Classificação das mutações de vírus HIV

Aluna: Mina Cintho
Orientador: João Eduardo Ferreira
MAC0499 - Trabalho de Formatura Supervisionado

Introdução

A interação entre a área de computação e biologia produz uma grande quantidade de dados que guardam muitas informações e conhecimento que não são facilmente extraídos por causa do grande volume de variáveis.

A tecnologia tem permitido a obtenção de grandes volumes de sequências de proteínas do vírus HIV, mas ainda é necessário explorar técnicas de análise de dados para relacionar as sequências a características importantes como a suscetibilidade dos vírus a drogas. Atualmente, essa classificação é feita “manualmente” de acordo com conhecimento intuitivo dos especialistas.

Esse trabalho tem como objetivo automatizar a classificação de vírus HIV a partir da aplicação de algoritmos de classificação em dados de pacientes infectados.

Proposta

Baseado em técnicas de reconhecimento de padrões, nos propusemos a implementar e aplicar os métodos de agrupamento Vizinho Mais Próximo e K-médias em torno de 13.000 dados de pacientes com o vírus HIV com o objetivo de obter classificações automáticas dos vírus de acordo com as mutações presentes nas sequências através das técnicas de agrupamento.

Fundamentos Biológicos

Mutações são alterações das seqüências de nucleotídeos do DNA propagadas pela produção de cópias. Podem afetar funções importantes das proteínas, pois as sequências de DNA são responsáveis pela determinação de suas sequências de aminoácidos.

Proteínas como a protease (Figura 1) e a transcriptase reversa (Figura 2) são importantes para a replicação do vírus do HIV e estão sujeitas a mutações que podem levar a diferentes respostas a diferentes drogas.

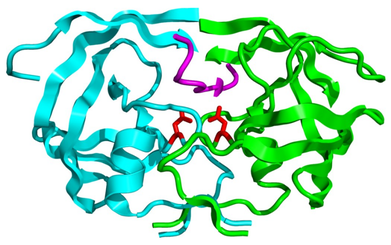

Figura 1: Protease

Fonte: Wikipedia
http://en.wikipedia.org/wiki/File:HIV_protease_1EBY.png

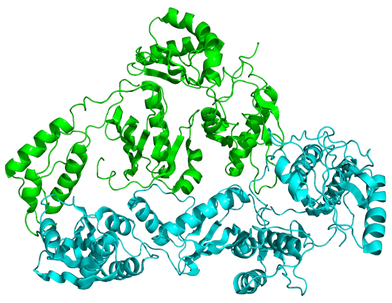

Figura 2: Transcriptase Reversa

Fonte: Wikipedia
http://en.wikipedia.org/wiki/File:Reverse_Transcriptase_1HMV.png

Fundamentos de Reconhecimento de Padrões

O agrupamento é a organização de uma coleção de padrões (dados), geralmente representados como vetores de medidas ou um ponto no espaço multidimensional, em grupos baseados em similaridade [1].

Similaridades podem ser representadas como coeficientes de correlação, medidas de associação ou medidas de distâncias, como a medida Euclidiana.

As dimensões representam atributos que caracterizam um padrão e, a partir dessa representação, pode-se tentar estabelecer similaridades entre os padrões.

Métodos de agrupamento hierárquicos geram grupos estabelecendo uma hierarquia que pode ser representada na forma de uma árvore ou dendrograma (Figura 3). No Vizinho Mais Próximo, por exemplo, a distância entre dois grupos é a menor das distâncias entre um elemento de um grupo e um elemento do outro grupo.

Já os métodos não-hierárquicos geram grupos com uma única partição (Figura 4). O K-Médias, por exemplo, une padrões de acordo com a maior proximidade ao centro de um grupo, ou centróide. Os centróides são definidos como as médias aritméticas de cada uma das dimensões das distâncias de todos os pontos de um grupo e são definidos aleatoriamente no início do algoritmo.

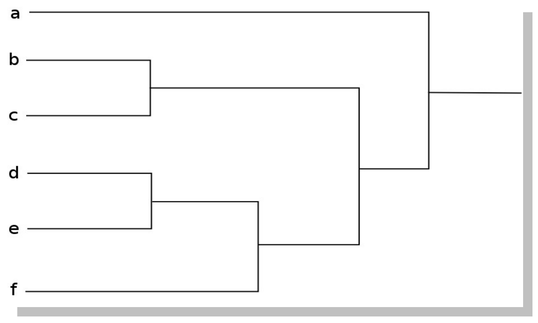

Figura 3: Exemplo de dendrograma

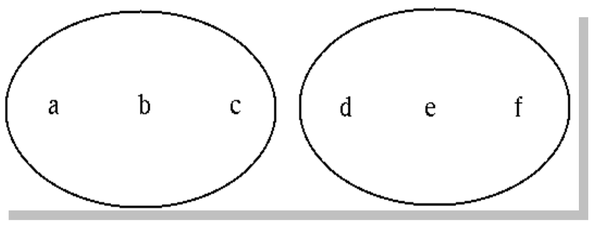

Figura 4: Exemplo de agrupamento não hierárquico

Resultados/Análise

Vizinho Mais Próximo

A partir da aplicação dos dados à implementação do método Vizinho Mais Próximo, foram construídos dendrogramas (Figura 5) com a ajuda do software Matlab [3].

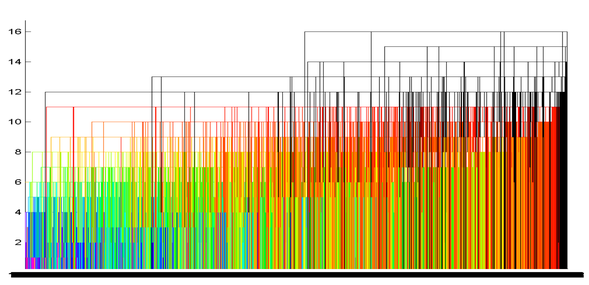

Figura 5: Dendrograma para dados de protease e transcriptase reversa

Para a melhor visualização, foram gerados arquivos texto para cada valor de dissimilaridade contendo a quantidade de grupos e sua constituição.

Também foram gerados gráficos mostrando em que regiões das proteínas há maior variação nas sequências de aminoácidos (Figura 6).

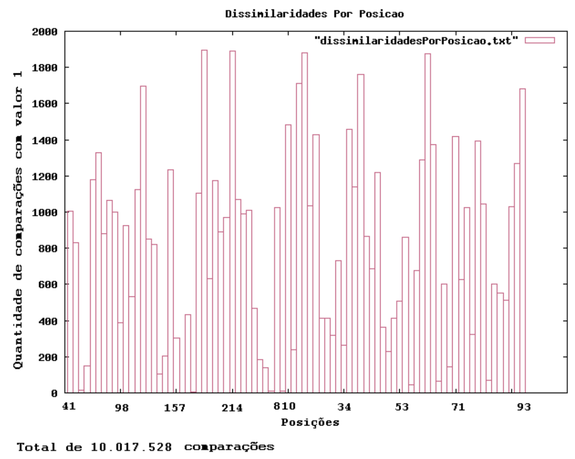

Figura 6: Gráfico de dissimilaridades por posição

K-médias

A escolha dos centróides no método K-médias é decisiva para a obtenção dos grupos no final do algoritmo, por isso foram utilizados os resultados do Vizinho Mais Próximo e o auxílio de um especialista para escolha dos centróides.

Várias combinações de centróides foram utilizadas. Algumas tentativas resultaram em alguns agrupamentos, mas não foi possível separar os dados de acordo com os diferentes subtipos.

Conclusão

As dissimilaridades do padrões receberam valores baixos em relação ao número de posições tendo em vista a métrica utilizada, refletindo a importância e conseqüente conservação das sequências do vírus.

Os métodos utilizados não obtiveram grupos que refletissem fielmente a classificação em subtipos possivelmente pelo fato de que alguns aminoácidos em algumas posições possuem um papel mais importante na manutenção das suas funcionalidades da proteína que outros aminoácidos em outras posições.

A distância euclidiana pode também não ter sido capaz de quantificar as dissimilaridades entre as sequências. Em uma próxima etapa do trabalho pretendemos alterar os cálculos de medida de dissimilaridades, por exemplo, dando pesos diferentes para as diferentes posições de aminoácidos nas proteínas.

Como trabalhos futuros pretendemos estudar a aplicação de outras formas de medidas de dissimilaridade, bem como outros métodos de classificação.

Referências Bibliográficas
[1] Jain, A. K., Murty, M. N., and Flynn, P. J., Data clustering: A review, ACM Computing Surveys 31, 264–323 (1999).
[2] Johnson, R.A., Wichern, D.W. (1982). Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice-Hall.
[3] The MathWorks, Inc., MATLAB 4.2, 24 Prime Park Way, Natick MA.
[4] Jain, A. K., Murty, M. N., and Flynn, P. J., \textbf{Data clustering: A review}, ACM Computing Surveys 31, 264–323 (1999).
[5] Laeyendecker O, Li X, Arroyo M, McCutchan F et al.(2006) \textbf{The Effect of HIV Subtype on Rapid Disease Progression in Rakai, 13th Conference on Retroviruses and Opportunistic Infections} Uganda,(abstract no. 44LB),