Relatório de Computação I
Genome Profiling

Bruno Daiki Yamada
Gustavo Arbex
Isabella Costa Maia

18 de novembro de 2015

Data do procedimento: 16 de novembro de 2015
Professor: Yoshiharu Kohayakawa

1 Objetivos

Esse experimento tem como objetivo verificar a similaridade entre diferentes materiais genéticos através de comparações de sequências de nucleotídeos de duas espécies por métodos computacionais.

2 Introdução

O exercício baseia-se, primordialmente, em duas funções:hash() e unhash(). Considera-se, em ambas, os nucleotídeos A C G T, na respectiva ordem, representáveis pelos números 0 1 2 3. A primeira delas, a partir de uma sequência determinada de nucleotídeos de tamanho k, devolve o 'hash' correspondente, ou seja, o número na base decimal equivalente ao número de base 4 formado pela série. A segunda função, como o nome sugere, realiza o processo inverso a partir do recebimento de um dado k-grama. Esse valor nada mais é do que a quantidade de nucleotídeos seguidos que desejamos usar para fins de comparação. Assim, quanto maioro k-grama, maior o rigor com a qual comparamos os strings.

3 Procedimento

Selecionamos dois tipos de sequências de nucleotídeos: genomas completos de vírus (que têm tipicamente cerca de 5000 nucleotídeos), e a sequência de um locus codificante de uma proteína. Entre os genomas de vírus, foram selecionados diversas variedades de Human Papiloma Virus, caxumba e um vírus do gênero Cotesia. Quanto a sequência de proteínas, foram escolhidas duas sequências

codificantes de triose fosfato isomerase, uma enzima importante na glicólise de duas espécies diferentes.

- *Cotesia congregata* - Genoma completo;
- *Human papillomavirus* tipo 16 - Genoma completo;
- *Human papillomavirus* tipo 18 - Genoma completo;
- *Human papillomavirus* tipo 38 - Genoma completo;
- *Human papillomavirus* tipo 60 - Genoma completo;
- Vírus da caxumba variedade MuV-IA - Genoma completo;
- Vírus da caxumba variedade L-Zagreb - Genoma completo;
- Sequência da triose fosfato isomerase de Drosophila;
- Sequência da triose fosfato isomerase de Schistosoma.

4 Resultados e Discussão

Ao observamos a tabela abaixo, verificamos o grau de semelhança entre todos os genomas testados quando comparados dois a dois. É possível notar o alto grau de semelhança entre as variedades de HPV e de Caxumba, quando comparadas entre si: o grau de semelhança entre HPV38 e HPV16 foi de 91% , assim como entre HPV38 e HPV60. Entre as caxumbas de variedades MuV- IA e L-Zagreb houve semelhança de 99%. A menor similaridade encontrada, por sua vez, foi entre as trioses fosfato isomerase, de apenas 33%.

```
C:\Users\cezar\Documents\Java>java CompareAllGenome 5 < genomes.txt
       Cote    HP16    HP18    HP38    HP60    CxMu    CxZa    TPID    TPIS
Cote   1,00    0,87    0,66    0,84    0,87    0,80    0,79    0,38    0,76
HP16   0,87    1,00    0,77    0,91    1,00    0,84    0,84    0,44    0,70
HP18   0,66    0,77    1,00    0,75    0,77    0,70    0,70    0,40    0,54
HP38   0,84    0,91    0,75    1,00    0,91    0,87    0,87    0,50    0,68
HP60   0,87    1,00    0,77    0,91    1,00    0,84    0,84    0,44    0,70
CxMu   0,80    0,84    0,70    0,87    0,84    1,00    0,99    0,56    0,64
CxZa   0,79    0,84    0,70    0,87    0,84    0,99    1,00    0,56    0,64
TPID   0,38    0,44    0,40    0,50    0,44    0,56    0,56    1,00    0,33
TPIS   0,76    0,70    0,54    0,68    0,70    0,64    0,64    0,33    1,00
```

5 Conclusão

Concluímos que para certas variedades de vírus, devido às mutações, a sequência é muito afetada, de maneira que essa análise rudimentar de similaridade é insuficiente para identificar se o genoma refere-se a uma mesma espécie. Isso explica a razão pela qual o parâmetro principal para relacionar espécies de vírus é justamente o formato tridimensional das proteínas, que é muito mais conservado apesar das mutações.Concluímos ainda que nossa análise difere da análise utilizada pela comunidade científica para identificar similaridade entre proteínas pois nossa similaridade teve como resultado um valor menor do que o esperado entre as proteínas TPI, que são altamente conservadas através das gerações.