Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Algoritmos para Bioinformática I - Primeiro Semestre 2020

Trabalho Prático 5 - Verificação do efeito dos ruídos nas matrizes de distâncias usadas nos algoritmos de hierarquização – estudo de caso

Aluno: Guilherme de Abreu Lima Buitrago Miranda - Matrícula: 2018054788

1. Introdução

Boa parte das vezes que se pretende mostrar a teoria da evolução de maneira gráfica, opta-se por uma árvore filogenética. Assim, não é incomum pensar, por exemplo, na árvore dos primatas, que mostra o ser humano com um ancestral em comum com os orangotangos. Contudo, poucas vezes se vê árvores sendo usadas para outras entidades biológicas, como as proteínas, por exemplo.

Além disso, também não é exatamente claro como essas árvores são obtidas, sobretudo para organismos tão diferentes dos animais. Assim, este trabalho busca entender tais questões. Em particular, pretende-se verificar o efeito dos ruídos nas matrizes de distâncias usadas nos algoritmos de hierarquização para a construção de tais árvores.

Dessa forma, primeiramente, mostra-se as entradas esperadas e as saídas obtidas. Em seguida, discute-se a respeito de cada uma das informações geradas e os métodos utilizados pelo algoritmo. Por último, debate-se as conclusões alcançadas a partir do trabalho.

2. Entradas e Saídas

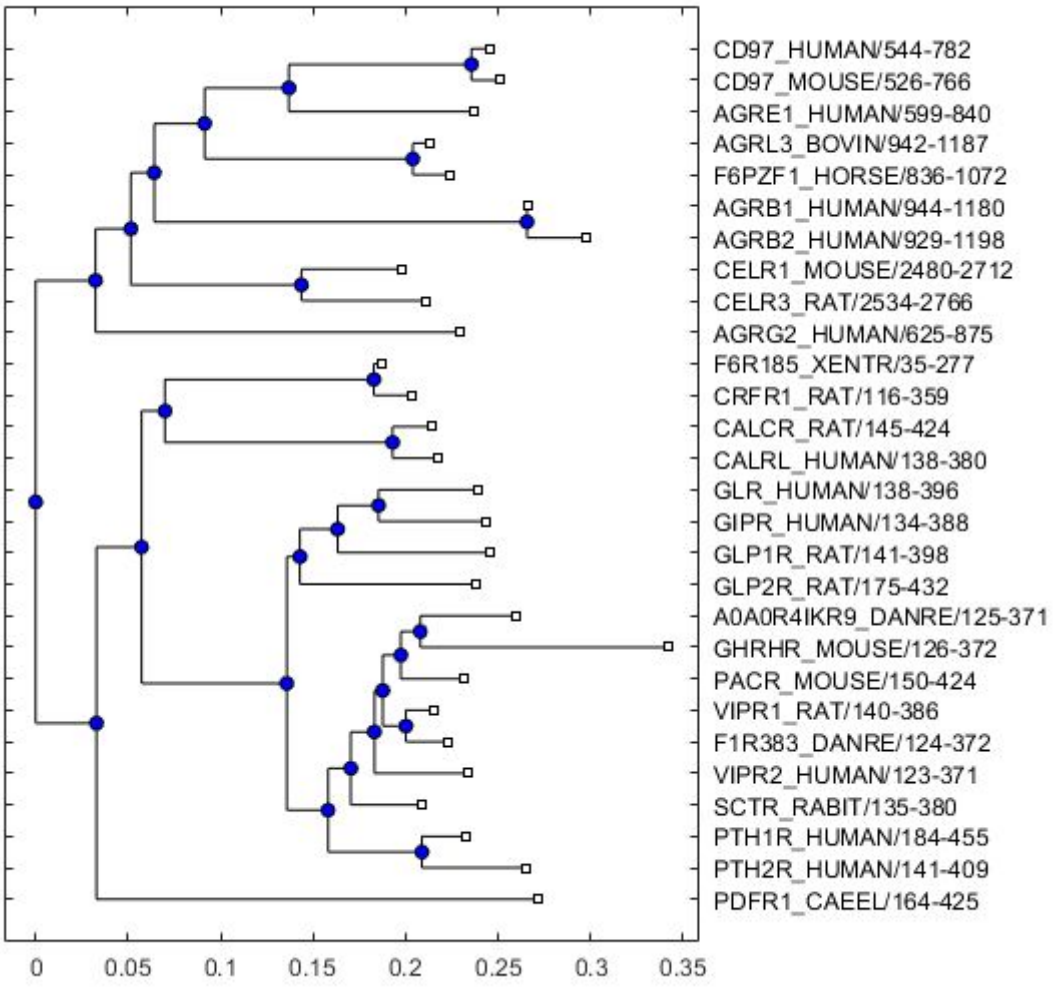
Utiliza-se, como principal entrada para o trabalho, o arquivo “PF00002.fasta”. Nele, são enumeradas uma série de proteínas (representadas por sequências de aminoácidos) de uma mesma família. Além disso, utiliza-se scripts fornecidos junto ao enunciado do trabalho, como o arquivo “montaA.m”.

Em seguida, aplica-se diferentes técnicas para representação dessas proteínas em um *vector space model*, a fim de se observar os pontos positivos e

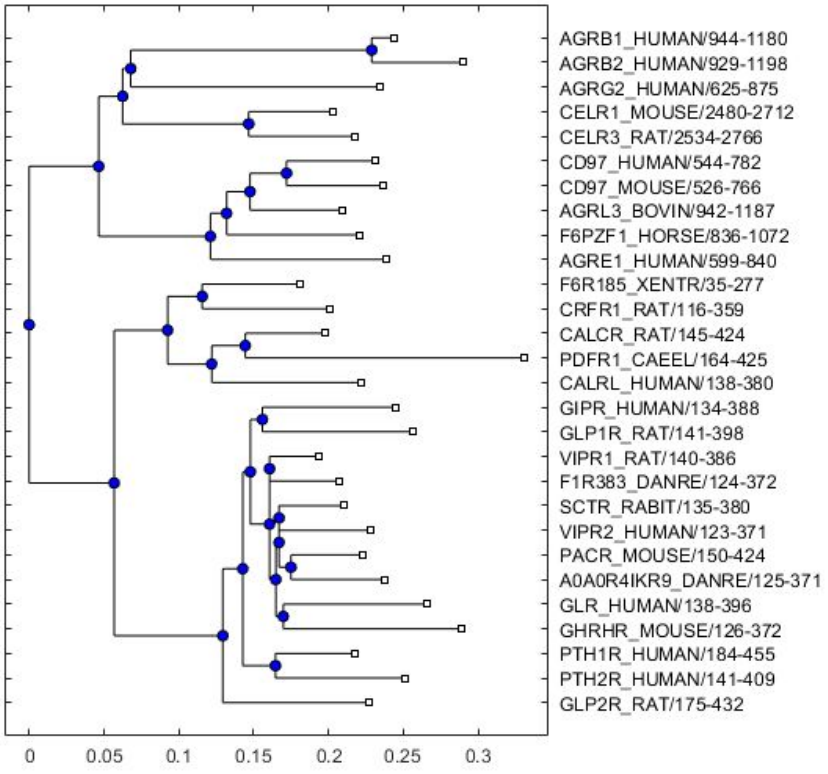
negativos de cada abordagem. O detalhamento de tais técnicas e sua respectiva avaliação será melhor tratado na seção 3 deste documento.

Por fim, obtém-se três árvores filogenéticas. São elas:

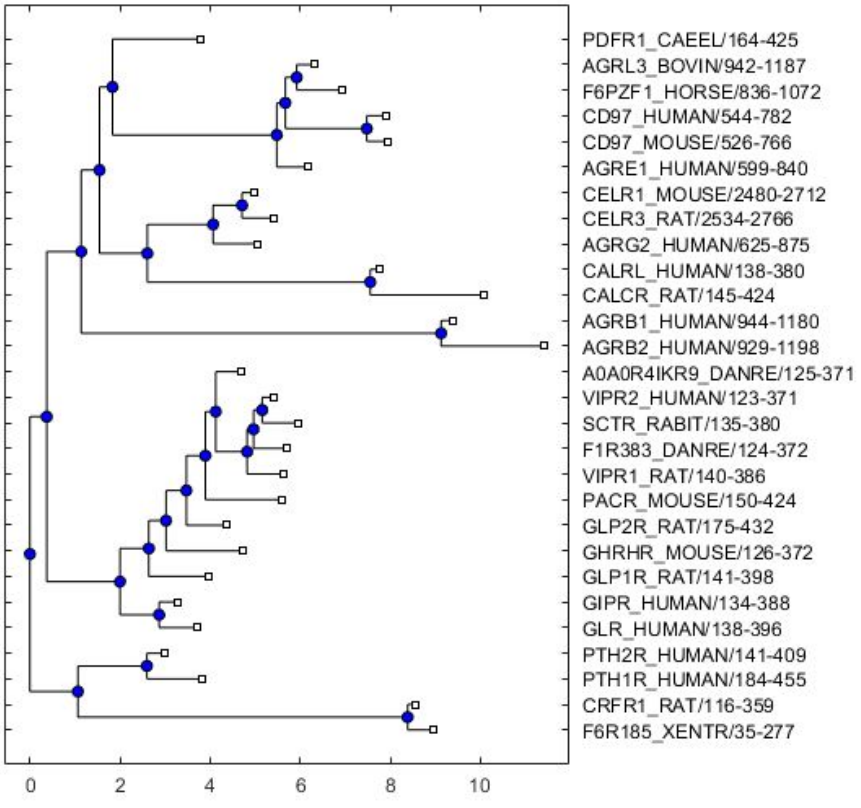
- Árvore 1:



- Árvore 2:


AGRB1_HUMAN/944-1180
AGRB2_HUMAN/929-1198
AGRG2_HUMAN/625-875
CELR1_MOUSE/2480-2712
CELR3_RAT/2534-2766
CD97_HUMAN/544-782
CD97_MOUSE/526-766
AGRL3_BOVIN/942-1187
F6PZF1_HORSE/836-1072
AGRE1_HUMAN/599-840
F6R185_XENTR/35-277
CRFR1_RAT/116-359
CALCR_RAT/145-424
PDFR1_CAEEL/164-425
CALRL_HUMAN/138-380
GIPR_HUMAN/134-388
GLP1R_RAT/141-398
VIPR1_RAT/140-386
F1R383_DANRE/124-372
SCTR_RABIT/135-380
VIPR2_HUMAN/123-371
PACR_MOUSE/150-424
A0A0R4IKR9_DANRE/125-371
GLR_HUMAN/138-396
GHRHR_MOUSE/126-372
PTH1R_HUMAN/184-455
PTH2R_HUMAN/141-409
GLP2R_RAT/175-432
0
0.05
0.1
0.15
0.2
0.25
0.3

- Árvore 3:


PDFR1_CAEEL/164-425
AGRL3_BOVIN/942-1187
F6PZF1_HORSE/836-1072
CD97_HUMAN/544-782
CD97_MOUSE/526-766
AGRE1_HUMAN/599-840
CELR1_MOUSE/2480-2712
CELR3_RAT/2534-2766
AGRG2_HUMAN/625-875
CALRL_HUMAN/138-380
CALCR_RAT/145-424
AGRB1_HUMAN/944-1180
AGRB2_HUMAN/929-1198
A0A0R4IKR9_DANRE/125-371
VIPR2_HUMAN/123-371
SCTR_RABIT/135-380
F1R383_DANRE/124-372
VIPR1_RAT/140-386
PACR_MOUSE/150-424
GLP2R_RAT/175-432
GHRHR_MOUSE/126-372
GLP1R_RAT/141-398
GIPR_HUMAN/134-388
GLR_HUMAN/138-396
PTH2R_HUMAN/141-409
PTH1R_HUMAN/184-455
CRFR1_RAT/116-359
F6R185_XENTR/35-277
0
2
4
6
8
10

3. Metodologia

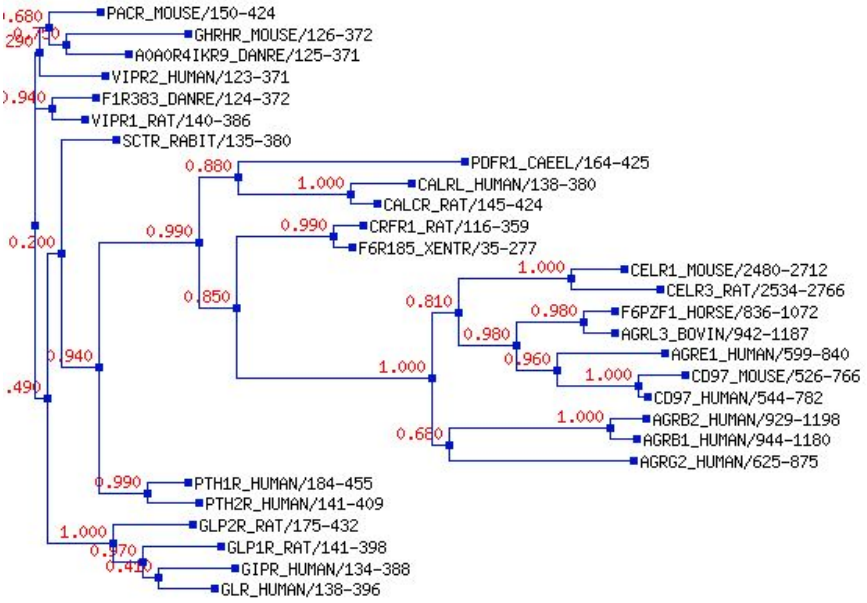
Após a obtenção das sequências na variável “seqs”, em primeiro lugar, executa-se o script fornecido “montaA.m” para obter a matriz A e os nomes das proteínas estudadas em separado.

Em seguida, inicia-se o método para obtenção da árvore 1. Primeiramente, executa-se o comando “seqpdist”, que calcula, par a par, a distância algébrica entre as sequências fornecidas. Posteriormente, a distância é elevada ao quadrado e, por fim, executa-se o comando seqneighjoin, que constrói a árvore filogenética utilizando o método de união de vizinhos.

Para a construção da árvore 2, em primeiro lugar, aplica-se a decomposição por valores singulares (SVD) na matriz de distâncias (calculada no método anterior). Além disso, define-se que serão usados 5 valores singulares para a construção da árvore. Repete-se, por fim, os cálculos para obtenção das distâncias supracitados e tem-se, como resultado, a árvore construída.

Por fim, para a árvore 3, tem-se a decomposição svd na matriz esparsa A. Em seguida, são executados os comandos para que se encontre as distâncias entre as sequências. Finalmente, a última árvore é obtida.

Abaixo, pode-se observar a árvore usada como referência, obtida no site do pfam:



Portanto, é bastante perceptível que, à medida com que os ruídos são retirados das matrizes de distâncias, os resultados obtidos para as árvores

filogenéticas melhoram. Em particular, é bastante significativo observar como essa melhoria é gradual; ou seja, quanto mais refina-se o método, melhores são seus resultados. Assim, tem-se que a árvore 3 é a mais fiel quando comparada à de referência, seguida pela árvore 2, e, finalmente, pela árvore 1.

4. Conclusões

O trabalho em questão foi bastante útil, primeiramente, para que se observasse a aplicação da técnica de decomposição SVD no contexto das proteínas. Conforme visto nas aulas, reconhecer padrões e semelhanças entre proteínas é fundamental para diversas pesquisas e descobertas no campo biológico. Dessa forma, ser capaz de unir o poder da mineração de dados ao estudo das proteínas foi muito proveitoso.

Além disso, o trabalho permite constatar, usando as ferramentas da bioinformática, a teoria da evolução proposta por Darwin. Com a construção da árvore filogenética da família de proteínas PF00002, é possível observar como tais estruturas evoluíram durante o tempo.

Ademais, é também interessante poder refletir a respeito das interações geradas a partir da evolução de tais proteínas com as espécies em questão, como em suas consequências para certas doenças, por exemplo.

5. Referências Bibliográficas

- Pfam database - https://pfam.xfam.org/
- seqpdist - Help Center - https://www.mathworks.com/help/bioinfo/ref/seqpdist.html
- seqneighjoin - Help Center - https://www.mathworks.com/help/bioinfo/ref/seqneighjoin.html

6. Anexos

Além dos scripts fornecidos junto ao enunciado e do arquivo “PF00002.fasta”, foi usada a seguinte sequência de comandos no trabalho:

```
seqs = fastaread('PF00002.fasta')
[nomes, A] = montaA(seqs);
Dist = seqpdist(seqs,'Method','alignment-score',...
```

```
      'Indels','pairwise-delete',...
      'ScoringMatrix','pam250',...
      'PairwiseAlignment',true);
Dist = squareform (Dist);
arvore1 = seqneighjoin(Dist, 'equivar', nomes);
view(arvore1)

[U, S, V] = svd(Dist);
k = 5;
Distk= U(:, 1:k)*S(1:k, 1:k)*V(:, 1:k)';
arvore2 = seqneighjoin(Distk, 'equivar', nomes);
view(arvore2)

[T, S, D] = svds(A);
l = 5;
Auxl = S(1:l, 1:l)*D(:, 1:l)';
Distl = pdist(Auxl');
Distl = squareform(Distl);
arvore3 = seqneighjoin(Distl, 'equivar', nomes);
view(arvore3)
```