

MAC499 - Trabalho de Formatura Supervisionado

Classificação do vírus HIV

Aluna: Mina Cintho
Orientador: João Eduardo Ferreira

Introdução

Computação e Biologia → Grande volume
de dados

Exemplos:

Sequenciamento de nucleotídeos

Sequenciamento de proteínas

ATA

GGA

CTCGGGATAACC

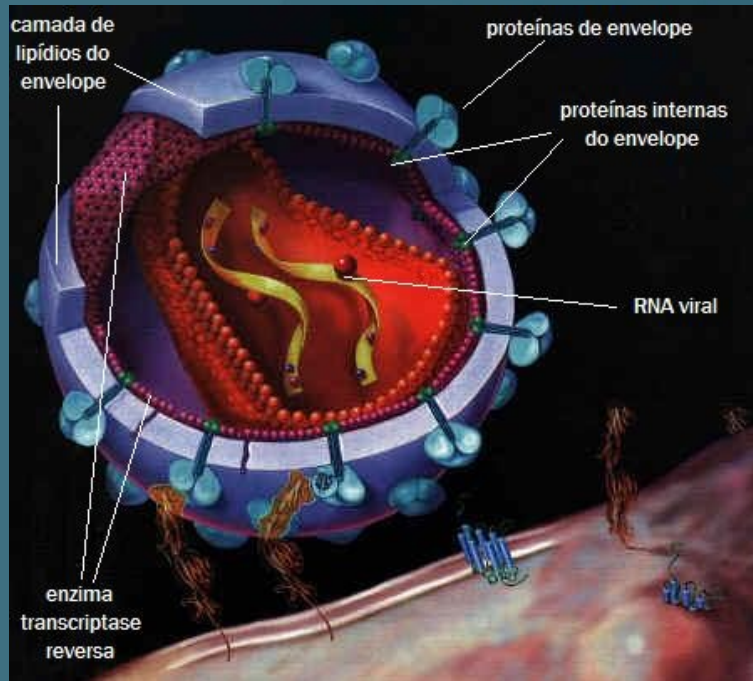
Introdução

Análise de dados → Informação

AGTACGATCATA
GACATGGA
TCGGGATAACC

Introdução

Exemplo: HIV Protease e Transcriptase Reversa



Replicação do vírus

http://insightpublicidade.files.wordpress.com/2008/04/hiv_virus.jpg

Introdução

Proteínas relacionadas à resistência a drogas

Decisões de tratamentos

Experiência de especialistas

Objetivo

Métodos de Classificação



Classificação automática



Decisão sobre tratamentos

Proposta

13.000 dados

K-médias

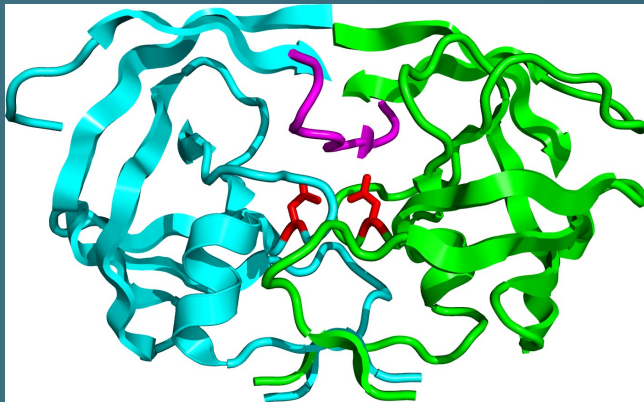
Vizinho Mais Próximo

Fundamentos Biológicos

Mutações

Propagação das mutações

Alteração das proteínas



Fonte: Wikipedia

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

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

ATA
GGA
CCGGGATAACC

Fundamentos de Reconhecimento de Padrões

Agrupamento por similaridades:

Coeficientes de correlação

Medidas de associação

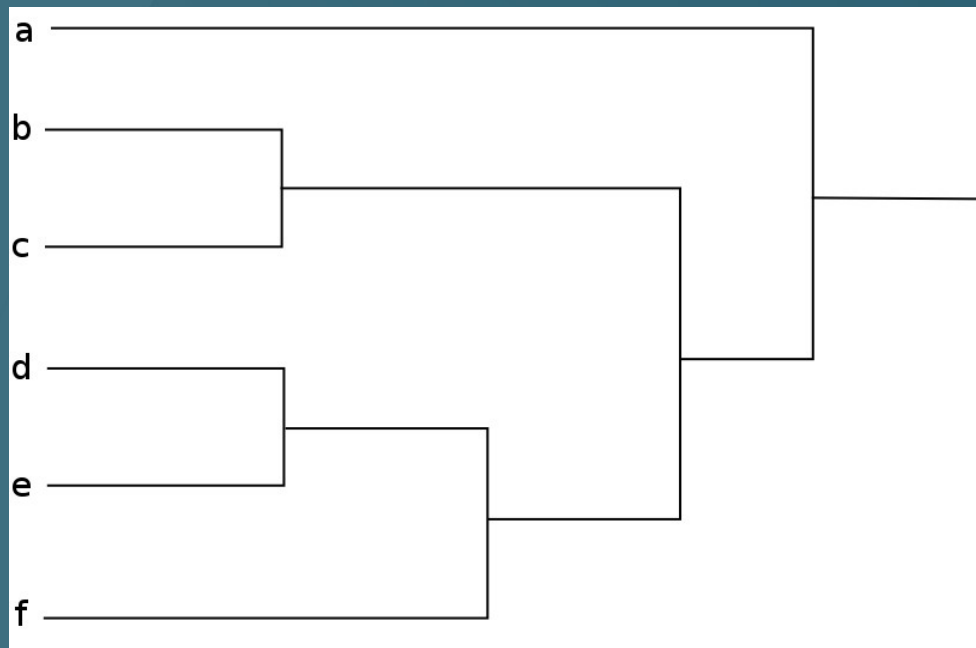
Distância

ATA

GGA

TCGGGATAACC

Métodos Hierárquicos



ATA

GGA

TCGGGATAACC

Métodos Hierárquicos:

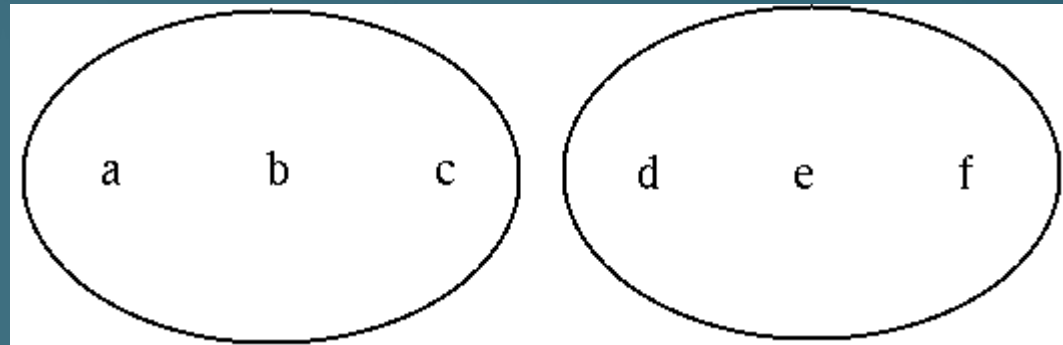
Vizinho Mais Próximo

Vizinho Mais Distante

Ligação Média Não Ponderada

Métodos Não-Hierárquicos:

K-Médias



Atividades Realizadas

Dados em Representação Binária

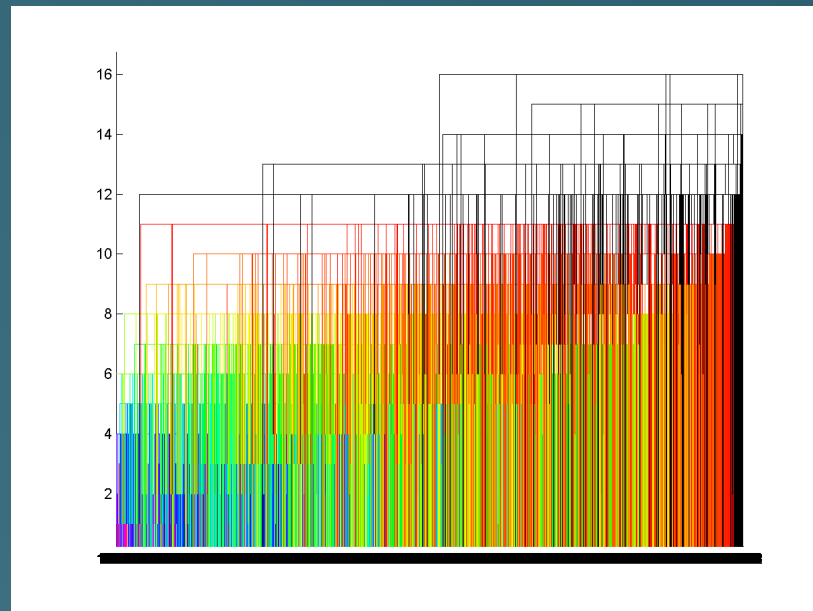
Medida de Dissimilaridade

CTAGTACTCATA
GAGGGA
CTCGGGATAACC

Atividades Realizadas

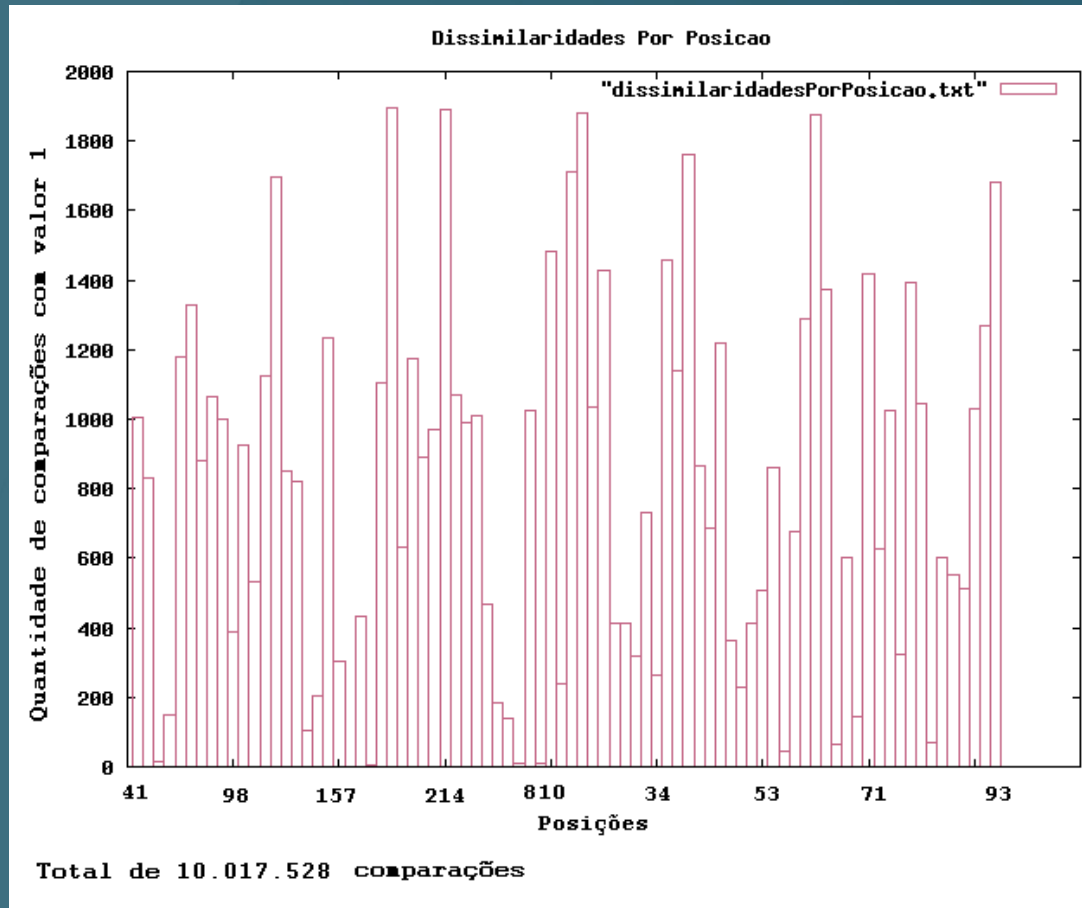
Vizinho Mais Próximo: MST

Dendrograma



ATA
GGA
TCGGGATAACC

Atividades Realizadas



Atividades Realizadas

K-Médias:

Escolha dos Centróides

Resultados do Vizinho Mais Próximo

Especialista

Conclusão

Outliers

Diferentes pesos

Outras medidas de dissimilaridades

Outros métodos

ATA

GGA

TCGGGATAACC

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., **Data clustering: A review**, ACM Computing Surveys 31, 264–323 (1999).
- [5] 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),