem as mutações TAM1 (thymidine-associated mutations 1) e as posições RT70 e RT219 definem as mutações TAM2 (thymidine-associated mutations 2) (Flandre *et al.*, 2004; Gonzales *et al.*, 2003; Hanna *et al.*, 2000; Marcelin *et al.*, 2004; Wolf *et al.*, 2003; Yahi *et al.*, 1999, 2000) e foram agrupadas separadamente, como esperado. Contudo, a posição RT67, que também caracteriza as TAM2 (Flandre *et al.*, 2004; Gonzales *et al.*, 2003; Hanna *et al.*, 2000; Marcelin *et al.*, 2004; Wolf *et al.*, 2003; Yahi *et al.*, 1999, 2000) foi agrupada com ambas TAMs, o que também foi descrito em (Sing *et al.*, 2005). Adicionalmente à posição RT67, a posição RT215 também aparecem com ambas TAMs, o que também foi observado em (Yahi *et al.*, 1999).

O Bimax também gerou *biclusters* para o subtipo B com as posições PR30 e PR88; RT67, RT 70 e RT219; e RT41, RT67, RT210 e RT215. Já as posições PR10, PR54, PR82 e PR90 foram colocadas juntas em um *bicluster* junto com a posição PR93.

O algoritmo de *biclustering* apresentou alguns *biclusters* com padrões de mutações parecidos nos diferentes subtipos, como por exemplo, os já citados por outros trabalhos, RT41 e RT215; RT70 e RT184; RT70 e RT219; RT67 e RT219; RT184 e RT215; RT67, RT70 e RT219; RT41, RT184 e RT215; PR20 e PR36; e PR15, PR20 e PR36. No entanto, algumas combinações de posições como PR30 e PR88 não foram encontradas como um *bicluster* separado, sem inclusão de outras posições, em todos subtipos. Isso pode indicar que esse par de posições pode estar relacionado com outras posições nos subtipos C e F ou pode ter sido reflexo da menor representatividade da variabilidade dos subtipos C e F, que possuíam conjuntos de sequências menores. Outra possibilidade é a diferença existente na probabilidade de ocorrência de uma mutação devido ao subtipo, e o fato de que os estudos de padrões em sua maioria são dados em conjuntos de sequências de subtipo B.

Além dos *biclusters* já encontrados por trabalhos anteriores, também foram encontrados vários outros padrões de mutações. Esses padrões de mutações incluem pares de posições de proteínas e combinações de

até 10 posições e são importantes, já que ocorrem em grande número de sequências.

Esses grupos e *biclusters* encontrados podem ser analisados quanto à correspondência com os resultados do tratamento de pacientes utilizando sequências de pacientes que têm a resistência ou susceptibilidade aos medicamentos conhecidos, analisando as posições importantes e encontrando a qual grupo a sequência pertence. Se os grupos e *biclusters* forem relacionados aos resultados de tratamentos, podem ser utilizados para predição de resistência ou susceptibilidade, a partir das sequências de proteínas.

Como os medicamentos inibidores de protease são criados de acordo com a estrutura da protease nos virus, os grupos também podem auxiliar no desenvolvimento de medicamentos, já que indicam algumas mutações que estão agrupadas e ocorrem juntas. Um medicamento que pudesse agir em diferentes grupos seria um medicamento com maior probabilidade de sucesso.

O entendimento dos padrões de mutações também é de interesse do estudo das vias de escape ao sistema imune e pesquisa em vacinas contra HIV. Um dos mecanismo de escape ao sistema imune é a seleção de mutações para evasão do reconhecimento imune. No entanto, os padrões de mutações do HIV podem atrapalhar na elucidação dos mecanismos de escape imune (Brumme *et al.*, 2008) que são relevantes na pesquisa sobre vacinas (Brumme *et al.*, 2009).

Como próximos passos desse trabalho estão incluídos a investigação da relação dos grupos e *biclusters* a sequências com resistência aos medicamentos conhecidos e a implementação e disponibilização das rotinas de binarização dos dados e visualização dos clusters e *biclusters* na forma integrada nos pacotes R.

Referências Bibliográficas

5**algoritmoBrasileiro()** algoritmoBrasileiro. Brazilian algorithm. URL http://forrest.ime.usp.br:3001/resistencia. Citado na pág. 2, 7

Alteri *et al.***(2009)** Claudia Alteri, Valentina Svicher, Caterina Gori, Roberta D'Arrigo, Massimo Ciccozzi, Francesca Ceccherini-Silberstein, Marina Selleri, Stefano Aviani Bardacci, Massimo Giuliani, Paola Elia, Paola Scognamiglio, Roberta Balzano, Nicoletta Orchi, Enrico Girardi, Carlo Federico Perno, e SENDIH Study Group. Characterization of the patterns of drug-resistance mutations in newly diagnosed hiv-1 infected patients naïve to the antiretroviral drugs. *BMC Infectious Diseases*, 9(1):111. Citado na pág. 14, 15, 40

AntiretroviralGuidelines(2012) AntiretroviralGuidelines. Panel on antiretroviral guidelines for adults and adolescents. http://www.aidsinfo.nih.gov/contentfiles/adultandadolescentgl.pdf, March 2012. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. US Department of Health and Human Services. Citado na pág. 7

Araújo *et al.***(2008)** Luciano Vieira Araújo, Ester C. Sabino e João Eduardo Ferreira. Hiv drug resistance analysis tool based on process algebra. páginas 1358–1363. Citado na pág. 23

Baeten *et al.***(2007)** Jared M Baeten, Bhavna Chohan, Ludo Lavreys, Vrasha Chohan, R Scott McClell, Laura Certain, Kishorchandra Mandaliya, Walter Jaoko e Julie Overbaugh. Hiv-1 subtype d infection is associated with faster disease progression than subtype a in spite of similar plasma hiv-1 loads. *J Infect Dis*, 195:1177–80. Citado na pág. 6

Bannister *et al.***(2006)** Wendy P Bannister, Lidia Ruiz, Clive Loveday, Stefano Vella, Kai Zilmer, Jesper Kjoer, Brygida Knysz, Andrew N Phillips, Amanda Mocroft e Jens D Lundgren. Hiv-1 subtypes and response to combination antiretroviral therapy in europe. *Antivir Ther*, 11(6):707–15. Citado na pág. 7

Beerenwinkel *et al.***(2001)** Niko Beerenwinkel, Thomas Lengauer, Joachim Selbig, Barbara Schmidt, Hauke Walter, Klaus Korn, Rolf Kaiser e Daniel Hoffmann. Geno2pheno: Interpreting enothypic hiv drug resistance test. *IEEE Intelligent Systems*, 16:35–41. Citado na pág. 7

Beerenwinkel *et al.***(2005)** Niko Beerenwinkel, Tobias Sing, Thomas Lengauer, Jörg Rahnenführer, Kirsten Roomp, Igor Savenkov, Roman Fischer, Daniel Hoffmann, Joachim Selbig, Klaus Korn, Hauke Walter, Thomas Berg, Patrick Braun, Gerd Fätkenheuer, Mark Oette, Jürgen Rockstroh, Bernd Kupfer, Rolf Kaiser e Martin Däumer. Computational methods for the design of effective therapies against drug resistant hiv strains. *Bioinformatics*, 21:3943–3950. Citado na pág. 22

Ben-Dor *et al.***(2003)** Amir Ben-Dor, Benny Chor, Richard Karp e Zohar Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of Computational Biology*, 10(3-4):373–384. Citado na pág. 2, 8

Bilu e Linial(2002) Yonatan Bilu e Michal Linial. The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classifications. *Journal of Computational Biology*, 9(2):193–210. Citado na pág. 8

Boden e Markowitz(1998) Daniel Boden e Martin Markowitz. Resistance to human immunodeficiency virus type 1 protease inhibitors. *Antimicrobial agents and chemotherapy*, 42(11):2775–2783. Citado na pág. 6

Bradley e Fayyad(1998) Paul S Bradley e Usama M Fayyad. Refining initial points for k-means clustering. *ICML*, 98:91–99. Citado na pág. 10

Brumme *et al.***(2008)** Zabrina L. Brumme, Mina John, Jonathan M. Carlson, Chanson J. Brumme, Mark A. Brockman, Luke C. Swenson, Iris Tao, Sharon Szeto, Pamela Rosato, Jennifer Sela, Carl M. Kadie, Nicole Frahm, Christian Brander, David W. Haas, Sharon A. Riddler, Richard Haubrich, Bruce D. Walker, P. Richard Harrigan, David Heckerman e Simon Mallal. Phylogenetic dependency networks: Inferring patterns of ctl escape and codon covariation in hiv-1 gag. *PLoS Computational Biology*, 4(11):e1000225. Citado na pág. 72

Brumme *et al.***(2009)** Zabrina L. Brumme, Mina John, Jonathan M. Carlson, Chanson J. Brumme, Dennison Chan, Mark A. Brockman, Luke C. Swenson, Iris Tao, Sharon Szeto, Pamela Rosato, Jennifer Sela, Carl M. Kadie, Nicole Frahm, Christian Brander, David W. Haas, Sharon A. Riddler, Richard Haubrich, Bruce D. Walker, P. Richard Harrigan, David Heckerman e Simon Mallal. Hla-associated immune escape pathways in hiv-1 subtype b gag, pol and nef proteins. *PLoS One*, 4(8):e6687. Citado na pág. 16, 72

Carlson *et al.***(2008)** Jonathan M. Carlson, Zabrina L. Brumme, Christine M. Rousseau, Chanson J. Brumme, Philippa Matthews, Carl Kadie, James I. Mullins, Bruce D. Walker, P. Richard Harrigan, Philip J. R. Goulder e David Heckerman. Phylogenetic dependency networks: inferring patterns of ctl escape and codon covariation in hiv-1 gag. *PLoS Computational Biology*, 4(11):e1000225. Citado na pág. 16

Cheng e Church(2000) Yizong Cheng e George M. Church. Biclustering of expression data. *Ismb*, 8: 93–103. Citado na pág. 17

Deeks(2003) Steven G Deeks. Treatment of antiretroviral-drug-resistant hiv-1 infection. *The Lancet*, 362: 2002–2011. Citado na pág. 6

Deforche *et al.***(2007)** Koen Deforche, Ricardo Jorge Camacho, Z Grossman, T Silander, M A Soares, Yves Moreau, Robert W Shafer, Kristel Van Laethemand A P Carvalho, B Wynhoven, P Cane, Joke Snoeck, J Clarke, S Sirivichayakul, K Ariyoshi, A Holguin, H Rudich, R Rodrigues, M B Bouzas, P Cahn, L F Brigido, V Soriano, W Sugiura, P Phanuphak, L Morris, J Weber, D Pillay, A Tanuri, P R Harrigan, J M Shapiro, D A Katzenstein, R Kantor e Annemie. Bayesian network analysis of resistance pathways against hiv-1 protease inhibitors. *Infection, Genetics and Evolution*, 7(3):382–390. Citado na pág. 26

Doherty *et al.***(2011)** Kathleen M. Doherty, Priyanka Nakka, Bracken M. King, Soo-Yon Rhee, Susan P. Holmes, Robert W. Shafer e Mala L. Radhakrishnan. A multifaceted analysis of hiv-1 protease multidrug resistance phenotypes. *BMC Bioinformatics*, 12(1):477. Citado na pág. 14, 15

Dupuis(1984) Claude Dupuis. Willi henning's impact on taxonomic thought. *Annual Review of Ecology and Systematics*, 15(1):1–25. Citado na pág. 7

D'Aquila(2000) Richard T D'Aquila. Limits of resistance testing. *Antvir Ther*, 5:71–6. Citado na pág. 7

Flandre *et al.***(2004)** Philippe Flandre, Diane Descamps, Véronique Joly, Vincent Meiffrédy, Catherine Tamalet, Jacques Izopet e Françoise Brun Vézinet. A survival method to estimate the time to occurrence of mutations: an application to thymidine analogue mutations in hiv-1-infected patients. *Journal of Infectious Diseases*, 189(5):862–870. Citado na pág. 71

Forgy(1965) E W Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769. Citado na pág. 17

Geretti *et al.***(2009)** Anna Maria Geretti, Linda Harrison, Hannah Green, Caroline Sabin, Teresa Hill, Esther Fearnhill, Deenan Pillay e David Dunn. Effect of hiv-1 subtype on virologic and immunologic response to start- ing highly active antiretroviral therapy. *Infect Dis*, 48(9):1296–305. Citado na pág. 7

Gifford *et al.***(2006)** Robert Gifford, Tulio de Oliveira, Andrew Rambaut, Richard E. Myers, Catherine V. Gale, David Dunn, Robert Shafer, Anne-Mieke Vandamme, Paul Kellam e Deenan Pillay. Assessment of automated genotyping protocols as tools for surveillance of hiv-1 genetic diversity. *AIDS*, 20:1521–1529. Citado na pág. 20

Gonzales *et al.***(2003)** Matthew J. Gonzales, Thomas D. Wu, Jonathan Taylor ans Ilana Belitskaya, Rami Kantor, Dennis Israelskiand Sunwen Chou, Andrew R. Zolopa, W. Jeffrey Fessel e Robert W. Shafer. Extended spectrum of hiv-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors. *AIDS*, 17(6):791. Citado na pág. 14, 26, 40, 71

Goodacre *et al.***(1998)** Royston Goodacre, Rebecca Burton, Naheed Kaderbhai, Andrew M. Woodward, Douglas B. Kell e Paul J. Rooney. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology*, 144(5):1157–1170. Citado na pág. 2, 8

Grim *et al.***(1998)** J Grim, J Novovicova, P Pudil, P Somol e F J Ferri. Initializing normal mixtures of densities. Em *Fourteenth International Conference on Pattern Recognition*, volume 1, páginas 886–890. Citado na pág. 10

Hanna *et al.***(2000)** George J. Hanna, Victoria A. Johnson, Daniel R. Kuritzkes, Douglas D. Richman, Andrew J. Leigh Brown, Anu V. Savara, J. Darren Hazelwood e T. D. Richard. Patterns of resistance mutations selected by treatment of human immunodeficiency virus type 1 infection with zidovudine, didanosine, and nevirapine. *Journal of Infectious Diseases*, 181(3):904–911. Citado na pág. 71

Hartigan e Wong(1979) John Hartigan e Manchek Wong. Algorithm as 136: A k-means clustering algorithm. *J of the Royal Statistical Society*, 28:100–108. Citado na pág. 2, 9, 10, 17, 23

Hirsch *et al.***(2000)** Martin S. Hirsch, Françoise Brun-Vézinet, Richard T. D'Aquila, Scott M. Hammer, Victoria A. Johnson, Daniel R. Kuritzkes, Clive Loveday, John W. Mellors, Bonaventura Clotet, Brian Conway, Lisa M. Demeter, Stefano Vella, Donna M. Jacobsen e Douglas D. Richman. Antiretroviral drug resistance testing in adult hiv-1 infection recommendations of an international aids society–usa panel. *JAMA*, 283:2417–26. Citado na pág. 7

Hoffman *et al.***(2003)** Noah G. Hoffman, Celia A. Schiffer e Ronald Swanstrom. Covariation of amino acid positions in hiv-1 protease. *Virology*, 314(2):536–548. Citado na pág. 14, 26, 40

Hu e Temin(1990) Wei-Shau Hu e Howard Martin Temin. Genetic consequences of packaging two rna genomes in one retroviral particle: pseudodi-ploidy and high rate of genetic recombination. Em *Proc Natl Acad Sci USA*, páginas 1556–60. Citado na pág. 6

Huang *et al.***(1997)** Xiaoqiu Huang, Mark D. Adams, Hao Zhou e Anthony R. Kerlavage. A tool for analyzing and annotating genomic sequences. *Genomics*, 46:37–45. Citado na pág. 20

Huff e Kahn(2001) Joel R Huff e James Kahn. Discovery and clinical development of hiv-1 protease inhibitors. *Advances in protein chemistry*, 56:213–251. Citado na pág. 6

Ideker *et al.***(2001)** Trey Ideker, Vesteinn Thorsson, Jeffrey A. Ranish, Rowan Christmas, Jeremy Buhler, Jimmy K. Eng, Roger Bumgarner, David R. Goodlett, Ruedi Aebersold e Leroy Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934. Citado na pág. 2, 8

Jain *et al.***(1999)** Anil K Jain, M Narasimha Murty e Patrick Joseph Flynn. Data clustering: A review. *ACM Computing Surveys*, 31:264–323. Citado na pág. 8, 9

Johnson e Wichern(1982) Richard Arnold Johnson e Dean W. Wichern. *Applied multivariate statistical analysis.* Prentice-Hall. Citado na pág. 8

Johnson *et al.***(2010)** Victoria A Johnson, Vincent Calvez, Huldrych F. Günthard, Roger Paredes, Deenan Pillay, Robert Shafer, Annemarie M. Wensing e Douglas D. Richman. 2011 update of the drug resistance mutations in hiv-1. *HIV Med*, 18:156–163. Citado na pág. 1, 6

Kanki *et al.***(1999)** Phyllis J. Kanki, Donald J. Hamel, Jean-Louis Sankalé, Chung cheng Hsieh, Ibou Thior, Francis Barin, Stephen A. Woodcock, Aïssatou Guèye-Ndiaye, Er Zhang, Monty Montano, Tidiane Siby, Richard Marlink, Ibrahima NDoye, Myron E. Essex e Souleymane MBoup. Human immunodeficiency virus t 1 subtypes differ in disease progression. *J Infect Dis.*, 179:68–73. Citado na pág. 6

Kaplan *et al.***(2004)** Noam Kaplan, Moriah Friedlich, Menachem Fromer e Michal Linial. A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics*, 5(1):196. Citado na pág. 2, 8

Kluger *et al.***(2003)** Yuval Kluger, Ronen Basri, Joseph T. Chang e Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4):703–716. Citado na pág. 2, 8, 17

Korber *et al.***(1993)** Bette T Korber, Robert M Farber, David H Wolpert e Alan S Lapedes. Covariation of mutations in the v3 loop of human immunodeficiency virus t 1 envelope protein: an information theoretic analysis. Em *Proc Natl Acad Sci*, volume 90, páginas 7176–80. Citado na pág. 16

Krasnogor e Pelta.(2004) Natalio Krasnogor e David A. Pelta. Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, 20(7):1015–1021. Citado na pág. 2, 8

Kriegel *et al.***(2009)** Hans-Peter Kriegel, Peter Kroger e Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans Knowl Discov Data*, 3:1–58. Citado na pág. 2, 10, 11, 21

Laethem *et al.***(1999)** K Van Laethem, K Van Vaerenbergh, J C Schmit, S Sprecher, P Hermans, V De Vroey, R Schuurman, T Harrer, M Witvrouw, E Van Wijngaerden, L Stuyver, M Van Ranst, J Desmyter, E De Clercq e A M Vandamme. Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed hiv-1 genotypic populations. *Journal of acquired immune deficiency syndromes*, 22(2):107–118. Citado na pág. 7

Laeyendecker *et al.***(2006)** Oliver Laeyendecker, X Li, Miguel Arroyo, Francine McCutchan, Ronald Gray, Maria Wawer, David Serwadda, F Nalugoda, Godfrey Kigozi, Thomas Quinn e Rakai Health Science Program. The effect of hiv subtype on rapid disease progression in rakai. Em *13th Conference on Retroviruses and Opportunistic Infections*, página abstract no. 44LB. Citado na pág. 6

Lathrop *et al.***(1998)** Richard H. Lathrop, Nicholas R. Steffen, Miriam P. Raphael, Sophia Deeds-Rubin, Michael J. Pazzani, Paul J. Cimoch, Darryl M. See e Jeremiah G. Tilles. Knowledge-based avoidance of drug-resistant hiv mutants. Em *Proceedings 15th Conference On Innovative Applications of Artificial Intelligence*, páginas 1071–1078, Menlo Park California. AAAI Press. Citado na pág. 2, 7

Lazzeroni e Owen(2002) Laura Lazzeroni e Art Owen. Plaid models for gene expression data. *Statistica sinica*, (1):61–86. Citado na pág. 17

Levy *et al.***(2004)** David N. Levy, Grace M Aldrovandi, Olaf Kutsch e George M. Shaw. Dynamics of hiv-1 recombination in its natural target cells. Em *Proc Natl Acad Sci USA*, páginas 4204–9. Citado na pág. 6

Liu e Wang(2003) Jinze Liu e Wei Wang. Op-cluster: Clustering by tendency in high dimensional space. Em *Proc. Third IEEE Int'l Conf*, páginas 187–194. Citado na pág. 2, 8

Liu *et al.***(2008)** Ying Liu, Eran Eyal e Ivet Bahar. Analysis of correlated mutations in hiv-1 protease using spectral clustering. *Bioiformatics*, 24(10):1243–1250. Citado na pág. 14, 15, 26

Lloyd(1982) Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137. Citado na pág. 2, 9, 17

losAlamos() losAlamos. The los alamos resistance database. URL http://hiv-web.lanl.gov. Citado na pág. 2, 7

MacQueen(1967) James MacQueen. Some methods for classification and analysis of multivariate observations. Em *Berkeley symposium on mathematical statistics and probability.*, volume 1, página 14. Citado na pág. 2, 9, 17

Madeira e Oliveira(2004) Sara C. Madeira e Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions*, 1:24–45. Citado na pág. 2, 8

Mansky(1998) Louis M Mansky. Retrovirus mutation rates abd their role in genetic variation. *J of General Virology*, 79:1337–45. Citado na pág. 1, 6

Marcelin *et al.***(2004)** Anne-Geneviève Marcelin, Constance Delaugerre, Marc Wirden, Pedro Viegas, Anne Simon, Christine Katlama e Vincent Calvez. Thymidine analogue reverse transcriptase inhibitors resistance mutations profiles and association to other nucleoside reverse transcriptase inhibitors resistance mutations observed in the context of virological failure. *Journal of Medical Virology*, 72(1):162–165. Citado na pág. 71

Mechelen *et al.***(2004a)** Iven Van Mechelen, Hans-Hermann Bock e Paul De Boeck. Two-mode clustering methods: a structured overview. *Statistical methods in medical research*, 13(5):363–394. Citado na pág. 2

Mechelen *et al.***(2004b)** Iven Van Mechelen, Hans-Hermann Bock e Paul De Boeck. Two-mode clustering methods:a structured overview. *Statistical Methods in Medical Research*, 413:363–94. Citado na pág. 10

Moore(1999) Andrew W Moore. Very fast em-based mixture model clustering using multiresolution kd-trees. *Advances in Neural information processing systems*, páginas 543–549. Citado na pág. 10

Murali e Kasif(2003) T M Murali e Simon Kasif. Extracting conserved gene expression motifs from gene expression data. Em *Pacific Symposium on Biocomputing*, volume 8, páginas 77–88. Citado na pág. 2, 8, 17

Muzammil *et al.***(2003)** Salman Muzammil, Patrick Ross e Ernesto Freire. A major role for a set of non-active site mutations in the development of hiv-1 protease drug resistance. *Biochemistry*, 42(3):631–638. Citado na pág. 7

Peeters(2003) René Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 131(3):651–654. Citado na pág. 11

Prelic *et al.***(2006)** Amela Prelic, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele e Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22:1122–1129. Citado na pág. xi, 11, 12, 13, 17

Preston *et al.***(1988)** Bradley D Preston, Bernard J Poiesz e Lawrence A Loeb. Fidelity of hiv-1 reverse transcriptase. *Science*, 242:1168–71. Citado na pág. 6

Quackenbush(2001) John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427. Citado na pág. 2, 8

Reuman *et al.***(2010)** Elizabeth C. Reuman, Soo-Yon Rhee, Susan P. Holmes e Robert W. Shafer. Constrained patterns of covariation and clustering of hiv-1 non-nucleoside reverse transcriptase inhibitor resistance mutations. *Journal of Antimicrobial Chemotherapy*, 65(7):1477–1485. Citado na pág. 14, 15, 21

Rhee *et al.***(2003)** Soo-Yon Rhee, Matthew J. Gonzales, Rami Kantor, Bradley J Betts, Jaideep Ravela e Robert W Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31(1):298–303. Citado na pág. 7, 14, 71

Rhee *et al.***(2004)** Soo-Yon Rhee, Tommy Liu, Jaideep Ravela, Matthew J. Gonzales e Robert W. Shafer. Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. *Antimicrobial Agents and Chemotherapy*, 48(8): 3122–3126. Citado na pág. 26

Rhee *et al.***(2009)** Soo-Yon Rhee, Walford Jeffrey Fessel, Tommy F. Liu, Natalia M. Marlowe, Charles M. Rowland, Richard A. Rode, Kristel Van Laethem Anne-Mieke Vandamme, Francoise Brun-Vezinet, Vincent Calvez, Jonathan Taylor, Leo Hurley, Michael Horberg e Robert W Shafer. Predictive value of hiv-1 genotypic resistance test interpretation algorithms. *The Journal of Infectious Diseases*, 200(3):453–63. Citado na pág. 8

Robertson *et al.***(2000)** D.L. Robertson, J.P. Anderson, J.A. Bradac, J.K. Carr, B. Foley, R.K. Funkhouser, F. Gao, B.H. Hahn, M.L. Kalish, C. Kuiken, G.H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, e B. Korber. Hiv-1 nomenclature proposal. a reference guide to hiv-1 classification. *Science*, 288(5463):55–6. Citado na pág. 7

Scherrer *et al.***(2011)** Alexandra U. Scherrer, Bruno Ledergerber, Viktor von Wyl, Jurg Boni, Sabine Yerly, Thomas Klimkait, Philippe Burgisser, Andri Rauch, Bernard Hirschel, Matthias Cavassini, Luigia Elzi, Pietro L. Vernazza, Enos Bernasconi, Leonhard Held, Huldrych F. Gunthard, e the Swiss HIV Cohort Study. Improved virological outcome in white patients infected with hiv-1 non-b subtypes compared to subtype b. *Clin Infect Dis*, 53(11):1143–52. Citado na pág. 7

Schinazi *et al.***(2000)** Raymond F. Schinazi, Brendan Larder e John W. Mellors. Mutations in retroviral genes associated with drug resistance:2000-2001 update. *Int'l AntiviralNews*, 8:65–92. Citado na pág. 2

Segal *et al.***(2001)** Eran Segal, Ben Taskar, Audrey Gasch, Nir Friedman e Daphne Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17:S243–S252. Citado na pág. 2, 8

Segal *et al.***(2003)** Eran Segal, Alexis Battle e Daphne Koller. Decomposing gene expression into cellular processes. Em *Pacific Symposium on Biocomputing*, volume 8, páginas 89–100. Citado na pág. 2, 8

Shafer *et al.***(2000a)** Robert W. Shafer, Duane R. Jung e Bradley J. Betts. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nature Medicine*, 6 (11):1290–1292. Citado na pág. 2

Shafer *et al.***(2000b)** Robert W. Shafer, Rami Kantor e Matthew J. Gonzales. The genetic basis of hiv-1 resistance to reverse transcriptase and protease inhibitors. *AIDS Rev*, 2:211–218. Citado na pág. 1, 6, 7, 26

Sheng *et al.***(2003)** Qizheng Sheng, Yves Moreau e Bart De Moor. Biclustering microarray data by gibbs sampling. *Bioinformatics*, 19(suppl 2):ii196–ii205. Citado na pág. 2, 8

Sing *et al.***(2005)** Tobias Sing, Valentina Svicher, Niko Beerenwinkel, Francesca Ceccherini-Silberstein, Martin Däumer, Rolf Kaisera, Hauke Tobias, Klaus Korn, Daniel Hoffmann, Mark Oette, Jurgen K. Rockstroh, Gert Fatkenheuer, Carlo-Federico Perno e Thomas Lengauer. Characterization of novel hiv drug resistance mutations using clustering, multidimensional scaling and svm-based feature ranking. *In Knowledge Discovery in Databases: PKDD*, páginas 285–296. Citado na pág. 14, 15, 22, 26, 40, 71

Slonim(2002) Donna K Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32(2002):502–508. Citado na pág. 2, 8

Sparks(1973) D. N. Sparks. Algorithm as 58: Euclidean cluster analysis. *Journal of the Royal Statistical Society.*, 22(1):126–130. Citado na pág. 10

Suzuki *et al.***(2010)** David T. Suzuki, Anthony JF Griffiths, Jeffrey H. Miller e Richard C. Lewontin. *An introduction to genetic analysis.* WH Freeman and Company, 10º edição. Citado na pág. xi, 5

Tanay *et al.***(2002)** Amos Tanay, Roded Sharan e Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 718(suppl 1):S136–S144. Citado na pág. 2, 8

Tanay *et al.***(2003)** Amos Tanay, Roded Sharan e Ron Shamir. Biclustering algorithms: A survey. Em *Handbook of Computational Molecular Biology.* in press. Citado na pág. 10

Tang e Shafer(2012) Michele W Tang e Robert W Shafer. Hiv-1 antiretroviral resistance: scientific principles and clinical applications. *Drugs*, 72(9):e1–25. Citado na pág. 7

Tavazoie *et al.***(1999)** Saeed Tavazoie, Jason D. Hughes, Michael J. Campbell, Raymond J. Cho e George M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285. Citado na pág. 2, 8

Taylor *et al.***(2008)** Barbara S. Taylor, Magdalena E. Sobieszczyk, Francine E. McCutchan, e Scott M. Hammer. The challenge of hiv-1 subtype diversity. *N Engl J Med*, 358(15):1590–1602. Citado na pág. 7

Team(2008) R Development Core Team. R: A language and environment for statistical computing. http://www.R-project.org, 2008. Citado na pág. 17, 23, 24

Telgarsky e Vattani(2010) Matus Telgarsky e Andrea Vattani. Hartigan's method: k-means clustering without voronoi. Em *International Conference on Artificial Intelligence and Statistics.*, volume 9, páginas 820–827. Citado na pág. 10

Williams *et al.***(2012)** Thomas Williams, Colin Kelley, Hans-Bernhard Broker, John Campbell, Robert Cunningham, David Denholm, Gershon Elber, Roger Fearick, Carsten Grammes, Lucas Hart, Lars Hecking, Peter Juhasz, Thomas Koenig, David Kotz, Ed Kubaitis, Russell Lang, Timothee Lecomte, Alexander Lehmann, Alexander Mai, Bastian Markisch, Ethan A Merritt, Petr Mikulık, Carsten Steger, Shigeharu Takeno, Tom Tkacik, Jos Van der Woude, James R. Van Zandt, Alex Woo e Johannes Zellner. Gnuplot 4.6: an interactive plotting program. http://gnuplot.sourceforge.net/, March 2012. Citado na pág. 18

Wolf *et al.***(2003)** Katharina Wolf, Hauke Walter, Niko Beerenwinkel, Wilco Keulen, Rolf Kaiser, Daniel Hoffmann, Thomas Lengauer, Joachim Selbig, Anne-Mieke Vandamme, Klaus Korn e Barbara Schmidt. Tenofovir resistance and resensitization. *Antimicrobial Agents and Chemotherapy*, 47(8):4836–4847. Citado na pág. 71

Wu *et al.***(2003)** Thomas D. Wu, Celia A. Schiffer, Matthew J. Gonzales, Jonathan Taylor, Rami Kantor, Sunwen Chou, Dennis Israelski, Andrew R. Zolopa, W. Jeffrey Fessel e Robert W. Shafer. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *Journal of Virology*, 77(8):4836–4847. Citado na pág. 14, 16, 26, 40

Yahi *et al.***(1999)** Nouara Yahi, Catherine Tamalet, Christian Tourrès, Natacha Tivoli, Franck Ariasi, Françoise Volot, Jean-Albert Gastaut, Hervé Gallais, Jacques Moreau e Jacques Fantini. Mutation patterns of the reverse transcriptase and protease genes in human immunodeficiency virus type 1-infected patients undergoing combination therapy: survey of 787 sequences. *Journal of Clinical Microbiology*, 37(12):4099–4106. Citado na pág. 14, 26, 40, 71

Yahi *et al.***(2000)** Nouara Yahi, Catherine Tamalet, Christian Tourrès, Natacha Tivoli e Jacques Fantini. Mutation l210w of hiv-1 reverse transcriptase in patients receiving combination therapy. *Journal of Biomedical Science*, 7(6):507–513. Citado na pág. 71

Yang *et al.***(2003)** Jiong Yang, Haixun Wang, Wei Wang e Philip Yu. Enhanced biclustering on expression data. Em *Third IEEE Symposium*, volume 9, páginas 321–327. Citado na pág. 2, 8