Classificação das mutações de vírus HIV

MAC0499 - Trabalho de Formatura Supervisionado

Aluna: Mina Cintho nºUSP 3746752

Orientador: João Eduardo Ferreira

Sumário

1 Introdução **3**

2 Fundamentos Biológicos **5**

2.1 Mutações 5

2.2 HIV e Resistência 5

2.3 Análise das sequências 7

3 Fundamentos de Reconhecimento de padrões **8**

3.1 Agrupamento 8

3.2 Métodos Hierárquicos 11

3.2.1 Métodos Aglomerativos Hierárquicos 12

3.3 Métodos Não-Hierárquicos 15

3.3.1 Método K-Médias 15

3.4 Conclusão 17

4 Atividades Realizadas **17**

4.1 Representação e manipulação dos dados 17

4.2 Medida de Dissimilaridade 18

4.3 Implementação do método Vizinho Mais Próximo 19

4.4 Implementação do método K-Médias 34

5 Conclusão **36**

6 Agradecimentos **38**

7 Parte Subjetiva **39**

8 Referências bibliográficas **41**

1 Introdução

Avanços nos estudos na área de bioinformática têm gerado uma grande quantidade de dados. Esses dados escondem muitas informações e conhecimento que não são facilmente extraídos devido ao grande volume de variáveis.

Um exemplo desse problema é a classificação de vírus em pacientes infectados com o HIV. O conhecimento e tecnologias atuais permitem que sejam obtidas grandes quantidades de sequências genéticas e de proteínas do vírus, sendo essas sequências importantes por possuírem informações relacionadas a características específicas dos vírus de cada paciente, como a possível resistência do vírus a uma determinada droga e sua suscetibilidade a outras.

Uma ferramenta capaz de detectar a relação entre as sequências de aminoácidos e a resistência a uma determinada droga, ou a uma combinação delas, seria bastante útil na tomada de decisão sobre o tratamento de um paciente. Essa ferramenta auxiliaria e automatizaria a classificação dos vírus, tornando mais rápida e validando a classificação realizada pelo médico. Partindo do conhecimento empírico, as sequências seriam analisadas através de um algoritmo que devolveria sua classificação.

Até o presente momento, essas sequências são interpretadas manualmente pelos especialistas, que contam com seu conhecimento e experiência para decidir quais drogas são mais adequadas para um paciente, ou seja, a quais drogas os vírus do paciente provavelmente são suscetíveis ou não e qual tratamento deve ser o mais eficaz no combate à doença.

Com o intuito de gerar uma classificação automática, baseado em técnicas de reconhecimento de padrões, este trabalho aplica métodos de agrupamento

em dados de pacientes infectados com o vírus HIV em tratamento, realizando a extração de informações importantes desses dados, como a presença ou ausência de padrões de mutações nos vírus. Essas informações podem contribuir para o desenvolvimento de estudos sobre o HIV e no estabelecimento de tratamentos contra a doença, bem como a classificação automatizada dos vírus.

Inicialmente são expostos os fundamentos biológicos e de reconhecimento de padrões necessários para a exploração e estudo do tema, seguidos da proposta de abordagem para o problema, os resultados, a análise dos resultados e conclusão do trabalho.

2 Fundamentos Biológicos

2.1 Mutações

Mutações são alterações das sequências de nucleotídeos do DNA, podendo ocorrer inserções, remoções ou substituições desses nucleotídeos. As mutações são permanentes, ou seja, são propagadas pelo processo de multiplicação do DNA, no qual são feitas cópias a partir das sequências originais, e podem gerar alterações na codificação de proteínas. As alterações se devem ao fato de que as sequências de nucleotídeos do DNA contêm a informação necessária para a determinação dos aminoácidos das proteínas e quando os nucleotídeos são modificados, os aminoácidos também podem ser modificados.

A sequência de aminoácidos de uma proteína interfere na interação entre ligações e posicionamento dos aminoácidos e quando ocorrem mutações é possível que aconteçam modificações na estrutura e consequentemente na função das proteínas, transformando sua atuação no metabolismo, já que as funções das proteínas estão altamente ligadas a sua estrutura tridimensional e as interações dessas com outras moléculas.

2.2 HIV e Resistência

O HIV, vírus da imunodeficiência humana, está suscetível às mutações que geram diversidade de sequências trazendo a variabilidade genética. Essa variabilidade possibilita a classificação dos vírus em tipos, subtipos e grupos [1]. Estudos têm sido realizados no sentido de verificar possíveis relações existentes entre essas classificações e a capacidade de transmissão, patogeni-

cidade e resposta a tratamentos [2-4].

Com a utilização de agentes antivirais no tratamento de doentes há seleção de mutantes resistentes à ação de drogas que fazem com que o tratamento seja ineficiente. Assim, os vírus resistentes prevalecem, e não há resposta ao tratamento. Porém, ainda pode haver outras drogas existentes às quais os vírus são suscetíveis. Se o médico tiver acesso a essa informação o tratamento pode ser realizado de forma mais específica, resultando em um tratamento muito mais eficaz.

A verificação e entendimento da possível ligação entre o material genético do vírus e resistência às drogas é uma informação de grande importância quando considerados os tratamentos antiretrovirais em pessoas infectadas. A resposta a tratamentos e a resistência de alguns vírus a certas drogas são um dos maiores obstáculos à supressão do HIV durante a *highly active antiretroviral therapy (HAART)* [5-11], tratamento que utiliza vários medicamentos em combinação. O surgimento de variantes resistentes às drogas tem limitado a efetividade a longo prazo e com o estabelecimento da relação entre resistência às drogas e tipos, subtipos ou grupos de HIV seriam possíveis tratamentos personalizados e mais eficientes.

Os médicos têm utilizado sua experiência no tratamento de pacientes para definir quais medicamentos cada indivíduo deve tomar. No momento não há estudos que utilizem métodos de análise de dados para verificar se há ou não grupos de mutações que ocorrem juntos, ou seja, que verifiquem a existência de tipos e subtipos entre os vírus, ou que verifiquem numericamente a correlação de mutações e resistência às drogas. As predições são dadas majoritariamente pela intuição do médico. Portanto, é importante que sejam

desvendadas as relações entre mutações e resistência baseando-se em métodos de análise de dados.

A terapia contra o vírus da AIDS atualmente é voltada para a inibição da transcriptase reversa (RT) e da protease (PR) que são de extrema importância para o vírus em sua replicação. A transcriptase reversa é utilizada para produção de DNA a partir do RNA do vírus que irá então se incorporar ao DNA da célula hospedeira. Já a protease é responsável pela clivagem de proteínas, gerando proteínas maduras que estarão presentes no vírion, partícula viral completa.

Por análises genéticas de vírus resistentes foi identificado um grande número de mutações nesses genes. A transcriptase reversa é uma das maiores responsáveis pela taxa de mutação ou variabilidade genética do HIV [12]. A alta taxa de erros na transcriptase reversa, 1 em 10.000 bases, e grande velocidade de replicação do vírus, 108-109 virions (partícula viral completa) por dia, favorecem o acontecimento de mutações e a seleção de vírus resistentes.

2.3 Análise das sequências

Para análise da presença ou não de mutações nas proteínas transcriptase reversa e protease de vírus são realizadas comparações com sequências já conhecidas e estudadas como a cepa HXB2 (GenBank Accession Number K03455) que é utilizada como padrão [16]. Essa comparação é feita pelo alinhamento das sequências, método no qual há comparação dos aminoácidos e computação de pontos para similaridades e dissimilaridades e o resultado é o alinhamento com maior valor de pontos. Do alinhamento podemos inferir

quais mutações estão presentes em cada vírus, sendo possível identificá-las em cada um dos pacientes.

3 Fundamentos de Reconhecimento de padrões

3.1 Agrupamento

A partir da obtenção das sequências de aminoácidos, do alinhamento e análise da presença ou não de mutações é gerada uma grande quantidade de dados contendo grande volume de informação. Para que essa informação seja extraída é necessária a realização da análise de dados, ou seja, a união, modelagem e transformação a fim de destacar a informação contida nesses dados.

A análise de dados é utilizada em diversos campos da ciência como a biologia, a computação e a física e seu objetivo é encontrar características importantes dos dados distinguindo as informações da aleatoriedade. Pode ser dividida em exploratória, em que há formulação de hipóteses e tomada de decisões, ou confirmatória, em que há validação de modelos [17]. Dentro da análise exploratória de dados, a técnica de agrupamento é empregada em casos nos quais não há grande quantidade de informação prévia e existem poucas hipóteses sobre os dados.

O agrupamento é a organização de uma coleção de padrões (dados), geralmente representados como vetores de medidas ou pontos no espaço multidimensional, em grupos baseados em similaridade [17]. Assim, 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.

Essa técnica em que se procura agrupar dados quando ainda não se tem grupos estabelecidos é chamada de classificação não-supervisionada. Na classificação supervisionada os padrões já estão contidos em grupos pré-estabelecidos e o objetivo é inserir novos elementos que ainda não estão agrupados.

Os grupos ou *clusters* resultantes do processo devem possuir padrões com propriedades em comum, ou seja, similares, sendo que padrões de um grupo devem ser mais similares a padrões do mesmo grupo do que de grupos distintos. Assim, a construção dos grupos pode ajudar na identificação de *outliers* e na sugestão de hipóteses de relacionamentos entre os dados e sua estrutura, 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 [18]. Essas escolhas influenciam na disposição dos dados e, consequentemente, podem influenciar nas formas dos grupos.

A medida de similaridade pode ser feita pela representação na forma de coeficientes de correlação, medidas de associação, como, por exemplo,

frequências ou na forma de medidas de distâncias. O cálculo de distâncias, que é frequentemente usado, pode ser realizado de várias maneiras:

Sendo x e y pontos n dimensionais:

- Distância de Manhattan: $D(x,y) = |x_1 - y_2| + |x_2 - y_2| + ... + |x_n - y_n|$
- Ângulo entre dois vetores: $\theta = arccos(\vec{x}.\vec{y}/|\vec{x}||\vec{y}|)$
- Distância Malahanobis: $D(x,y) = \sqrt{(\vec{x} - \vec{y})\Sigma^{-1}(\vec{x} - \vec{y})}$, sendo Σ^{-1} a matriz de covariância
- Medida Euclideana:

$$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}.$$

$$D(x,y) = \sqrt{(x - y)'(x - y)}.$$

Além das diferentes representações dos cálculos de medidas de distâncias também existem diferentes formas e técnicas para a criação dos agrupamentos. Isso porque inerente à técnica do agrupamento há o problema da existência do grande número de combinações possíveis para a formação dos grupos, mesmo quando um pequeno número de padrões é considerado. Dessa forma, não é possível simular todas as combinações possíveis para a escolha de uma delas e é preciso utilizar outras técnicas.

Os métodos de agrupamento podem então ser divididos em hierárquicos, e não-hierárquicos. Métodos hierárquicos geram grupos em uma relação que estabelece uma hierarquia entre os padrões e que pode ser representada na forma de uma árvore chamada de dendrograma, como na Figura 1. Já os

métodos não-hierárquicos geram grupos sem relação de hierarquia, com uma única partição (Figura 2).

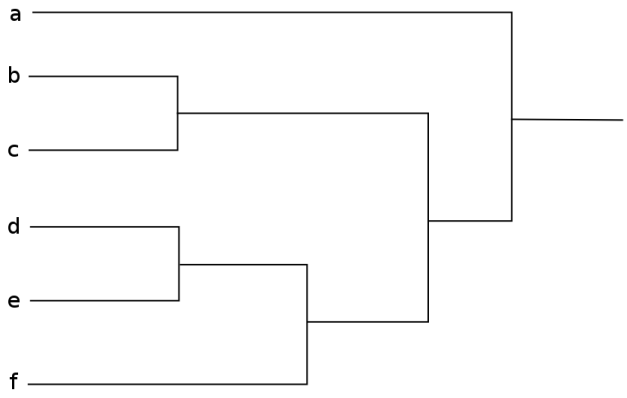

Figura 1: Exemplo de dendrograma

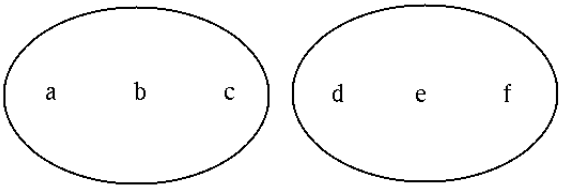

Figura 2: Exemplo de agrupamento não-hierárquico

3.2 Métodos Hierárquicos

Dentro dos métodos hierárquicos temos os aglomerativos e os divisivos. Os aglomerativos iniciam com cada padrão pertencendo a grupos distintos que são continuamente unidos até que se tenha um único grupo contendo todos os elementos. Já os divisivos têm, no início, um único grupo contendo

todos os padrões, e esse grupo é dividido em grupos menores até que se tenha o mesmo número de elementos e grupos.

As divisões ou uniões de agrupamentos em ambos os casos são realizadas de acordo com as medidas de similaridades dos elementos. No caso de métodos aglomerativos, inicialmente são agrupados os dois elementos mais similares ou menos distantes, então são recalculadas as medidas como se esses elementos fossem um único e então novamente são unidos os dois mais similares ou menos distantes e até que se tenha um grupo único. Nos métodos divisivos, o grupo é dividido em dois subgrupos de forma que os novos subgrupos sejam o menos similar entre si possível, são recalculadas as similaridades ou distâncias, divididos os grupos novamente e assim por diante até que se tenham as mesmas quantidades de grupos e elementos.

Os passos de junções realizados pelos algoritmos hierárquicos podem ser visualizados nos dendrogramas gerados. Os "galhos" representam os grupos que se unem nos nós em diferentes níveis de fusão que indicam as similaridades ou distâncias, como podemos verificar na Figura 1 em que são unidos {b, c} e {d,e} inicialmente, então em um nível acima o grupo {d,e} é unido a {f}, para depois ser unido a {b,c} que em somente um nível mais acima é unido a {a}.

3.2.1 Métodos Aglomerativos Hierárquicos

Os métodos aglomerativos hierárquicos são representados basicamente pelo Vizinho Mais Próximo (*single linkage*), o Vizinho Mais Distante (*complete linkage*) e o método da ligação média não ponderada (*average linkage* ou UPGMA (*Unweighted Pair Group Method with Arithmetic mean*)). Nos

três métodos só se distingue a maneira como é calculada a distância entre os grupos quando é realizada uma junção. No Vizinho Mais Próximo a distância entre dois grupos é a menor das distâncias entre um elemento de um grupo e um elemento do outro grupo. No método Vizinho Mais Distante a distância entre dois grupos é dada pela maior das distâncias entre um elemento de um grupo e um elemento do outro grupo. Já no método da ligação média não ponderada as distâncias são dadas pelas distâncias médias entre os pares de elementos dos grupos.

O algoritmo geral para métodos aglomerativos hierárquicos é dado por [18]:

1. Comece com N grupos, cada um contendo uma única entidade e uma matriz simétrica NxN de distâncias (ou similaridades) $D = d_{ik}$.

2. Procure pelo par de grupos mais próximos na matriz de distâncias (similaridades). Seja d_{uv} a distância entre esses grupos U e V.

3. Una os grupos U e V. Nomeie o novo grupo (UV). Atualize as entradas da matriz apagando as linhas e colunas correspondentes aos grupos U e V e adicionando uma linha e uma coluna com as distâncias entre o grupo (UV) e os outros grupos.

4. Repita os passos 2 e 3 por $N-1$ vezes. (Todos os objetos estarão em um único grupo ao final do algoritmo). Guarde a identidade dos grupos que são unidos e os níveis (distâncias ou similaridades) nos quais as uniões são realizadas.

Os algoritmos aglomerativos hierárquicos são similares, no entanto não

produzem resultados idênticos quando utilizados com os mesmos dados e medidas de similaridades ou distâncias. O algoritmo do Vizinho Mais Próximo, por exemplo, não consegue distinguir grupos próximos porque se utiliza da menor distância para uni-los. Porém, diferentemente da maioria dos outros métodos, consegue construir grupos de formatos não-elípticos, pois tem tendência a reconhecer grupos alongados, conhecidos como *chaining*. O *chaining* pode influenciar no agrupamento uma vez que seus elementos das extremidades podem ser bastante distantes ou pouco similares.

Em oposição ao algoritmo Vizinho Mais Próximo que identifica grupos alongados [19], o algoritmo Vizinho Mais Distante identifica grupos fortemente ligados ou compactos [20].

Além dessas diferenças, os métodos Vizinho Mais Próximo e Vizinho Mais Distante não têm seus resultados alterados quando as distâncias ou similaridades têm seus valores alterados mantendo-se a ordem relativas dos dados, ao contrário do que ocorre no método da Ligação Média Não Ponderada em que o resultado é alterado.

Esses algoritmos não consideram fontes de variações ou erros, como *outliers*, podendo ter a determinação dos grupos influenciada por esses dados. Caso ocorra um agrupamento incorreto nas etapas iniciais do algoritmo, não há correção e se faz necessário um exame cuidadoso dos grupos gerados. Para se obter maior segurança quanto aos resultados, diversos algoritmos podem ser testados, bem como formas de cálculo de distâncias ou similaridades, sendo observadas as consistências das informações obtidas. Ainda, a estabilidade pode ser testada inserindo-se pequenas perturbações aos dados, já que, se bem distintos, os grupos não devem ser alterados.

3.3 Métodos Não-Hierárquicos

Algoritmos de agrupamento não-hierárquicos obtêm uma única partição dos dados ao invés de dendrogramas como nos algoritmos hierárquicos [17]. Nesse método é necessária a definição, antecipada ou durante o processo, do número de grupos que se deseja ter no final. Além disso, pelo fato de que não é necessária a utilização da matriz de distâncias e nem a sua manipulação e armazenamento, métodos não hierárquicos podem ser utilizados para dados muito mais numerosos do que métodos hierárquicos.

Métodos não-hierárquicos iniciam de uma partição inicial de itens em grupos ou de um conjunto inicial de *seed points* que irão formar os núcleos dos grupos que devem ser escolhidos de forma não tendenciosa, aleatoriamente.

3.3.1 Método K-Médias

Um dos algoritmos mais populares e comuns entre os não-hierárquicos é o K-Médias, ou *K-Means*, que une os padrões de acordo com a maior proximidade ao centro, também chamado de centróide, de um grupo. Para tanto, o centro é definido como sendo a média aritmética de cada uma das dimensões das distâncias de todos os pontos do grupo. Os centróides são definidos aleatoriamente no início ou são criados tantos grupos aleatórios quanto se queira.

O algoritmo geral para o método K-Médias é dado por [18]:

1. Particione os padrões em K grupos iniciais

2. Siga pela lista de itens, inserindo-os ao grupo cujo centróide é o mais próximo. Recalcule o centróide do grupo recebendo o novo item e

retirando itens removidos.

3. Repita o passo 2 até que não haja mais inserções a serem feitas

Para testar a estabilidade dos resultados, pode-se executar repetidamente o algoritmo com partições iniciais distintas. Além disso, as informações obtidas em um primeiro resultado podem ser utilizadas para rearranjar os elementos em ordem de acordo com os agrupamentos hipotéticos do primeiro resultado. Outra possibilidade é criar uma tabela com centróides e variâncias entre grupos.

O K-Médias é fácil de ser implementado e possui complexidade O(n), sendo n o número de padrões. Porém, é possível perceber que os grupos obtidos pelo método são dependentes das escolhas iniciais dos grupos ou dos centróides. Como consequência disso, muitas variantes do algoritmo foram criadas, inclusive na tentativa de escolher uma boa partição inicial.

Uma forma de variante fornece a possibilidade de se separar ou unir os grupos resultantes de acordo com distâncias limites pré-determinadas. Assim, se dois grupos estão mais próximos do que certa distância eles são unidos, e caso um grupo tenha elementos mais distantes do que certo valor, ele será dividido.

Uma segunda desvantagem dos algoritmos não-hierárquicos é a fixação do número de grupos, que pode interferir caso sejam escolhidos como centróides dois elementos que devem pertencer a um mesmo grupo ou caso haja um *outlier* que pode criar um grupo com itens dispersos. Além disso, mesmo que haja exatamente K grupos, os dados do menor grupo podem não ser representativos o suficiente para se conseguir representar o grupo e um novo

grupo artificial pode ser forçadamente criado.

3.4 Conclusão

Algoritmos hierárquicos são mais versáteis do que não-hierárquicos. Por exemplo, o Vizinho Mais Próximo funciona bem em dados não-isotrópicos (não uniformes em todas as direções) com grupos bem separados, em forma de cadeia e concêntricos. Já algoritmos como o K-Médias funcionam bem com dados isotrópicos [21]. No entanto, o tempo e complexidade dos algoritmos não-hierárquicos são menores [22].

Em aplicações em dados biológicos não há experimentos que consolidem o uso de um método específico para agrupamento. Dessa forma, o presente trabalho aplica diferentes métodos, analisando seus resultados para verificar qual o mais adequado para esse tipo específico de dados.

4 Atividades Realizadas

4.1 Representação e manipulação dos dados

Numa primeira análise, foram utilizados 14393 dados de pacientes infectados com o vírus HIV contendo sua identificação, classificação do subtipo do vírus e sequências parciais da transcriptase reversa e protease. As sequências de protease e transcriptase reversa tiveram posições pré-selecionadas de acordo com estudos que caracterizam algumas posições nas sequências de aminoácidos dessas proteínas como sendo chaves para a classificação do vírus (38 posições da transcriptase reversa e 44 posições da protease).

Numa segunda análise, 13213 sequências completas de transcriptase reversa (335 posições) e protease (99 posições) foram utilizadas com identificação e classificação de subtipo do vírus.

Todas as sequências continham para cada posição os valores # caso a posição contivesse o mesmo aminoácido da sequência padrão ou a sigla de um aminoácido caso possuísse outro aminoácido diferente da sequência padrão.

Com intuito de facilitar a análise de dados, foi empregada a representação binária para as sequências de transcriptase reversa e protease de forma que quando em uma posição havia a sigla de um aminoácido resultante de mutação essa posição era representada pelo valor 1 e quando a posição possuía um aminoácido igual ao da sequência padrão esse símbolo era trocado pelo valor 0. Assim, os dados podem ser interpretados como vetores em um espaço N dimensional, com N variando com o tamanho das sequências de aminoácidos, nos quais as coordenadas valem 0 ou 1.

4.2 Medida de Dissimilaridade

A medida de dissimilaridade aplicada nos dados foi a medida Euclideana, sendo que as sequências eram comparadas duas a duas em cada uma das posições sendo atribuídos os valores:

$$D(x,y) = \sum_{j=1}^{n} (x_j - y_j)^2 = \begin{cases} 0 & \text{caso} \quad x_j = y_j = 1 \quad ou \quad x_j = y_j = 0 \\ 1 & \text{caso} \quad x_j \neq y_j \end{cases}$$

tal que x e y são sequências binárias de dados N dimensionais, para j va-

lendo de 1 a N. As dissimilaridades entre as sequências são então dadas pelo somatório dos valores atribuídos a cada uma das comparações nas posições.

4.3 Implementação do método Vizinho Mais Próximo

Dentre os métodos de agrupamento hierárquicos foi escolhido o Vizinho Mais Próximo para implementação. Como toda informação necessária para o agrupamento com o método Vizinho Mais Próximo está na MST construída a partir dos dados e algoritmos que encontram a MST são eficientes para tal propósito [21], o Vizinho Mais Próximo foi implementado como a busca pela MST do algoritmo de Kruskal com o armazenamento das informações de inserção dos vértices na MST (ordem de entrada, aresta utilizada na inserção e custo da aresta).

No algoritmo as arestas do grafo representam as dissimilaridades entre os dados que são representados pelos vértices. As arestas são colocadas em ordem crescente de tamanho e são percorridas nessa ordem, de forma que se o vértice ainda não pertence a MST (não pertence a nenhum grupo) é inserido na MST (é inserido no mesmo grupo do outro vértice da aresta). Como se trata de um grafo completo, todos os elementos pertencem a um único grupo no final da execução do algoritmo e todos os vértices são visitados.

As informações de inserção na MST são importantes para o algoritmo de agrupamento, pois determinam quais grupos estão sendo unidos ao longo da execução do algoritmo e qual o valor da dissimilaridade utilizada em tal união (qual o valor do "galho"no dendrograma).

Algoritmo para o método Vizinho Mais Próximo com Kruskal:

1. Dados são inseridos no grafo
2. São calculadas as dissimilaridades dos dados
3. São colocadas em uma lista e ordenadas as arestas do grafo
4. Busca-se a aresta de menor custo cujo vértice ainda não esteja na MST
5. O vértice é inserido na MST e a aresta utilizada tem seu custo armazenado
6. Segue-se inserindo os vértices na MST até que todos os vértices sejam inseridos

Da aplicação do algoritmo sobre os dados foram gerados dendrogramas com o auxílio do programa Matlab para dados de protease e transcriptase reversa (Figura 3) , apenas para sequências de protease (Figura 4) e apenas para sequências de transcriptase reversa (Figura 5).

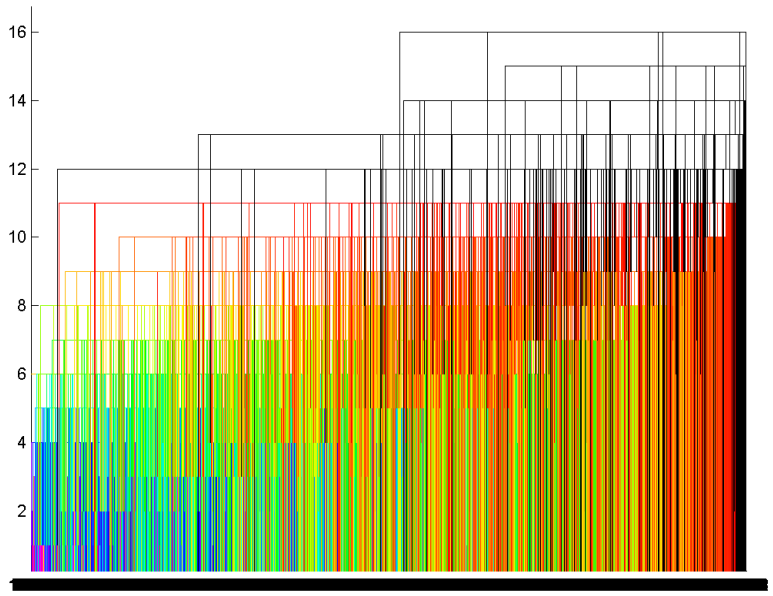

Figura 3: Dendrograma para dados de protease e transcriptase reversa (sequências parciais)

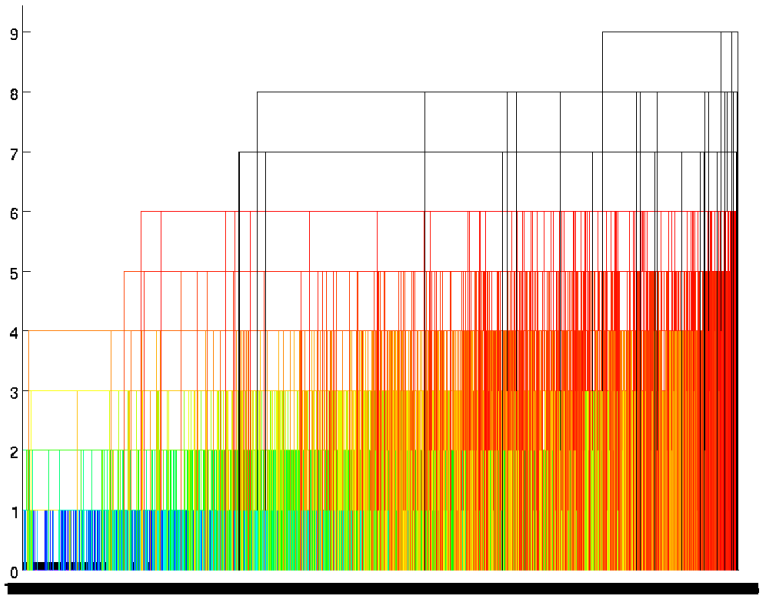

Figura 4: Dendrograma para dados de protease (sequências parciais)

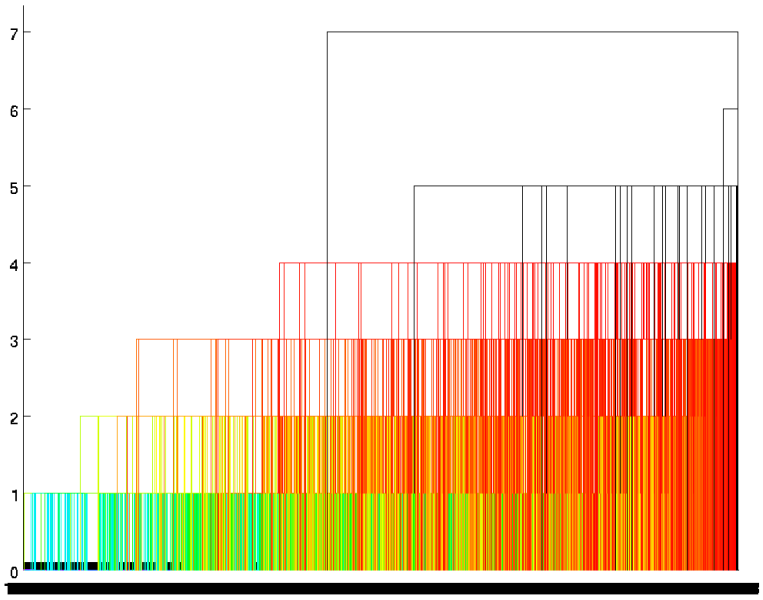

Figura 5: Dendrograma para dados de transcriptase reversa (sequências parciais)

Nota-se que os "galhos" do dendrograma não são muito grandes (as dissimilaridades são baixas), o que significa que, com a forma de métrica utilizada, os dados podem ser representados por pontos próximos uns dos outros e não há grande dissimilaridade entre eles. Portanto, as sequências de proteínas são bastante parecidas e conservadas, o que deve ser verdadeiro uma vez que são bastante importantes para o vírus.

Como em um dendrograma com essa quantidade de elementos a visualização dos grupos é difícil, arquivos de texto ajudaram na exploração dos agrupamentos gerados pelo algoritmo. Fazendo cortes em diferentes níveis de um dendrograma obtemos diferentes agrupamentos dos dados (Figura 6), assim, arquivos foram gerados para cada valor de dissimilaridade. Os arquivos continham a quantidade de grupos em cada nível de dissimilaridade e a constituição dos grupos (Tabelas 1-5).

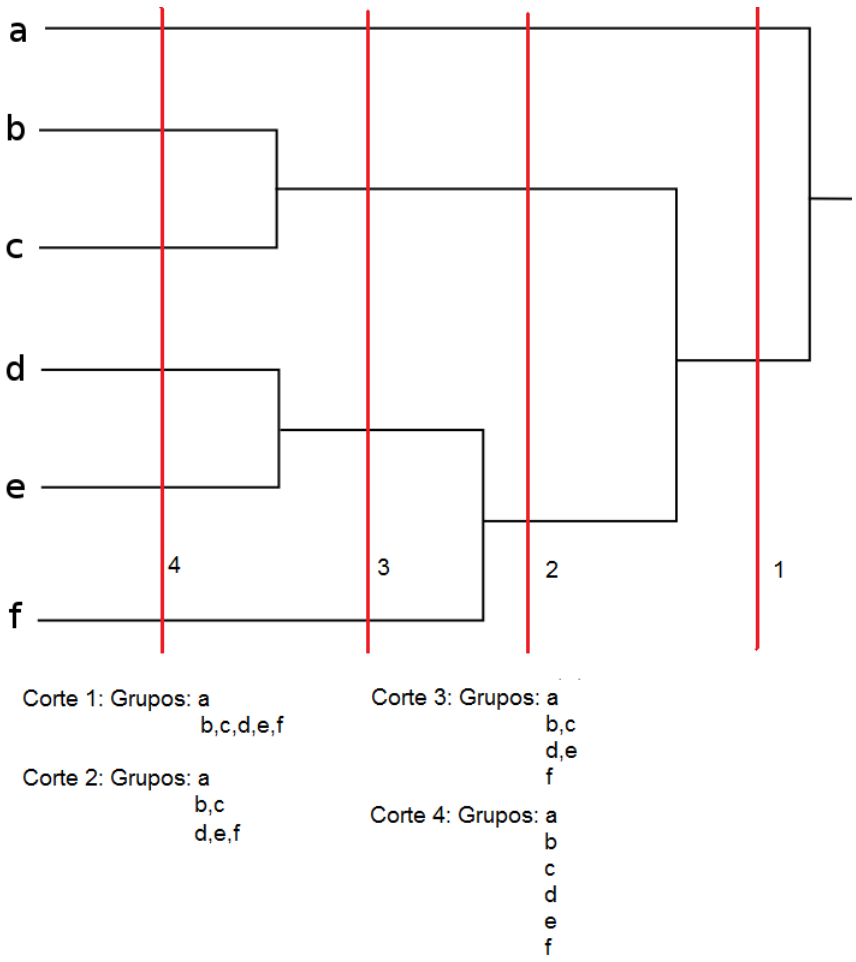

Figura 6: Impressão dos clusters gerados pelo Vizinho Mais Próximo

Tabela 1: Método Vizinho Mais Próximo com sequências de Transcriptase Reversa

19 Centróides, 13213 Dados		
Dissimilaridade de Corte	Quantidades de Grupos	Quantidades Elementos
1	12901	1
	117	2
	3	3
	1	4
	1	5
	1	60
2	12774	1
	172	2
	6	3
	2	4
	1	5
	1	64
4	12307	1
	315	2
	16	3
	6	4
	1	5
	1	7
	1	65
	1	1277
6	11151	1
	419	2
	28	3
	5	4
	3	5
	1	9
	1	53
	1	944
	1	99

Tabela 2: Continuação Método Vizinho Mais Próximo com sequências de Transcriptase Reversa

19 Centróides, 13213 Dados		
Dissimilaridade de Corte	Quantidades de Grupos	Quantidades Elementos
8	8919	1
	423	2
	28	3
	6	4
	1	5
	1	3335
10	6299	1
	314	2
	22	3
	3	4
	1	6208

Tabela 3: Método Vizinho Mais Próximo com sequências de Protease

13213 Dados		
Dissimilaridade de Corte	Quantidades de Grupos	Quantidades Elementos
1	8913	1
	335	2
	39	3
	11	4
	5	5
	1	7
	1	8
	1	14
	1	3415
2	86583	1
	310	2
	37	3
	17	4
	5	5
	2	6
	1	7
	1	8
	3	11
	1	64
	1	28
	1	29
	1	5625
4	3005	1
	185	2
	23	3
	4	4
	1	9734
6	938	1
	44	2
	5	3
	1	12172
8	212	1
	5	2
	1	12991
10	40	1
	1	13173

Tabela 4: Método Vizinho Mais Próximo com Sequências Parciais de Protease e Transcriptase Reversa

14393 Dados		
Dissimilaridade de Corte	Quantidades de Grupos	Quantidades Elementos
1	12243	1
	293	2
	27	3
	6	4
	4	5
	1	6
	1	87
	1	1346
2	10466	1
	347	2
	35	3
	12	4
	7	5
	2	6
	2	7
	2	8
	1	9
	1	11
	1	2983
4	6563	1
	305	2
	35	3
	9	4
	3	5
	4	6
	1	33
	1	7007
6	3246	1
	210	2
	22	3
	2	4
	2	5
	2	6
	1	33
	1	10631

Tabela 5: Continuação Método Vizinho Mais Próximo com Sequências Parciais de Protease e Transcriptase Reversa

14393 Dados		
Dissimilaridade de Corte	Quantidades de Grupos	Quantidades Elementos
8	1197	1
	81	2
	9	3
	3	4
	1	5
	1	31
	1	12959
10	344	1
	20	2
	2	3
	1	14003

Os arquivos gerados mostram que os dados tendem a formar grupos unitários nos níveis menores de dissimilaridade e grupos abrangendo a maioria das sequências nos níveis maiores de dissimilaridades. Assim, observa-se que não há formação de grupos com a utilização dessa métrica e essa técnica para esses dados. Esse resultado pode ter sido influenciado pela medida de similaridade utilizada, não sendo essa capaz de medir adequadamente as similaridades entre os dados.

No auxílio à análise de dados, também foram construídos gráficos contendo a frequência de ocorrência de mutações com mudança de aminoácidos para cada posição. Ou seja, a cada comparação entre duas sequências, eram verificadas quais posições possuíam aminoácidos diferentes, resultando em um gráfico que mostra quais posições possuem maior variação de aminoácidos.

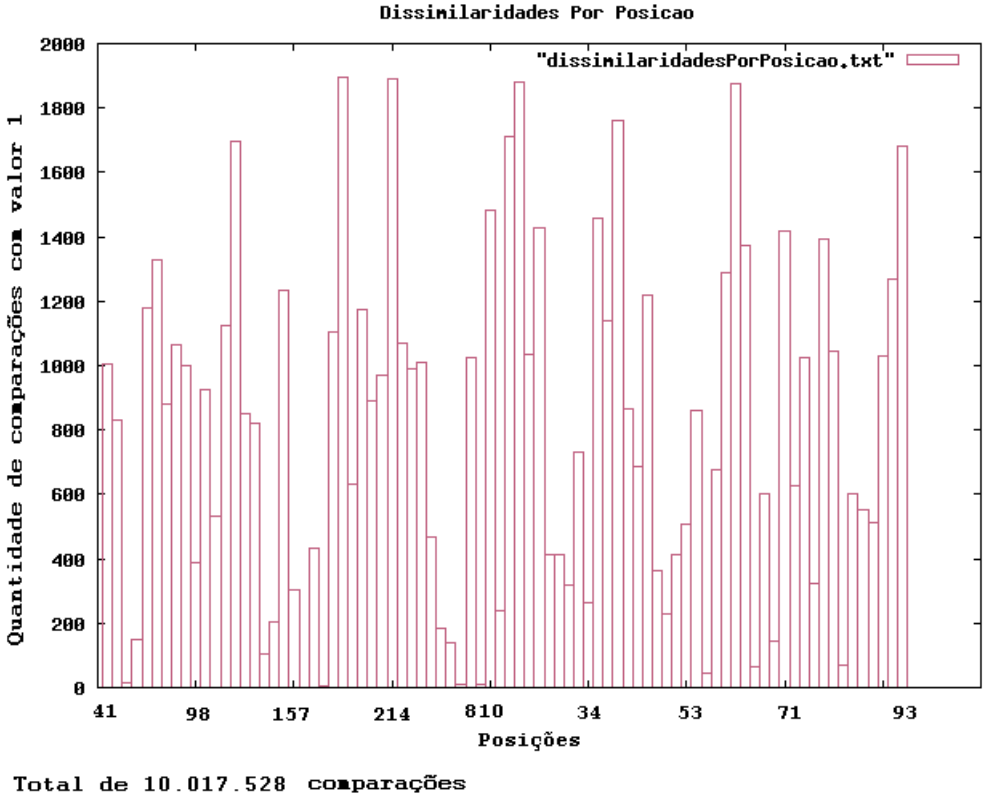

Figura 7: Gráfico de disimilaridades por posição com protease e transcriptase reversa

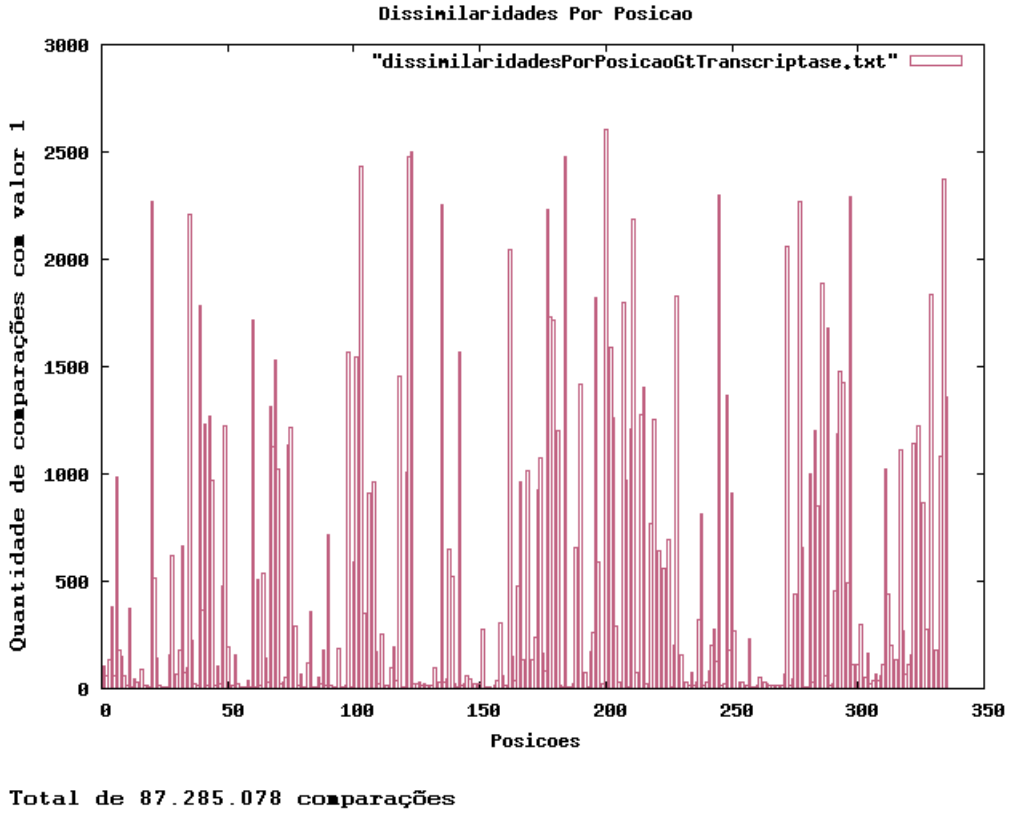

Figura 8: Gráfico de disimilaridades por posição com transcriptase reversa

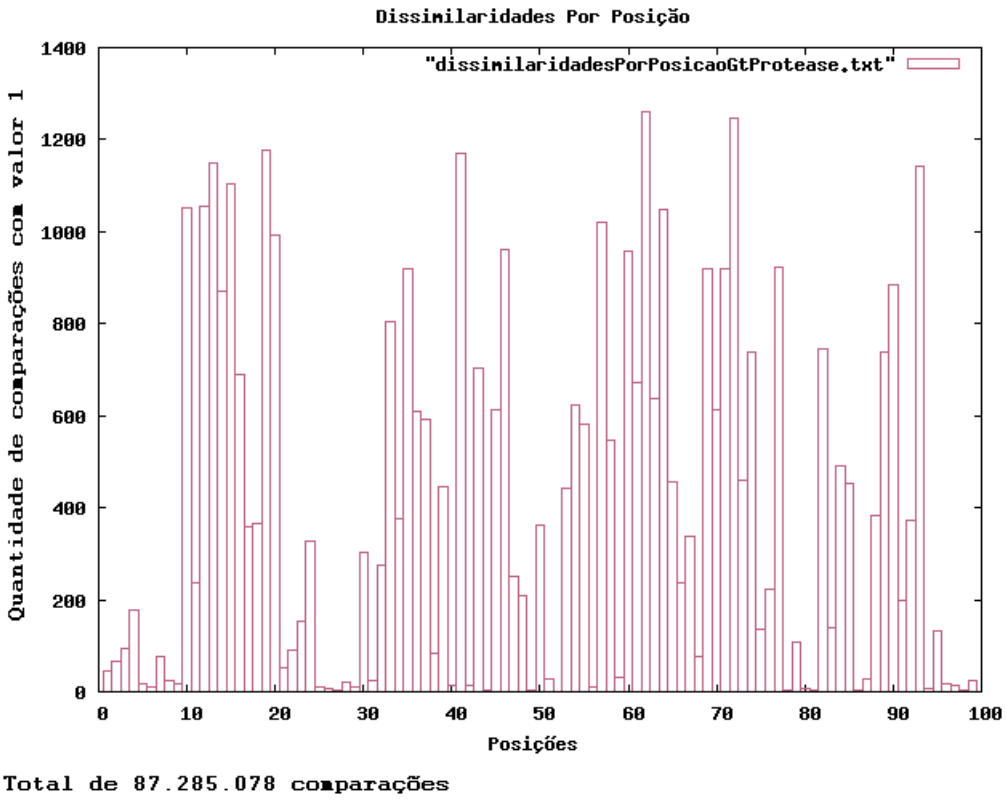

Figura 9: Gráfico de disimilaridades por posição com protease

4.4 Implementação do método K-Médias

Como algoritmo não-hierárquico, foi escolhido K-Médias que é um dos mais utilizados como técnica de agrupamento. A implementação do algoritmo K-Médias foi dada segundo [18].

A dificuldade da utilização desse método é dada na escolha correta dos centróides, uma vez que essa decisão interfere no resultado final, podendo gerar agrupamentos artificiais caso os centróides não sejam bem escolhidos. É necessário escolher cuidadosamente os centróides de forma que cada uma das sequências escolhidas represente bem os grupos existentes.

Como ponto de partida para a escolha dos centróides utilizou-se os resultados do experimento com o método Vizinho Mais Próximo, sendo que as sequências agrupadas com maiores e sequências com menores valores de dissimilaridades foram escolhidas como centróides. Com essa escolha de centróides o agrupamento não foi bem sucedido, pois não formou grupos distintos e sim um grande grupo abrangendo todos os dados.

Em um segundo experimento, um especialista selecionou algumas sequências dentre as da tentativa anterior de acordo com sua experiência e conhecimento. Para sequências parciais também não foram obtidos grupos bem definidos nessa tentativa de agrupamento (Tabela 6). Já com as sequências completas (Tabelas 7 e 8), houve a formação de alguns grupos, mas ainda não se obteve os grupos esperados.

Tabela 6: Método K-Médias com Sequências Parciais de Transcriptase Reversa e Protease

14393 Dados		
Centróides	Quantidades de Grupos	Quantidades Elementos
XB - 6 centróides	1	13213
XC - 6 centróides	1	14391
	1	2
XF - 6 centróides	1	14391
	1	2
XXF - 6 centróides	1	13213
XB,XC,XF,XFF - 19 centróides	1	14391
	1	2

Tabela 7: Método K-Médias com Sequências de Transcriptase Reversa

13213 Dados		
Centróides	Quantidades de Grupos	Quantidades Elementos
XB - 6 centróides	1	13186
	1	2
	1	25
XC - 6 centróides	1	13213
XF - 6 centróides	1	2177
	1	10977
	1	59
XXF - 6 centróides	1	13213
XB,XC,XF,XFF - 19 centróides	1	12648
	1	367
	1	155
	1	24
	1	11
	1	4
	1	2
	2	1

Tabela 8: Método K-Médias com Sequências de Protease

13213 Dados		
Centróides	Quantidades de Grupos	Quantidades Elementos
XB - 6 centróides	1	13206
	1	7
XC - 6 centróides	1	13213
XF - 6 centróides	1	6785
	1	6411
	1	17
XXF - 6 centróides	1	13213
XB,XC,XF,XFF - 19 centróides	1	13203
	1	10

Em ambas as aplicações, os dados tenderam a se unir em um único grupo. Isso pode ter acontecido pelo fato de o tipo de cálculo de dissimilaridade utilizado não ser o mais adequado para os dados e não expressar bem as dissimilaridades.

5 Conclusão

Os agrupamentos obtidos dos métodos Vizinho Mais Próximo e K-Médias podem refletir a conservação das sequências de transcriptase reversa e protease na multiplicação do vírus HIV pela importância que essas duas proteínas possuem para sua replicação. A escolha da medida de Distância Euclideana pode não ter sido capaz de retratar bem as diferenças entre as sequências e pode ter impedido a obtenção de agrupamentos verdadeiros.

Na tentativa de se obter grupos consistentes de ambos os métodos, estudos futuros podem utilizar outras métricas de distância ou outros tipos de cálculos de similaridades a fim de destacar as diferenças entre as sequências.

Outra possibilidade é utilizar o cálculo de Distância Euclideana, mas com pesos diferentes para as diferentes posições de acordo com estudos sobre os aminoácidos que constituem as sequências dessas proteínas.

As aplicações de outros métodos também podem ser mais bem sucedidas para esse tipo e distribuição de dados.

6 Agradecimentos

Agradeço à doutora Ester Sabino que foi importante no desenvolvimento desse trabalho.

7 Parte Subjetiva

O trabalho de formatura me permitiu ter contato com um estudo acadêmico com aplicação prática direta, o que foi bastante motivador, bem como a temática do HIV que é bastante interessante e importante de ser estudada.

Durante o trabalho, foi de grande importância a interação entre o orientador, a orientada e a especialista na área médica. Como as duas áreas são bastante distintas, é necessário que haja disposição e determinação para que se consiga entender os conceitos e os diferentes pontos de vista.

Várias disciplinas lecionadas ao decorrer do curso foram importantes para o desenvolvimento desse estudo, tais como:

1. MAC0110 - Introdução à Computação, MAC0122 - Princípios de Desenvolvimento de Algoritmos e MAC 323 - Estruturas de Dados que forneceram os fundamentos para programação

2. MAC0211 - Laboratório de Programação I e MAC0242 - Laboratório de Programação II que mostraram utilidade de ferramentas como o Latex e mostraram na prática como acontece o desenvolvimento de programas

3. MAC0460 – Aprendizagem computacional: modelos, algoritmos e aplicações que mostrou algumas abordagens para problemas de classificação e clusterização

4. MAC0328 - Algoritmos em Grafos que introduziu o conceito de grafos e a aplicação em diferentes problemas

Apesar de ainda não ter sido atingido os resultados exatamente como

desejado, os ajustes de alguns parâmetros e a utilização de outros métodos poderão auxiliar na obtenção do objetivo almejado.

No momento, esse estudo está sendo explorado com a aplicação de outros métodos e abordagens, como por exemplo a utilização de árvores binárias nas quais os nós internos representam a presença ou não de certas mutações nas sequências e nos níveis das folhas se tem a formação de diferentes grupos.

Outro método de análise que está sendo aplicado é a utilização da representação dos dados como sequências binárias ordenadas. Esta análise permite levar em consideração a posição em que ocorre a mutação e pode auxiliar na elucidação dos agrupamentos existentes.

A bioinformática, a análise de dados e o reconhecimento de padrões se mostraram áreas de pesquisa bastante interessantes e uma direção interessante a ser tomada no futuro.

8 Referências bibliográficas

[1] Morgado, M.G., Sabino, E.C., Shpaer, E.G., Bongertz, V., Brigido, L., Guimaraes, M.D., Castilho, E.A., Galvao-Castro,B., Mullins, J.I., Hendry, R.M. and et al. **V3 region polymorphisms in HIV-1 from Brazil: prevalence of subtype B strains divergent from North American/European prototype and detection of subtype F** AIDS Res Hum Retroviruses, 10 (1994) 569-76.

[2] Baeten JM, Chohan B, Lavreys L, Chohan V, McClelland RS, Certain L, Mandaliya K, Jaoko W, Overbaugh J. (2007) **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.

[3] Laeyendecker O, Li X, Arroyo M, McCutchan F et al.(2006) **The Effect of HIV Subtype on Rapid Disease Progression in Rakai, 13th Conference on Retroviruses and Opportunistic Infections** Uganda,(abstract no. 44LB),

[4] Kanki P.J., Donald J. Hamel, Jean-Louis Sankalé, Chung-cheng Hsieh, Ibou Thior, Francis Barin, Stephen A. Woodcock, Aïssatou Guêye-Ndiaye, Er Zhang, Monty Montano, Tidiane Siby, Richard Marlink, Ibrahima NDoye, Myron E. Essex, and Souleymane MBoup (1999) **Human Immunodeficiency Virus Type 1 Subtypes Differ in Disease Progression, Journal of Infectious Diseases** Volume 179 Number 1.

[5] Carpenter CCJ, Fischl MA, Hammer SM, Hirsch MS, Jacobsen DM, Katzenstein DA, et al. **Antiretroviral therapy for HIV infection in 1998.** JAMA 1998, 280:78–86.

[6] Shafer RW, Winters MA, Palmer S, Merigan TC. **Multiple concurrent reverse transcriptase and protease mutations multidrug resistance of HIV-1 isolates from heavily treated patients.** Ann Intern Med 1998, 128:906–911.

[7] Ledergerber B, Egger M, Opravil M, Telenti A, Hirschel B, Battegay M, et al. **Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study.** Lancet 1999, 353:863–868.

[8] Finzi D, Hermankova M, Pierson T, Carruth LM, Buck C, Chaisson RE, et al. **Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy.** Science 1997, 278:1295-1300

[9] Ross L, Johnson M, DeMasi R, Liao Q, Graham N, Shaefer M, et al. **Viral genetic heterogeneity in HIV-1 infected individuals is associated with increasing use of HAART and higher viremia.** AIDS 2000, 14:813–819.

[10] Conway B, Wainberg MA, Hall D, Harris M, Reiss P, Cooper D, et al. **Development of drug resistance in patients receiving combinations of zidovudine, didanosine and nevirapine.** AIDS 2001, 15:1269–1274.

[11] Vella S. and Palmisano L. **Antiviral therapy: state of the HAART.** Antiviral Res 2000, 45:1–7.

[12] Preston, B.D., B.J. Poiesz, and L.A. Loeb. (1988) **Fidelity of HIV-1 reverse transcriptase.** Science. 242(4882): p. 1168-71

[13] Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature. 1995; 373:123-6.

[14] Wei, X., S.K. Ghosh, M.E. Taylor, V.A. Johnson, E.A. Emini, P. Duestch, J.D. Lifson, S. Bonhoeffer, M.A. Nowak, B.H. Hahn, and G.M. Shaw. 1995. **Viral dynamics in human immunodeficiency virus type 1 infection.** Nature (Lond.). 373: 117-122 [Medline] .

[15] Coffin JM. **HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy.** Science 1995;267:483-489.

[16] Korber B.T., Foley B. F., Kuiken C.I. , Pillai S. K., and Sodroski J. G., (1998) **"Numbering Positions in HIV Relative to HXB2CG,"in Human Retroviruses and AIDS**. Report LA-UR 99-1704, B. T. Korber et. al., Ed. Los Alamos, NM: Los Alamos National Laboratory, pp. III-102;III-111.

[17] Jain, A. K., Murty, M. N., and Flynn, P. J., **Data clustering: A review**, ACM Computing Surveys 31, 264–323 (1999).

[18] Johnson, R.A., Wichern, D.W. (1982). **Applied multivariate statistical analysis.** Englewood Cliffs, NJ: Prentice-Hall.

[19] NAGY, G. 1968. **State of the art in pattern recognition.** Proc. IEEE 56, 836–862.

[20] BAEZA-YATES, R. A. 1992. **Introduction to data structures and algorithms related to information retrieval.** In Information Retrieval: Data Structures and Algorithms,W.B. Frakes and R. Baeza-Yates, Eds. Prentice- Hall, Inc., Upper Saddle River, NJ, 13–27.

[21] GOWER,J.C. AND ROSS, G. J. S. 1969. Minimum spanning rees and single-linkage cluster analysis. Appl. Stat. 18, 54–64.

[22] The MathWorks, Inc., MATLAB 4.2, 24 Prime Park Way, Natick MA.